Nemotron-Labs-Audio-Visual Flamingo: Open Audio-Visual Intelligence for Long and Complex Videos

Sreyan Ghosh[1,2,*] Arushi Goel[1,*] Kaousheik Jayakumar[2] Lasha Koroshinadze[2]
Nishit Anand[2] Siddharth Gururani[1] Hanrong Ye[1] Pritam Biswas[1] Yuanhang Su[1]
Ehsan Hosseini-Asl[1] Sang-gil Lee[1] Zhifeng Kong[1] Jaehyeon Kim[1] Sungwon Kim[1]
Karan Sapra[1] S Sakshi[2] Ramani Duraiswami[2] Dinesh Manocha[2] Andrew Tao[1]
Mohammad Shoeybi[1] Bryan Catanzaro[1] Ming-Yu Liu[1] Wei Ping[1]

[1]NVIDIA, USA [2]University of Maryland, USA

Code Model Project Page Dataset Demo

Abstract. We present **Nemotron-Labs-Audio-Visual Flamingo** (AV-Flamingo), a fully open state-of-the-art audio-visual large language model (AV-LLM) for joint understanding and reasoning over audio, images, and long-form videos. Unlike prior AV-LLMs that primarily focus on short clips, AV-Flamingo is designed for understanding and reasoning over long and complex real-world (audio-visual) videos. To support this, we make three key contributions: **(i)** Audio-Visual-Skills, a large-scale collection of real-world videos with ≈7M caption and question-answer training instances designed to emphasize temporal, compositional, and cross-modal audio-visual reasoning; **(ii)** a novel three-stage curriculum that progressively trains the model from short-range perception to long-horizon multi-event reasoning; and **(iii)** Temporal Audio-Visual Interleaved Chain-of-Thought, a reasoning framework that explicitly grounds intermediate reasoning steps to timestamps in long audio-visual streams, improving temporal alignment and interpretability. Extensive experiments across 15+ AV, omni-modal, audio, and vision benchmarks show that AV-Flamingo outperforms similarly sized open models by clear margins and remains highly competitive with, and in some cases surpasses, much larger open-weight and closed models, particularly on long and complex real-world audio-visual understanding and reasoning tasks. Beyond benchmark performance, AV-Flamingo exhibits strong real-world utility and transfers well to unseen tasks, highlighting its robustness and generalization ability.

1. Introduction

Video constitutes 82% of global internet traffic, and users spend over 2.5 hours per day watching online videos on average (Kumar, 2026). These videos are often long, diverse, and information-rich, combining complex visuals, graphics, speech, sounds, and music in a continuous stream. This closely mirrors how humans perceive the world: as an ongoing audio-visual sequence. Yet, despite the remarkable progress of Large Language Models (LLMs) in many real-world tasks, their ability to understand and reason over videos in a human-like manner remains relatively underexplored. Hereafter, unless stated otherwise, we use *video* to refer to audio-visual sequences.

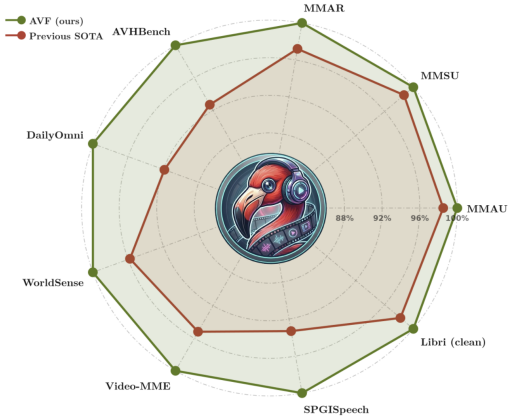

Figure 1: Comparison of AVF with current SOTA model on various benchmarks.

Most video understanding models still do not process

*Project-Leads. Ordering was decided with a coin toss.

audio jointly with visuals (Zhang et al., 2023, 2025; Zhu et al., 2025). However, audio–including speech, sounds, and music–is a fundamental modality and has even been described as "50% of the movie-going experience" (Lucas, 1992).

Recent audio-visual and omni-modal LLMs have shown promise (Xu et al., 2025a; Team, 2026b), but most focus primarily on short-video understanding (Chowdhury et al., 2024). In contrast, much of the video content humans consume is medium- to long-form, such as movies, television, lectures, and documentaries.

A central challenge for long-form audio-visual understanding is the lack of suitable data: publicly available resources are often audio-only, video-only, or concentrated on short-form videos with captions and QA annotations. Moreover, current models still exhibit substantial weaknesses in joint AV perception (Gong et al., 2024; Zhou et al., 2026; Goel et al., 2026; Chowdhury et al., 2025; Fu et al., 2025a), which we attribute to both architectural limitations and the scarcity of large-scale, high-quality paired audio-visual data (Seth et al., 2025; Selvakumar et al., 2025). Recent mechanistic analysis attributes this partly to a visual bias in which the deeper layers privilege vision and suppress latent audio representations (Selvakumar et al., 2026). In practice, many existing omni models are trained separately for audio and visual understanding, and are expected to acquire cross-modal reasoning only implicitly. Finally, the strongest models for audio-visual understanding are either closed-source (Team et al., 2024) or only open-weight (Xu et al., 2025b; Team, 2026b), with their training data, code, and methodology undisclosed. This limits both open progress and understanding of how such systems are built.

Main Contributions. In this work, we present **Nemotron-Labs-Audio-Visual Flamingo (AV-Flamingo or AVF)**, a fully open multimodal large language model for joint audio-visual understanding of long and complex real-world videos. AV-Flamingo is a first step toward scaling open audio-visual intelligence beyond academic benchmarks by leveraging internet-scale audio-visual data and targeted post-training for reasoning. To enable this, AV-Flamingo centers on three technical components. First, we introduce **Audio-Visual-Skills (AV-Skills)**, a large-scale dataset for joint audio-visual understanding and reasoning. Unlike prior omni models that rely heavily on single-modality data and hope to learn cross-modal reasoning implicitly, AV-Skills is explicitly curated for cross-modal learning. It contains videos collected from diverse sources, paired with captions and QA annotations designed to test and train audio-visual reasoning, totaling ≈7M caption and QA training instances, including ≈4.8M QA pairs, across both short and long videos. Second, we propose **Temporal Audio-Visual Interleaved Chain-of-Thought (TAVIT)**, a reasoning framework that explicitly grounds intermediate reasoning steps to timestamps in long audio-visual streams. Third, we develop a novel three-stage training curriculum consisting of pre-training, mid-training, and post-training with different data mixtures. AV-Flamingo outperforms similarly sized or larger models across more than 15 audio-visual, omni-modal, audio, and vision benchmarks. In summary, our main contributions are:

1. We introduce **Nemotron-Labs-Audio-Visual Flamingo**, a fully open frontier AVLLM for joint audio-visual understanding and reasoning over long and complex real-world videos.
2. We present a scalable recipe for next-generation AV-LLMs, including internet-scale audio-visual data curation, targeted capability expansion, and temporally grounded reasoning for long videos.
3. We open-source the model, training, and inference code, and associated techniques to support future research in open audio-visual large language models.
4. AV-Flamingo outperforms similar-sized AV and omni-modal LLMs on 15+ benchmarks, and remains highly competitive with, and in some cases surpasses, much larger open-weight and closed models. Additionally, it shows substantially stronger robustness on long and complex real-world videos.

2. Related Works

Audio-Visual Large Language Models. Recent progress in multimodal LLMs has produced increasingly capable vision-language models, from foundational work like Flamingo (Alayrac et al., 2022), BLIP-2 (Li

et al., 2023), and LLaVA (Liu et al., 2023b) to recent systems such as InternVL 2.5 (Chen et al., 2024d), Qwen3-VL (Bai et al., 2025), VideoLLaMA 3 (Zhang et al., 2025), and Qwen3.5 (Team, 2026a). In parallel, audio-language models such as LTU (Gong et al., 2023b), SALMONN (Tang et al., 2023), Qwen2-Audio (Chu et al., 2024), and Audio Flamingo 3 (Goel et al., 2025) have advanced speech, sound, and music understanding. However, integrating audio and visual streams into a single model remains qualitatively harder, requiring cross-modal alignment, temporal synchronization, and handling of diverse input combinations. Early AV-LLMs such as Video-LLaMA (Zhang et al., 2023) used separate Q-Formers for each modality; Video-LLaMA 2 (Cheng et al., 2024) improved on this with a spatial-temporal convolution connector; video-SALMONN (Sun et al., 2024; Tang et al., 2025) proposed multi-resolution causal Q-Formers; Dolphin (Guo et al., 2025) introduced multi-scale adapters; and TriSense (Li et al., 2025d) proposed query-based adaptive modality reweighting. On the omni-modal frontier, proprietary systems like GPT-4o (Hurst et al., 2024) and Gemini 1.5 Pro (Team et al., 2024) set high benchmarks, while open models have rapidly closed the gap: Qwen2.5-Omni (Xu et al., 2025a) introduced the thinker-talker architecture, Qwen3-Omni (Xu et al., 2025b) and Qwen3.5-Omni (Team, 2026b) scaled this further, Phi-4-Multimodal (Abouelenin et al., 2025) demonstrated Mixture-of-LoRAs for compact omni-modal models, and OmniVinci (Ye et al., 2025) introduced shared omni-modal alignment with a 24M-sample synthetic pipeline. However, most of these systems focus on short-to-medium clips rather than long-form, temporally grounded audio-visual reasoning. Furthermore, many remain closed-source or open-weights-only, hindering reproducibility – a gap AV-Flamingo addresses.

Long-Form Video Understanding and Temporal Reasoning. Understanding long videos remains one of the central open challenges in multimodal AI, as most current video LLMs are still limited to clips shorter than one minute. Recent progress in long-form video understanding has focused largely on vision-only settings, often without audio (Li et al., 2024b; Cheng et al., 2024), and many approaches rely on context reduction to handle long videos, which can hurt performance in real-world scenarios (Zhang et al., 2024a; Qian et al., 2024; Li et al., 2025c). Recent omni models are a notable exception, with some beginning to support long audio-visual sequences Team (2026b); Deshmukh et al. (2026). However, benchmarks like Video-MME (Fu et al., 2025a) and MMOU (Goel et al., 2026) reveal systematic performance degradation as video duration increases – MMOU shows the best proprietary models achieve only 64.2% accuracy on long audio-visual reasoning, while open-source models plateau at 46.8%. Separately, temporal chain-of-thought reasoning has emerged as a promising direction: Temporal CoT (Arnab et al., 2025) narrows attention progressively for long-video QA, Open-o3-Video (Meng et al., 2025) integrates spatiotemporal evidence into reasoning chains, UniTime (Liu et al., 2024) interleaves timestamp tokens with video tokens, and Audio Flamingo 3 (Goel et al., 2025) introduced on-demand thinking for audio events. In the audio domain, R1-AQA (Li et al., 2025a) and Step-Audio-R1 (Tian et al., 2025) have shown CoT gains, though primarily on short audio. However, no prior work jointly grounds intermediate reasoning steps to timestamps in both audio and visual streams, and we propose to bridge this gap with AV-Think.

Data Curation. High-quality data remains a major bottleneck in building multimodal foundation models, including for audio (Panayotov et al., 2015; Gemmeke et al., 2017; Galvez et al., 2021), video (Miech et al., 2019; Bain et al., 2022; Chen et al., 2024b), and especially long-form understanding. This challenge is even more pronounced for audio-visual models (Liu et al., 2025b), where high-quality joint audio-visual supervision is rarely available in the open (Geng et al., 2025). Most existing resources are either unimodal or, when audio-visual, are largely limited to short videos and recognition-oriented tasks. For example, many current audio-visual LLMs are trained primarily on short-form audio-visual data (Sun et al., 2024; Cheng et al., 2024; Xu et al., 2025a; Fu et al., 2025b), while models such as OmniVinci rely heavily on separate unimodal audio and video corpora. To bridge this gap, we introduce Audio-Visual-Skills, a large-scale dataset designed to support temporal, compositional, and cross-modal reasoning across 13 skill categories in long, real-world videos.

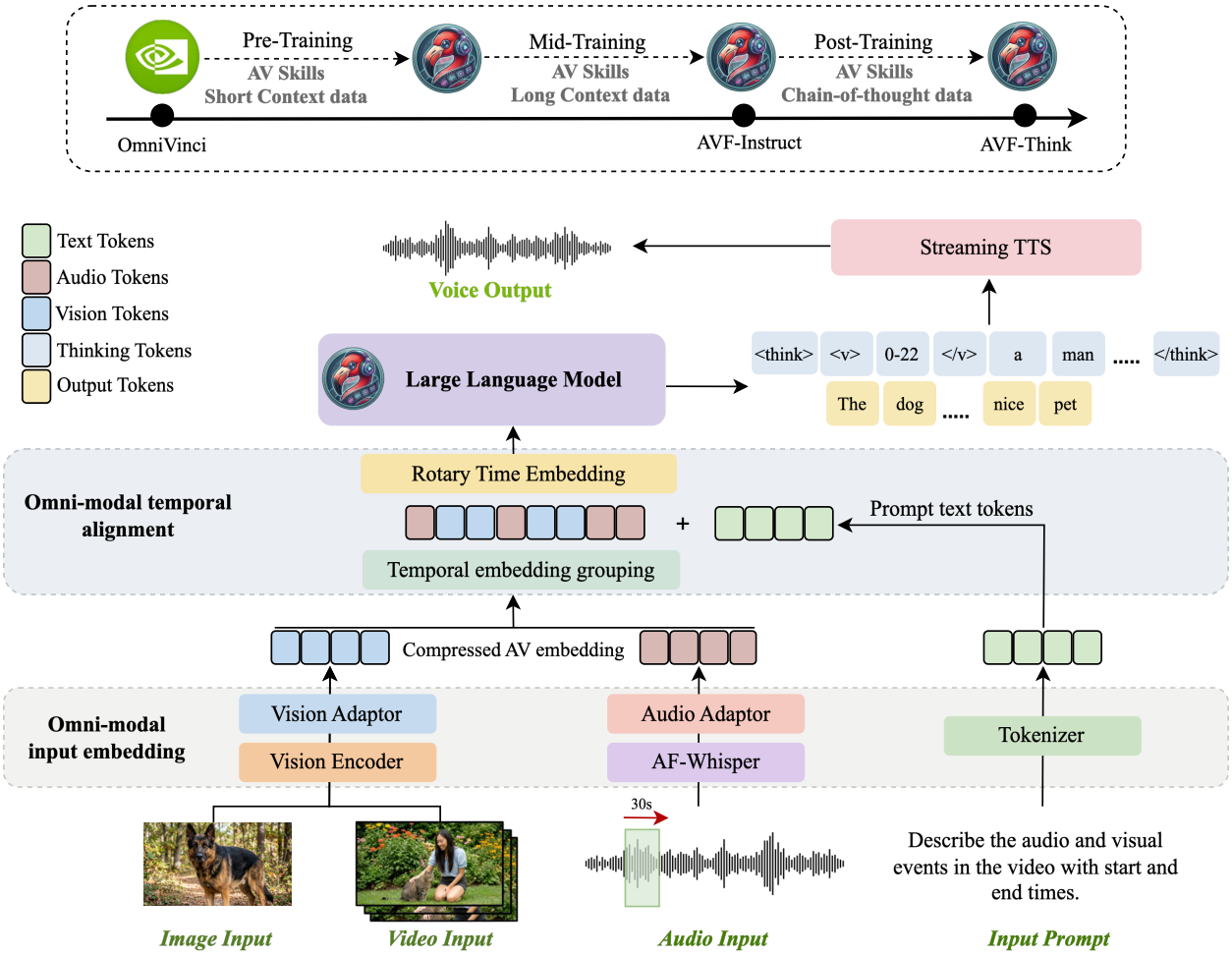

Figure 2: **AV-Flamingo training and architecture. Top:** Starting from OmniVinci, AV-Flamingo is trained in three stages: pre-training on AV-Skills short-context data, mid-training on AV-Skills long-context data, and post-training on chain-of-thought data, producing AVF-Instruct and AVF-Think. **Bottom:** AV-Flamingo accepts image, video, audio, and text inputs. Visual and audio streams are encoded through separate encoders and adaptors, compressed into audio-visual embeddings, temporally grouped, and aligned with prompt text tokens using Rotary Time Embedding before being processed by the LLM.

3. Methodology

3.1 Architecture

In this section, we discuss the architecture for Audio-Visual Flamingo as shown in Figure 2. AVF is an AVLM designed to jointly reason over visual and auditory inputs (both or either) while supporting voice-based interaction through speech synthesis. The architecture is similar to OmniVinci and consists of five key components: i) a SigLip vision encoder that extracts rich spatial features from visual inputs (*e.g.*, images, video frames), ii) AF-Whisper (borrowed from the Audio Flamingo series), an audio encoder with a sliding-window feature extraction mechanism for processing long-form audio, iii) a cross-modal interleaving and temporal alignment module that fuses video and audio representations along the time axis, iv) a text-only LLM that serves as the central reasoning backbone, and v) a streaming text-to-speech (TTS) module that enables real-time voice output. We discuss each component in detail below.

SigLip Vision Encoder. To encode visual inputs, we employ the SigLip (Zhai et al., 2023) vision encoder,

which produces dense feature representations from images or video frames, V. Given a raw input, we first extract feature maps from the pre-trained SigLip model. These feature maps are then passed through a "Spatial-Scale-then-Compress" Dynamic S2 module, following the design principles of Liu et al. (2025a) and Ye et al. (2025). The Dynamic S2 module operates by first encoding the input image at multiple spatial scales, then compressing the multi-scale representations into a compact token sequence. This multi-scale strategy enables AVF to encode higher-resolution images and longer videos with more frames while preserving fine-grained spatial and temporal details – critically, without a proportional increase in the total number of tokens consumed by the LLM. As a result, we obtain a feature representation for a single image (or each frame of a video) denoted as $h_v = f_v(V)$, where $h_v \in \mathbb{R}^{HW \times d_v}$, H and W are the spatial height and width of the compressed visual feature map, and d_v is the hidden dimension of the vision encoder.

AF-Whisper Audio Encoder. Following Audio Flamingo 3, Next, and Music Flamingo, we adopt the Whisper-based AF-Whisper audio encoder from AF-Next (Goel et al., 2025; Ghosh et al., 2025a). For each audio input, A, we first resample it to 16 kHz mono to standardize the sampling rate across diverse audio sources. The raw resampled waveform is then transformed into a 128-channel log-mel spectrogram using a window size of 25 ms and a hop size of 10 ms, producing a time-frequency representation suitable for the downstream encoder. This mel-spectrogram is processed by the encoder (Goel et al., 2025; Ye et al., 2025), which applies a sliding-window mechanism to handle audio of arbitrary length: the spectrogram is segmented into non-overlapping 30-second chunks, each of which is independently encoded and then concatenated along the temporal axis. This yields audio features denoted as $h_a = f_a(A)$, where $h_a \in \mathbb{R}^{N \times d_a}$, N is the total number of 30-second audio chunks, and d_a is the hidden dimension of the Whisper encoder. The sliding-window design allows AVF to process long-form audio (e.g., full-length podcast episodes or movie soundtracks) without truncation or excessive memory consumption.

Vision and Audio Adaptors. For each modality, we have projection layers, specifically, a 2-layer MLP for audio, denoted $\mathcal{A}(\cdot)$, and a 2-layer MLP for vision, denoted $\mathcal{V}(\cdot)$. These projections map the encoder outputs into the LLM's embedding space: the projected audio embeddings are $a = \mathcal{A}(h_a)$ and the projected image/video embeddings are $v = \mathcal{V}(h_v)$.

Cross-Modal Alignment of Vision and Audio. A central challenge in multimodal architectures is the fusion of visual and audio tokens before feeding into the LLM. Following OmniVinci (Ye et al., 2025), we enforce a cross-modal alignment between the two modalities through temporal interleaving. The projected video and audio token sequences are first divided along the time dimension into multiple synchronized chunks and then interleaved according to their timestamps, so that visual tokens from a given temporal window are placed adjacent to the audio tokens from the same window. This interleaved arrangement allows the LLM's self-attention mechanism to naturally attend across co-occurring visual and auditory events. Finally, before injecting these interleaved token embeddings into the LLM, we augment them with periodic temporal information via Constrained Rotary Time Embeddings (CRTE) (Goel et al., 2024; Ye et al., 2025). This encodes absolute temporal position using sinusoidal rotary transformations, enabling the model to distinguish the ordering and relative timing of multimodal events even after interleaving.

Large Language Model (LLM). The core reasoning component of AVF is a large language model that processes the interleaved multimodal token sequence and generates textual responses. Following OmniVinci (Ye et al., 2025), we employ Qwen2.5-7B (Team., 2025) as our base LLM, which comprises 7B parameters, 36 hidden layers, and 16 attention heads per layer. The LLM receives the temporally aligned, interleaved vision-audio embeddings as input prefix tokens, followed by any text-based instructions or queries, and autoregressively generates a textual response. To handle long-context training for long videos of upto 15 minutes long as discussed in Section 3.3, we use hybrid sequence parallelism (Fang and Zhao, 2024) along with fully-sharded data parallel (Rasley et al., 2020). This approach reduces memory usage by splitting the multi-modal sequence dimension across GPUs, using Ulysses (Jacobs et al., 2023) for intra-node communication and Ring-Attention (Liu et al., 2023a) for inter-node communication. Optionally, to support voice-to-voice interaction, similar to Audio Flamingo 3 (Goel et al., 2025) and OmniVinci (Ye et al., 2025), Audio-Visual Flamingo incorporates a streaming TTS module. For more details, we refer our readers to Goel et al. (2025).

Streaming TTS. To support voice-to-voice interaction, similar to Audio Flamingo 3 (Goel et al., 2025) and

OmniVinci (Ye et al., 2025), Audio-Visual Flamingo incorporates a streaming TTS module. The module is implemented as a decoder-only transformer that predicts the next audio token conditioned on incoming subword text tokens from the LLM and previously generated audio tokens. For more details, we refer our readers to Goel et al. (2025).

3.2 Training Data

Data remains one of the most critical yet least openly discussed components in building open foundational omni-modal and audio-visual models (Ye et al., 2025). While recent efforts have improved the availability of large-scale audio-only (Galvez et al., 2021; Ghosh et al., 2026) and image/video-only datasets (Schuhmann et al., 2022; Zhang et al., 2024b), comparable resources for joint audio-visual learning remain limited, particularly for tasks requiring integrated understanding of sound and vision and cross-modal reasoning. Existing audio-visual QA datasets, often derived from foundational benchmarks centered on recognition tasks such as AVSR, event classification, or AV localization (Chen et al., 2020; Tian et al., 2018), are insufficient for training models with strong expert-level reasoning abilities (Goel et al., 2026).

In AVF, we place particular emphasis on developing reasoning and problem-solving capabilities through large-scale, high-quality audio-visual QA data. Inspired by the Audio Flamingo series, in addition to existing unimodal image/audio/video datasets (largely borrowed from OmniVinci, with some additions, detailed stats in Table 2), we introduce **Audio-Visual-Skills (AV-Skills)**, a large-scale dataset of real-world videos with diverse durations, sourced from both existing datasets and the open internet, paired with high-quality captions and QA annotations. Unlike prior data mixtures that rely heavily on unimodal supervision, AV-Skills is curated specifically for joint cross-modal audio-visual understanding, with a focus on teaching models complex skills beyond recognition, including reasoning over temporally extended, compositional, and real-world long-form videos.

As a first step in data curation, we identify the key skills required for strong real-world audio-visual understanding. To do so, we evaluate frontier omni-modal and audio-visual models, including the Qwen-Omni series, Gemini, and VideoLLaMA, on challenging benchmarks such as WorldSense and MMOU, and analyze their outputs to identify systematic capability gaps. In particular, we convert benchmark QA pairs into open-ended form, since multiple-choice settings can allow models to exploit option biases or elimination heuristics. Our analysis reveals both gaps in core skills, such as counting and temporal understanding, and distributional weaknesses caused by limited training exposure to diverse video types. For instance, much of the data used in existing open models is dominated by vlog-style talking-head videos, with relatively limited coverage of environmental sounds, background music, and complex real-world audio-visual scenes.

To address these limitations, we curate two large-scale training sets: **AV-Skills-Short**, consisting of videos of up to 60 seconds, and **AV-Skills-Long**, consisting of videos between 60 seconds and 15 minutes. These datasets are built from two sources: publicly available datasets and raw videos collected from the open internet, which we subsequently annotate synthetically. For public datasets, we primarily use YouTube-8M (Abu-El-Haija et al., 2016), HD-VILA (Xue et al., 2022), InternVid (Wang et al., 2024b), and VidChapters (Yang et al., 2023) for short-form data, and HarmonySet (Zhou et al., 2025), LSMDC (Rohrbach et al., 2016), MMTrail (Chi et al., 2024), MovieClips (Bose et al., 2022), and MiraData (Ju et al., 2024) for longer-form data. The remaining data is collected from the open internet with an emphasis on categorical diversity (Goel et al., 2026), spanning domains such as podcasts, city tours, interviews, sports, and more (Fig. 4). Our final AV-Skills-Short dataset comprises 100K hours of video and 3.8M training instances, including 1M captions and 2.8M QA pairs (detailed stats in Table 2). The long-video training data comprises approximately 140K hours of video and 3.2M training instances, including 1.2M captions and 2.0M QA pairs. For **AV-Skills-Short**, we focus on the following skills (examples shown in Section B): (i) **Relation Reasoning** involves understanding the relationships, interactions, or distinctions between entities in a video, including people, objects, roles, and their relative configurations. (ii) **Emotion Change** focuses on tracking how a person's or character's emotional state evolves over time, including the initial emotion, the final emotion, and the transition between them. (iii) **Temporal Reasoning** captures the chronological structure of a video, including event order, what happens before or after a target event, and how actions unfold over time. (iv) **Spatial Perception**

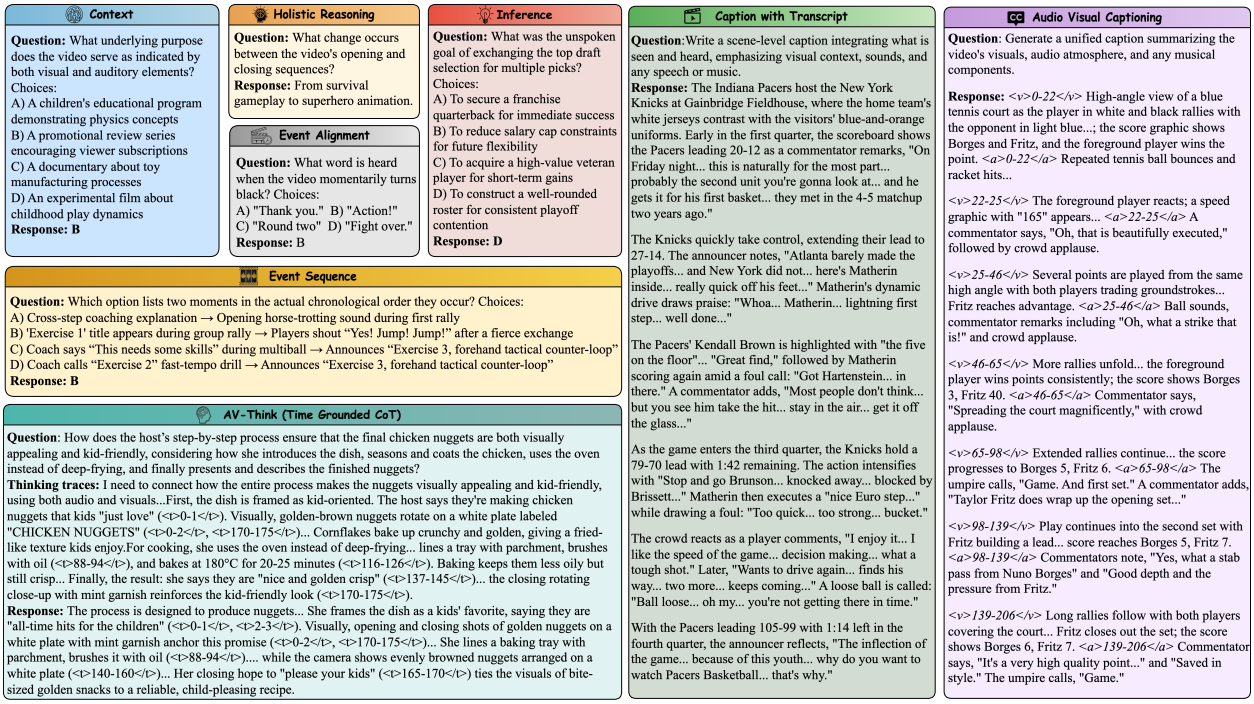

Figure 3: Examples of types of questions in Audio-Visual-Skills. More examples are shown in Fig 5 and Table 3, and details are provided in Appendix B.

requires perceiving spatial properties in the visual scene, such as direction, orientation, relative position, alignment, and placement of people or objects. (v) **Causal Reasoning** aims to infer why an event, action, behavior, or outcome occurs by connecting observed effects to their underlying causes. (vi) **Hallucination Detection** evaluates whether a question is grounded in information actually present in the video, requiring the model to reject false premises or unsupported claims. (vii) **Audio Counting** tests the ability to count the number of times a particular sound, spoken word, instrument, or other audio event occurs in the video. (viii) **Video Counting** measures the ability to count the number of times a particular object, entity, action, or visual event appears in the video.

For **AV-Skills-Long**, we focus on the following skills (examples shown in Fig. 3 and Table 2): (i) **Needle-in-the-Haystack Reasoning:** Answering questions about a specific but important moment in a long video, or a localized event that still requires broader audio-visual context to identify correctly. (ii) **Temporal Reasoning:** ***a) Temporal Referring:*** Identify what happens immediately before or after a target event. ***b) Temporal Order:*** Determine the correct order of multiple audio-visual events. ***c) Temporal Attribute:*** Reason about how scene dynamics such as intensity, pace, or atmosphere evolve over time. (iii) **Sub-scene Understanding:** Understanding and describing a meaningful intermediate portion of a long video based on the events that precede and follow it. (iv) **Holistic Reasoning:** Inferring high-level goals, motivations, or outcomes by integrating audio-visual evidence distributed across a substantial part of the video. (v) **Counting:** Counting the number of times a particular audio-visual event, interaction, or recurring pattern appears throughout the video. (vi) **Audio-Visual Referring:** Linking an event in one modality to its counterpart in the other, such as identifying the visual scene associated with a sound or the sound associated with a visual event. (vii) **Topic-level Reasoning:** Identifying the main activity, objective, or overall theme of the video by synthesizing evidence across the full timeline. (viii) **Detailed Captioning:** Generating a comprehensive audio-visual description of the full video, capturing key actions, scene changes, sounds, and spoken content without redundancy. (ix) **Event Sequence Reasoning:** Determining the relative ordering of key audio-visual events across the video timeline. (x) **Audio-Visual Event Alignment:** Identifying the exact sound that coincides with a visual moment, or the exact visual event synchronized with a target audio cue. (xi) **Inference:** Drawing conclusions about implicit intentions, causes, or outcomes

by combining multiple audio-visual clues that are not stated directly. (xii) **Comparative Reasoning:** Identifying important differences or similarities between two audio-visual events, scenes, or presentations in the same video. (xiii) **Context Understanding:** Inferring the broader situational setting, background context, or scene conditions by jointly reasoning over visual details, sounds, and speech.

AV-Think. Finally, we introduce **Temporal Audio-Visual Interleaved Chain-of-Thought (TAVIT)**, a reasoning framework that teaches the model to ground intermediate reasoning steps to timestamps in interdependent audio and visual streams (example in Fig. 3). Prior work on CoT for video understanding remains limited, largely focusing on video-only settings or shallow reasoning traces in recent omni models. Existing approaches often yield only modest gains, and sometimes even degrade performance (Team, 2026b), partly because current audio or audio-visual CoT data is mostly restricted to short clips and simple QA pairs with reasoning appended post hoc.

We argue that explicit reasoning is most useful for long, complex real-world videos, where evidence is distributed across multiple, overlapping, and temporally dispersed events. To address this gap, we construct **AV-Think**, a dataset of question-answer-reasoning triplets for temporally grounded audio-visual reasoning. Unlike prior work, AV-Think explicitly interleaves audio and visual evidence within timestamped reasoning steps, bridging reasoning accuracy with temporal grounding.

AV-Think is curated from challenging long-form videos, including trailers, movie recaps, mystery stories, and multi-party conversations, paired with questions requiring extended temporal reasoning. We ground reasoning to time because **(i)** timestamped thoughts help the model navigate and reason over long audio-visual streams, and **(ii)** grounding intermediate reasoning in temporal events can improve recognition performance (Kumar et al., 2026). To build the dataset, we first generate time-stamped captions for each video using a TAC-style pipeline, and then prompt an LLM over these captions to synthesize reasoning triplets (see Prompt in Fig 20). AV-Think contains approximately 24K training samples, with reasoning chains averaging 635.7 words.

3.3 Training Curriculum

We design a three-stage curriculum to train AVF. In each stage, we employ a distinct data mixture designed to gradually increase context length and complexity of tasks. We hypothesize that different capabilities emerge at different stages of training: foundational auditory and visual skills are best acquired early on unimodal and short-context data, during which the model also acquires weak audio-visual alignment. On the other hand, strong audio-visual alignment, long-context understanding, and more complex skills like temporal understanding require later-stage specialization with long, real-world, complex data. We provide the full data mixing ratios in Table 2 and describe the training technique, including training hyperparameters, in Section 4.

Pre-training. This stage consists of two phases designed to establish foundational capabilities. In **Stage 1: Initialization stage**, we initialize AVF from the pre-trained OmniVinci (Ye et al., 2025) checkpoint, which provides a strong starting point with already-aligned vision, audio, and language representations. In **Stage 2: Short-Context Training**, we perform full fine-tuning of the model on a mixture of unimodal datasets–spanning audio-only and vision-only tasks such as classification, captioning, ASR, and image/video question answering–alongside our newly curated AV-Skills-Short data. The unimodal data ensures that modality-specific capabilities inherited from OmniVinci are preserved and strengthened, while the AV-Skills-Short data introduces cross-modal reasoning abilities unique to Audio-Visual Flamingo. The maximum a-v duration in this stage is capped at 5 minutes, keeping the focus on high-quality short- and medium-length examples where skill-specific supervision is most reliable and scalable. The total context length in this stage is capped at 16K tokens.

Mid-training. Building on the capabilities acquired after stage 2 of pre-training, this stage focuses on extending AVF's understanding to longer audio-visual inputs. We expand the data mixture with our AV-Skills-Long data, focused on long-form captioning, time-aware annotations that require precise temporal grounding, and long-context question answering datasets. To maintain a balanced training distribution, we

Table 1: Performance comparison of AVF model variants on Omni, Audio, Video, and Speech Understanding benchmarks (WER ↓ and ACC ↑). We highlight closed source, open weights, and open source models.

Task	Dataset	Model	Metrics	Results
Omni Understanding	**WorldSense**	Qwen2.5-O OmniVinci AVF-Instruct AVF-Think	ACC ↑	45.4 48.2 **50.3** 51.6
	DailyOmni	OmniVinci AVF-Instruct AVF-Think	ACC ↑	66.5 **72.4** 73.9
	OmniBench	Gemini-1.5 Pro AVF-Instruct AVF-Think	ACC ↑	47.6 **48.5** 50.6
	MMOU	Gemini-2.5 Pro Minicpm-o 4.5 AVF-Instruct AVF-Think	ACC ↑	**64.2** 46.8 56.9 60.2
	AVHBench *A->V Hall. / V->A Hall.*	Gemini-2.0 Flash AVF-Instruct AVF-Think	ACC ↑	**83.3** | 63.3 77.0 | **81.1** 79.0 | 85.9
Audio Understanding	**MMAR**	OmniVinci AVF-Instruct	ACC ↑	58.4 **60.1**
	MMSU	Gemini 1.5 Pro AVF-Instruct	ACC ↑	60.7 **61.5**
	MMAU-v05.15.25 (test) *Sound / Music / Speech / Avg*	Audio Flamingo 3 OmniVinci AVF-Instruct	ACC ↑	75.83 | 74.47 | 66.97 | 72.42 73.57 |73.07| 68.17 | 71.60 **77.97** |**73.17**| **69.33** | **73.49**
	CMM Hallucination	Gemini 2.5 Pro AVF-Instruct	ACC ↑	82.0 **86.7**
Video Understanding	**Video-MME** *w/o subtitles / w/ subtitles*	NVILA OmniVinci AVF-Instruct	ACC ↑	64.2 | - 67.3 | 68.6 **70.7** | **71.2**
	LongVideoBench	OmniVinci NVILA AVF-Instruct	ACC ↑	**62.0** 58.7 60.1
	MVHBench	OmniVinci AVF-Instruct	ACC ↑	70.6 **71.7**
Automatic Speech Recognition	**LibriSpeech (en)** *test-clean / test-other*	Phi-4-mm | Qwen2.5-O AVF-Instruct	WER ↓	1.67 | **3.4** **1.64** | 3.5
	SPGISpeech (en)	Qwen2-A-Inst AVF-Instruct	WER ↓	3.0 **2.8**
	TEDLIUM (en)	Phi-4-mm AVF-Instruct	WER ↓	**2.9** 3.0
	GigaSpeech (en)	Phi-4-mm AVF-Instruct	WER ↓	**9.78** 10.2
	VoxPopuli (en)	Phi-4-mm AVF-Instruct	WER ↓	5.9 **5.8**

additionally include a down-sampled subset of the AV-Skills-Short data from the pre-training stage, ensuring that previously acquired skills are retained while the model learns to reason over extended temporal contexts. The maximum audio/video duration in this stage is increased to 15 minutes, and the total context length is expanded to 32K tokens. The fully trained model resulting from this process is referred to as **AVF-Instruct**.

Post-training. Finally, we train AVF for temporally grounded CoT on the AV-Think dataset. Starting from AVF-Instruct, we first perform SFT on AV-Think that encourages the model to produce structured, step-by-step rationales anchored to specific instances in the audio-visual input. We then train the model using GRPO-based RL (Shao et al., 2024) (more details in Section E) to further strengthen the model's reasoning quality. The final model obtained from this stage is referred to as **AVF-Think**.

4. Experiments

Experimental Setup. We perform pre-training, long-context training, and post-training of Audio Visual Flamingo on 512 NVIDIA H100 GPUs. Further details on batch size, learning rates, and optimizers for each stage of training are in Appendix C.

Baselines. We evaluate our proposed model against recent SOTA open-source and proprietary MLLMs (Large Multimodal Language Models, including Qwen2.5-O(mni) (Xu et al., 2025a), Qwen3-O(mni) (Xu et al., 2025b), OmniVinci (Ye et al., 2025), Gemini (2.0 Flash, 1.5 Pro, 2.5 Flash and 2.5 Pro) (Team et al., 2024; Team, 2025) as well as GPT-4o (Hurst et al., 2024). Furthermore, for audio and speech understanding, we also compare against specialized Large Audio Language Models (LALMs) such as Audio Flamingo (Kong et al., 2024), Audio Flamingo 2 (Ghosh et al., 2025b), Audio Flamingo 3 (Goel et al., 2025), Qwen-A(udio) (Chu et al., 2023), Qwen2-A(udio) (Chu et al., 2024), Qwen2-A(udio)-(Inst)ruct, R1-AQA Li et al. (2025a), Pengi (Deshmukh et al., 2024), Phi-4-mm (Abouelenin et al., 2025), Baichun Audio (Li et al., 2025b), Step-Audio-Chat (Tian et al., 2025), LTU (Gong et al., 2023b), LTU-AS (Gong et al., 2023a), SALMONN (Tang et al., 2023) and AudioGPT (Huang et al., 2023). Similarly, for video understanding, we compare against specialized vision language models (VLMs) such as VILA (Lin et al., 2024), LongVILA (Chen et al., 2024c), NVILA (Liu et al., 2025a), InternVL (Chen et al., 2024e), Qwen2-VL (Wang et al., 2024a), LLaVA-OneVision (Li et al., 2024a), and Florence-VL (Chen et al., 2024a).

Evaluation Datasets. We evaluate our model on a diverse set of benchmarks spanning *audio understanding* (MMAR (Ma et al., 2025), MMSU (Wang et al., 2026), MMAU (v05.15.25) (Sakshi et al., 2024)), which covers sound, music, and speech reasoning; *video understanding* (Video-MME (Fu et al., 2025a), ActivityNetQA (Yu et al., 2019), LongVideoBench (Wu et al., 2024)), capturing visual comprehension as well as temporal reasoning; *omni/multimodal understanding* (WorldSense (Hong et al., 2026) and DailyOmni (Zhou et al., 2026)), evaluating joint reasoning across audio and video signals; and ASR (LibriSpeech (clean and other) (Panayotov et al., 2015), SPGISpeech (O'Neill et al., 2021), TEDLIUM (Hernandez et al., 2018), GigaSpeech (Chen et al., 2021), Common Voice 15), which span clean, noisy, and diverse real-world speech conditions.

5. Results

Audio Understanding and Reasoning. AVF-Instruct shows strong transfer to audio-only understanding and reasoning despite being trained as a joint audio-visual model. It improves over OmniVinci on MMAR (60.1 vs. 58.4) and exceeds Gemini 1.5 Pro on MMSU (61.5 vs. 60.7). On MMAU, AVF-Instruct achieves the best overall average score (73.49), outperforming both AF3 and OmniVinci, with especially strong gains on sound and speech understanding. For ASR, AVF achieves the best result on LibriSpeech test-clean (1.64 WER), SPGISpeech (2.8), and VoxPopuli (5.8), while remaining close to the strongest baselines on others.

Video and Omni Understanding. AVF-Instruct also demonstrates consistent gains on both video-only and omni-modal evaluation. On Video-MME, it achieves the best performance both without subtitles (70.7) and with subtitles (71.2), outperforming both NVILA and OmniVinci. On WorldSense and DailyOmni, AVF-Instruct again sets the strongest results, indicating better joint perception and reasoning across modalities. On LongVideoBench, while OmniVinci remains slightly stronger, AVF-Instruct still outperforms NVILA and remains competitive on long-form video understanding, suggesting that it scales effectively to challenging

real-world video settings. Results on MMOU emphasize AVF is best-in-class for long and complex audio-visual understanding. Further qualitative examples are provided in Table 7 and our project page. **Ablations:** Table 5 summarizes the compact AVF training recipe. Table 6 provides the appendix ablation on our AV-Skills curriculum.

6. Conclusion, Limitations and Future Work

We presented AV-Flamingo, a fully open AV-LLM for joint understanding and reasoning over long and complex real-world videos. AV-Flamingo combines AV-Skills, a large-scale dataset for joint cross-modal learning, a three-stage curriculum for scaling from short-context perception to long-horizon reasoning, and TAVIT, a reasoning framework that grounds intermediate thoughts to timestamps in audio and visual streams. Experiments show strong performance across diverse benchmarks, with especially notable gains on long and complex real-world videos.

AV-Flamingo still has several limitations. AV-Skills is built from public datasets and open-internet videos, which may introduce source bias and potential overlap with prior training data. In addition, reasoning over very long and highly dense videos remains challenging, especially when evidence is sparse or temporally dispersed. Finally, current benchmarks do not fully capture open-ended real-world deployment. In future work, we plan to expand AV-Skills to broader domains and more challenging long-form settings, improve long-context reasoning, and develop more realistic evaluation protocols for open audio-visual systems.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames. *arXiv preprint arXiv:2507.02001*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022. URL https://arxiv.org/abs/2104.00650.

Beijing DataTang Technology Co., Ltd. aidatatang_200zh: A free Chinese Mandarin speech corpus. https://www.openslr.org/62/, 2018. OpenSLR resource 62.

Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies, 2022. URL https://arxiv.org/abs/2210.11065.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, 2017.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: A multilingual corpus for end-to-end speech translation. *Computer speech & language*, 66:101155, 2021.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pages 3670–3674. ISCA, August 2021. doi: 10.21437/interspeech.2021-1965. URL http://dx.doi.org/10.21437/Interspeech.2021-1965.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Jiuhai Chen, Jianwei Yang, Haiping Wu, Dianqi Li, Jianfeng Gao, Tianyi Zhou, and Bin Xiao. Florence-vl: Enhancing vision-language models with generative vision encoder and depth-breadth fusion, 2024a. URL https://arxiv.org/abs/2412.04424.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers, 2024b. URL https://arxiv.org/abs/2402.19479.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos, 2024c. URL https://arxiv.org/abs/2408.10188.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024d.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024e. URL https://arxiv.org/abs/2312.14238.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Xiaowei Chi, Yatian Wang, Aosong Cheng, Pengjun Fang, Zeyue Tian, Yingqing He, Zhaoyang Liu, Xingqun Qi, Jiahao Pan, Rongyu Zhang, Mengfei Li, Ruibin Yuan, Yanbing Jiang, Wei Xue, Wenhan Luo, Qifeng Chen, Shanghang Zhang, Qifeng Liu, and Yike Guo. Mmtrail: A multimodal trailer video dataset with language and music descriptions, 2024. URL https://arxiv.org/abs/2407.20962.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, pages 52–70. Springer, 2024.

Sanjoy Chowdhury, Hanan Gani, Nishit Anand, Sayan Nag, Ruohan Gao, Mohamed Elhoseiny, Salman Khan, and Dinesh Manocha. Aurelia: Test-time reasoning distillation in audio-visual llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22899–22910, October 2025.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks, 2024. URL https://arxiv.org/abs/2305.11834.

Jiarui Fang and Shangchun Zhao. Usp: A unified sequence parallelism approach for long context generative ai. *arXiv preprint arXiv:2405.07719*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025a.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. Vita: Towards open-source interactive omni multimodal llm, 2025b. URL https://arxiv.org/abs/2408.05211.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage, 2021. URL https://arxiv.org/abs/2111.09344.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017. IEEE.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos, 2025. URL https://arxiv.org/abs/2411.19772.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, et al. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*, 2025a.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities, 2025b. URL https://arxiv.org/abs/2503.03983.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music, 2026. URL https://arxiv.org/abs/2604.10905.

John J. Godfrey and Edward Holliman. Switchboard-1 release 2 (ldc97s62). Linguistic Data Consortium, Philadelphia, 1993. Web Download.

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. Omcat: Omni context aware transformer. *arXiv preprint arXiv:2410.12109*, 2024.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, et al. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. *arXiv preprint arXiv:2603.14145*, 2026.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. Av-odyssey bench: Can your multimodal llms really understand audio-visual information?, 2024. URL https://arxiv.org/abs/2412.02611.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding, 2023a. URL https://arxiv.org/abs/2309.14405.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023b.

Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models, 2025. URL https://arxiv.org/abs/2504.02061.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE, 2024.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation*, page 198–208. Springer International Publishing, 2018. ISBN 9783319995793. doi: 10.1007/978-3-319-99579-3_21. URL http://dx.doi.org/10.1007/978-3-319-99579-3_21.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2026. URL https://arxiv.org/abs/2502.04326.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head, 2023. URL https://arxiv.org/abs/2304.12995.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions, 2024. URL https://arxiv.org/abs/2407.06358.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities, 2024. URL https://arxiv.org/abs/2402.01831.

Naveen Kumar. 93 video marketing statistics 2026 [latest data & trends]. https://www.demandsage.com/video-marketing-statistics/, April 2026. DemandSage. Published April 9, 2026. Accessed April 14, 2026.

Sonal Kumar, Prem Seetharaman, Ke Chen, Oriol Nieto, Jiaqi Su, Zhepei Wang, Rithesh Kumar, Dinesh Manocha, Nicholas J. Bryan, Zeyu Jin, and Justin Salamon. Tac: Timestamped audio captioning, 2026. URL https://arxiv.org/abs/2602.15766.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. Baichuan-audio: A unified framework for end-to-end speech interaction, 2025b. URL https://arxiv.org/abs/2502.17239.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024b.

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving llm video understanding with 16 frames per second, 2025c. URL https://arxiv.org/abs/2503.13956.

Zinuo Li, Xian Zhang, Yongxin Guo, Mohammed Bennamoun, Farid Boussaid, Girish Dwivedi, Luqi Gong, and Qiuhong Ke. Watch and listen: Understanding audio-visual-speech moments with multimodal llm. *arXiv preprint arXiv:2505.18110*, 2025d.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024. URL https://arxiv.org/abs/2312.07533.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*, pages 4095–4106, 2024.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2025a. URL https://arxiv.org/abs/2412.04468.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025b.

George Lucas. Sound is half the experience of seeing a film. *The New York Times*, May 1992. Quoted in C. Hodenfield, "In the Action With 'Star Wars' Sound," Section 2, p. 24.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix, 2025. URL https://arxiv.org/abs/2505.13032.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, 2019. URL https://arxiv.org/abs/1906.03327.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition, 2021. URL https://arxiv.org/abs/2104.02014.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description, 2016. URL https://arxiv.org/abs/1605.03705.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Ramaneswaran Selvakumar, Ashish Seth, Nishit Anand, Utkarsh Tyagi, Sonal Kumar, Sreyan Ghosh, and Dinesh Manocha. MULTIVOX: A benchmark for evaluating voice assistants for multimodal interactions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28481–28493, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1447. URL https://aclanthology.org/2025.emnlp-main.1447/.

Ramaneswaran Selvakumar, Kaousheik Jayakumar, S Sakshi, Sreyan Ghosh, Ruohan Gao, and Dinesh Manocha. Do audio-visual large language models really see and hear? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 5892–5902, June 2026.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. EGOILLUSION: Benchmarking hallucinations in egocentric video understanding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28461–28480, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1446. URL https://aclanthology.org/2025.emnlp-main.1446/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.5-omni: Scaling up, toward native omni-modal agi, March 2026b. URL https://qwen.ai/blog?id=qwen3.5-omni.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos, 2018. URL https://arxiv.org/abs/1803.08842.

Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013. IEEE.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*, 2020.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark, 2026. URL https://arxiv.org/abs/2506.04779.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024b. URL https://arxiv.org/abs/2307.06942.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions, 2022. URL https://arxiv.org/abs/2111.10337.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale, 2023. URL https://arxiv.org/abs/2309.13952.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In *Proc. ICASSP*. IEEE, 2022.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering, 2019. URL https://arxiv.org/abs/1906.02467.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553, 2023.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Zitang Zhou, Ke Mei, Yu Lu, Tianyi Wang, and Fengyun Rao. Harmonyset: A comprehensive dataset for understanding video-music semantic alignment and temporal synchronization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3152–3162, 2025.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2026. URL https://arxiv.org/abs/2505.17862.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A. Audio-Visual Flamingo Training Datasets

Table 2 summarizes all datasets used to train Audio-Visual Flamingo, including total hours, number of training pairs, and the number of epochs (passes over the dataset) used at each training stage. Similar to (Ye et al., 2025; Goel et al., 2025), we convert all foundational datasets (captioning, classification, etc.) into QA formats, using the same set of prompts for each task.

AV-Safety QA. In addition to the skill-focused subsets, Table 2 includes AV-Safety QA, a set of 92K question–answer pairs over 536 hours of video used during long-context SFT. This subset pairs audio-visual inputs with questions that probe unsafe, harmful, or policy-violating requests (e.g., requests to identify private individuals, generate harassing descriptions, or extract sensitive information from the audio-visual content), together with target responses that safely decline or redirect while remaining helpful on the benign portion of the request. Its purpose is to preserve safety alignment while the model acquires long-context audio-visual capabilities.

Table 2: List of pre-training and fine-tuning datasets together with their training composition.

Dataset	Hours	Num. Instances	AVF-Short-SFT	AVF-Long-SFT	CoT-Training
AV-Skills-Short (Ours)	100k hrs	3.8M	1.0	0.5	-
Long AV Captioning (Ours)	51k hrs	1.2M	-	2.0	-
Long Comparative AVQA (Ours)	8.3K hrs	169K	-	2.0	-
Long Context AVQA (Ours)	8.3K hrs	175K	-	2.0	-
Long Subscene AVQA (Ours)	6.8K hrs	157K	-	2.0	-
Long Counting AVQA (Ours)	6.2K hrs	136K	-	2.0	-
Long Needle AVQA (Ours)	8K hrs	169K	-	2.0	-
Long Referring AVQA (Ours)	7.2K hrs	160K	-	2.0	-
Long Topic AVQA (Ours)	7.3K hrs	172K	-	2.0	-
Long Holistic Reas. AVQA (Ours)	8.9K hrs	180K	-	2.0	-
Long Event Alignment AVQA (Ours)	7.9K hrs	160K	-	2.0	-
Long Event Sequence AVQA (Ours)	7.4K hrs	153K	-	2.0	-
Long Inference AVQA (Ours)	4.5K hrs	63K	-	2.0	-
Long Temporal AVQA (Ours)	9K hrs	200K	-	2.0	-
AV-Safety QA (Ours)	536 hrs	92K	-	2.0	-
AF3-training mix (Goel et al., 2025)	308k hrs	48.6M	1.0	-	-
AudioSkills-XL (Goel et al., 2025)	-	9700K	1.0	0.5	-
MF-Skills (Ghosh et al., 2025a)	-	3M	1.0	0.5	-
Image-text (Ye et al., 2025)	-	3M	0.5	-	-
Video-text (Ye et al., 2025)	-	3M	0.5	-	-
CHiME (Vincent et al., 2013)	342 hrs	30k	1.0	1.0	-
AliMeeting (Yu et al., 2022)	118.75 hrs	387k	1.0	1.0	-
EMILIA (He et al., 2024)	5000 hrs	1.7M	1.0	1.0	-
MuST-C (Cattoni et al., 2021)	500 hrs	245k	1.0	1.0	-
CoVoST (Wang et al., 2020)	2880 hrs	5M	1.0	1.0	-
Aidatatang (Beijing DataTang Technology Co., Ltd., 2018)	139.39 hrs	165k	1.0	1.0	-
Aishell-1 (Bu et al., 2017)	150.85 hrs	120k	1.0	1.0	-
Multi-talker Switchboard (Godfrey and Holliman, 1993)	109.9 hrs	76.6K	1.0	1.0	-
Audio-Think-Time (Ours)	3954 hrs	43K	-	2.0	-
Video-Think-Time (Ours)	1238 hrs	22K	-	2.0	-
AV-Think (Ours)	1500 hrs	24K	-	-	2.0

B. Dataset Examples and Category Distribution

Figure 4 shows the category distribution of the 15 major video categories used for training. Figure 5 and Table 3 provide representative AV-Skills-Long and AV-Skills-Short examples.

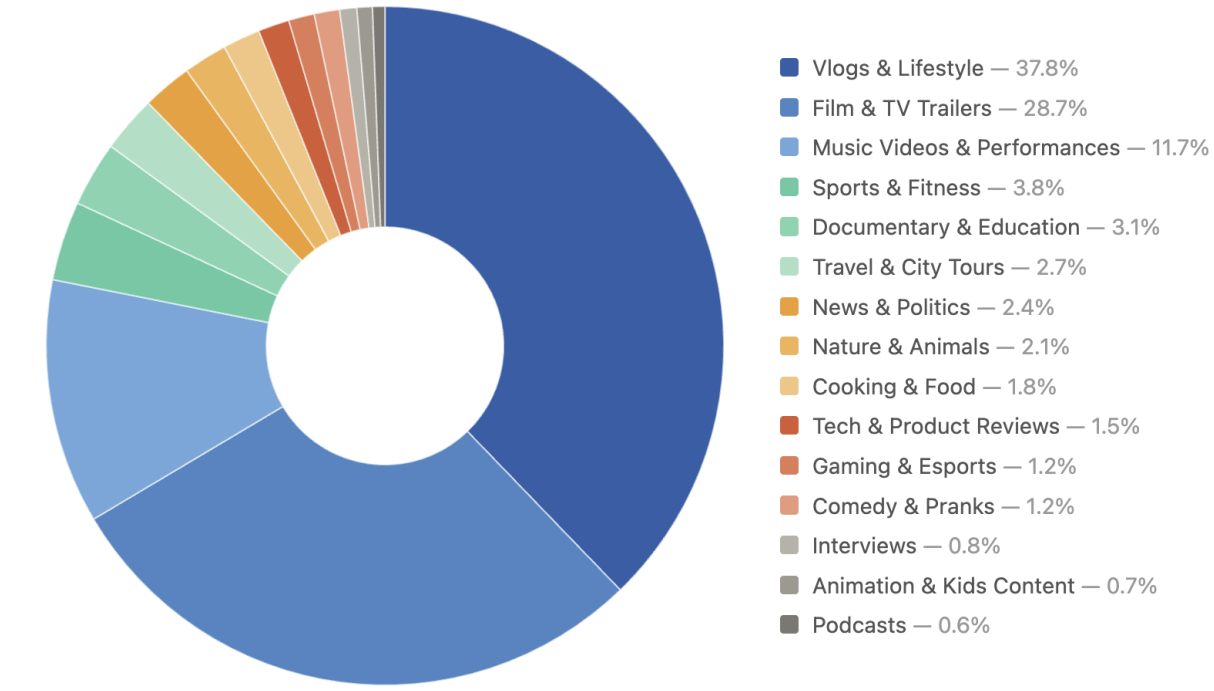

Figure 4: Category distribution of the videos used in the training datasets.

C. Audio-Visual Flamingo Training Details

In this section, we present the training settings of our model across all stages, each with specific configurations. Details are in Table 4; a compact stage-level recipe is summarized in Table 5.

D. Additional Ablation Study

Effect of AV-Skills-Short and AV-Skills-Long. This section isolates the contribution of the curated AV-Skills data across the training curriculum. We compare three models: (i) **OmniVinci**, the base checkpoint from which AVF is initialized; (ii) **AVF-Stage1**, our model after pre-training (Stage 1 + Stage 2) on AV-Skills-Short alongside unimodal data; and (iii) **AVF-Instruct**, the final model obtained after mid-training on AV-Skills-Long.

As shown in Table 6, training on AV-Skills-Short (AVF-Stage1) yields consistent improvements over OmniVinci across both audio-visual and video-only benchmarks, confirming that curated AV-Skills data injects cross-modal reasoning that cannot be captured by unimodal supervision alone. Adding AV-Skills-Long during mid-training (AVF-Instruct) further improves reasoning over extended temporal contexts and tighter audio-visual grounding.

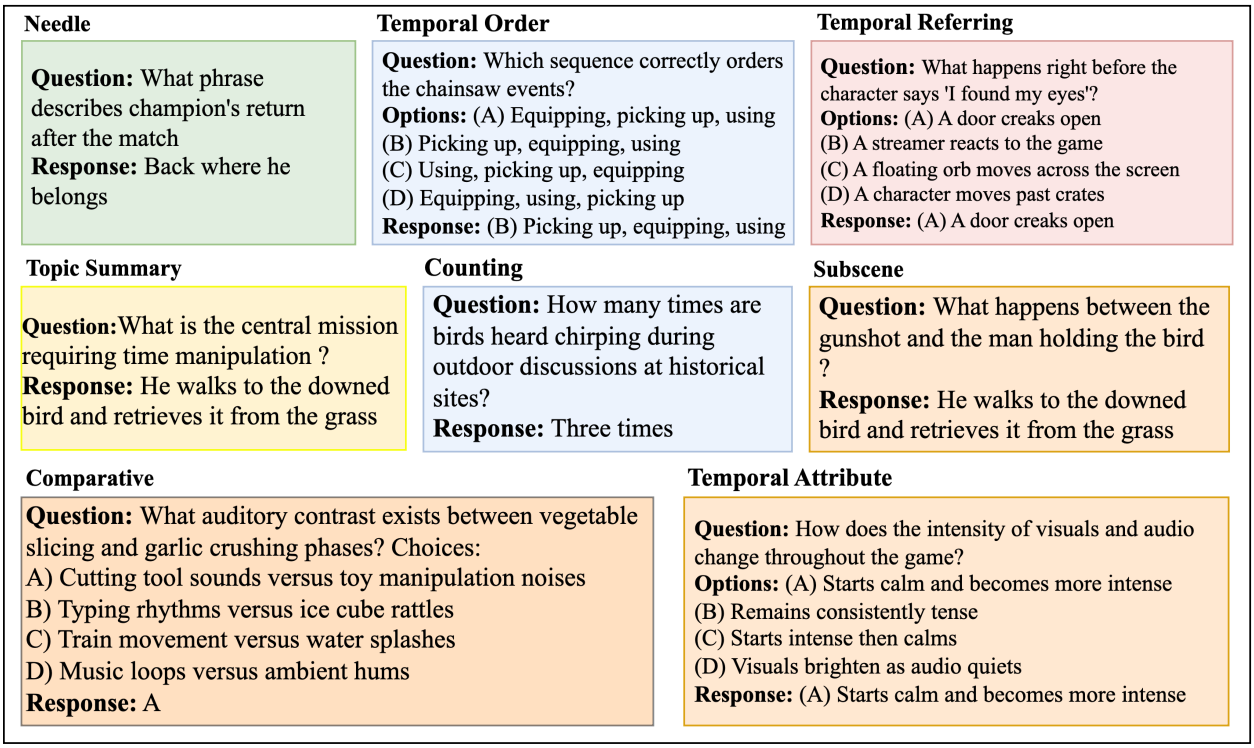

Figure 5: **Additional QA pair examples from AV-Skills-Long.** Representative samples spanning eight long-form skill categories used during mid-training: Needle-in-the-Haystack, Temporal Order, Temporal Referring, Topic Summary, Counting, Sub-scene, Comparative, and Temporal Attribute. Examples include both open-ended and multiple-choice formats.

E. GRPO-based RL training Details

GRPO for Audio-Visual Understanding and Reasoning. Building on recent advances in Group Relative Policy Optimization (GRPO), we adopt the standard GRPO framework to train AV-Flamingo. GRPO removes the need for an explicit value function and instead estimates the advantage using the average reward of multiple sampled outputs for the same input. For each input question or caption prompt q, the policy model generates a group of candidate outputs $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$, along with their corresponding rewards $\{r_1, r_2, \ldots, r_G\}$ computed using rule-based reward functions such as format, answer accuracy, and structured reward matching. The policy π_θ is then optimized using the GRPO objective:

$$\mathcal{J}(\theta)=\mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i,\ \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)},1-\epsilon,1+\epsilon\right)A_i\right)-\beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right)\right], \tag{1}$$

where ϵ is the clipping range for the importance sampling ratio, β is the coefficient for the KL regularization term that keeps the learned policy close to the reference policy, and G denotes the group size, i.e., the number of sampled candidate outputs per input, which we set to 5 in our experiments. To stabilize optimization, rewards are normalized within each group to compute the advantage:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}.$$

We next describe the reward functions used for AV-Flamingo.

Format Reward. To encourage well-structured outputs, we use a regex-based format reward similar to prior reasoning work. The model is required to produce intermediate reasoning within `<think> </think>` tags,

followed by the final response within <answer> </answer> tags. Outputs that strictly follow this format receive a reward of 1, and 0 otherwise. This binary reward encourages the model to consistently generate structured reasoning traces and final answers.

Accuracy Reward. For question-answering instances, we use an *accuracy reward* based on the final answer extracted from the <answer> </answer> tags. Given a question and its ground-truth answer, the reward is computed by matching the normalized predicted answer with the normalized reference answer. This reward directly encourages correct final responses for audio-visual QA.

Structured Reward for Open-Ended Captions and Responses. For open-ended captioning and free-form response generation, strict string matching is not suitable due to the long-form and semantically diverse nature of valid outputs. To address this, we design a structured reward that first parses both the reference output and the generated output into structured JSON using an LLM, and then compares the resulting fields. Specifically, given a ground-truth caption or response, we prompt an LLM to extract salient structured metadata such as scene type, participants, environment, topic, events, emotional tone, and key conclusions. For example, for a podcast video discussing artificial intelligence, the parsed JSON may contain fields such as:

{"**Scene**": podcast studio, "**Participants**": two individuals, speaker in light blue shirt; listener in black blazer, "**Environment**": professional studio, beige walls, microphones, "**Topic**": artificial intelligence, "**Main Discussion**": limitations of neural networks for human-level intelligence, "**Speaker 1 View**": skeptical of scaling neural networks alone, "**Speaker 2 View**": emphasizes physical-world understanding for reasoning and planning, "**Tone**": analytical, technical, collaborative, "**Key Themes**": human-level AI, neural network limitations, reasoning, planning, collaboration in research community }

We then parse the model-generated caption or response into the same schema and compute a structured matching reward by measuring overlap between the predicted and reference fields. The final reward is obtained by averaging field-level matches across categories, allowing semantically rich open-ended outputs to be evaluated in a structured and consistent way.

The overall reward function in GRPO combines the format reward with the accuracy reward for question-answering data, and the format reward with the structured reward for open-ended captioning and response-generation data. This enables AV-Flamingo to improve both answer correctness and the quality of long-form grounded audio-visual generations.

F. AV-Think and TAVIT Details

This section documents the AV-Think generation constraints in Section F.1 and representative Temporal Audio-Visual Interleaved Chain-of-Thought (TAVIT) examples in Section F.2.

F.1 AV-Think Generation Constraints

For each long-form video, the generator receives timestamped audio segments, including speech transcriptions and non-speech event descriptions, together with timestamped visual captions. It produces open-ended QA-reasoning triplets, multiple-choice variants with plausible distractors, and an integrated audio-visual caption. We explicitly require questions to depend on both modalities, demand multi-hop temporal reasoning rather than surface-level transcription, and ground each intermediate reasoning step with <t>...</t> timestamp tags; questions answerable from audio alone or vision alone are excluded.

F.2 TAVIT Examples

Table 7 shows representative TAVIT examples with the question, grounded thinking chain, and ground-truth response. Table 8 provides five additional TAVIT samples pairing the merged audio-visual caption (with source YouTube link) with the corresponding chain-of-thought question and ground-truth response.

Table 8: Five overlapping samples from TAVIT: merged audio-visual caption (with YouTube link), chain-of-thought question, and ground-truth (GT) response. All entries are truncated; full text available in the dataset.

[YouTubeID] Caption	**Question**	**GT Response**
[youtube.com/watch?v=-0EraoJmdhk] A bartender shakes a cocktail shaker containing a pink liquid. The liquid is poured through a fine mesh strainer into a martini glass. A close-up shows the finished pink cocktail garnished with an orange peel; text overlays display "Cosmopolitan", its Hindi transliteration, and "DRINK IT EASY 2.0". The bartender sets down an empty shaker, removes its lid, then measures and pours vodka, lemon juice, cranberry juice, and triple sec with bilingual text callouts for each ingredient. Ice cubes are added, the shaker sealed and shaken vigorously. The cocktail is strained into a frosted martini glass and finished with a twisted orange peel garnish before the final beauty shot.	How do the changes in the lo-fi hip-hop soundtrack support the visual progression from ingredient preparation to the final presentation and branding of the Cosmopolitan cocktail, and what does this reveal about the intended mood and viewing experience of the video as a whole?	The soundtrack and visuals are tightly coordinated so that the music's evolving texture mirrors the cocktail's journey from raw ingredients to a polished, branded product, creating a relaxed, tutorial-meets-aesthetic experience rather than a purely instructional one. At the very beginning, the ambient pad without drums (0–2 s) plays over the first glimpse of shaking and pouring and the hero close-up of the finished Cosmopolitan with multilingual overlays (0–8 s), framing the drink as aspirational, not just a recipe. When the main synth melody and full beat enter (2–18 s), the visuals shift into step-by-step recipe work: measuring and pouring each spirit with bilingual callouts, adding ice, sealing, and shaking. The steady, looping groove makes these repetitive actions feel smooth and unhurried. A slightly thinned bridge (27–35 s) coincides with the refinement phase: the frosted glass, straining, and the orange-peel garnish. Finally, during the gentle fade-out (35–44 s), the camera lingers on the completed cocktail before cutting to a branded outro screen with social-media handles and a website. [. . .]
Continued on next page		

Table 8 – continued from previous page

[YouTubeID] Caption	Question	GT Response
[youtube.com/watch?v=-0ItZ8fGUNc] A close-up of a dark bowl of rich, reddish-brown stew garnished with coriander; text reads "Pork Sorpotel". A dark pan shows whole cloves, black peppercorns ("4–5"), green cardamoms ("3"), half a cinnamon stick, cumin seeds, and 3–4 dried red chillies added one by one. The spices are shaken on an induction cooktop, transferred into warm water to soak for 30–40 minutes, then blended into a liquid paste in a Wonderchef blender. Oil is poured into a black pot; boiled, diced pork and pork liver are sautéed until golden brown and set aside. Ginger, garlic, green chillies, onions, and salt are sautéed until the onions brown; the red spice paste and pork stock are added and stirred. Sugar, tamarind pulp, vinegar, and salt are added; the pot is covered and simmered for 15–20 minutes before the finished dish is ladled into a serving bowl and garnished with fresh coriander.	How does the combination of the steadily building EDM soundtrack and the evolving visuals of the Pork Sorpotel recipe work together to guide the viewer from raw spices and meat through to the final plated dish, and what does this reveal about how the video balances methodical instruction with an energetic, appetizing mood?	The video uses the EDM track as a continuous emotional backbone that parallels the visual progression from raw ingredients to finished Pork Sorpotel, turning a methodical recipe into something that feels energetic and appetizing rather than slow or tedious. At the start, the music is in a steady, moderate groove but with a light arrangement (0–2 s), matching the clean title close-up of the garnished dish. As whole spices are toasted and added one by one (3–16 s), the beat maintains energy, making the repetitive individual shots feel purposeful. When the pork and liver are sautéed to golden brown (38–49 s), the EDM transitions to a more energetic section, lifting the energy as the recipe moves from preparation into active cooking. A brief breakdown (75–91 s) accompanies the final seasoning steps—sugar, tamarind, vinegar—signalling an imminent climax. The full groove returns as the pot simmers (91–104 s), then fades as the dish is plated and garnished, completing the parallel arcs of music and recipe. [. . .]
[youtube.com/watch?v=–ccsQMua_s] The *Fox & Friends* logo appears on an orange-and-blue graphic. Biden is shown at the Resolute Desk in the Oval Office; on-screen text reads "BIDEN DELIVERS FAREWELL ADDRESS FROM OVAL OFFICE". Peter Doocy reports live from a snowy White House lawn. Biden warns of "a dangerous concentration of power in the hands of a very few ultra-wealthy people" and "an oligarchy taking shape in America". Biden invokes Eisenhower's military-industrial complex, coining "tech-industrial complex". A ghost image of Biden is overlaid on a shot of his black-clad family seated alongside Kamala Harris and Doug Emhoff. CNN poll graphics appear: Biden job approval 36%, with 61% calling his presidency a failure. Studio hosts discuss Biden's legacy while a split screen shows him alongside the *Fox & Friends* panel.	How does the segment use both President Biden's own farewell warnings and the studio's visual and conversational framing—especially the ghostly overlay of him on his black-clad family, the polling graphics, and the contrast with Trump and tech leaders—to recast his stated concerns about oligarchy and a tech-industrial complex into a narrative of a failed, bitter presidency out of step with the country's future?	The segment systematically reframes Biden's farewell warnings from principled concerns into evidence of a failed, embittered presidency by weaving together his own words with highly curated visuals and studio commentary. Biden is visually anchored at the Resolute Desk as a formal, traditional president (1–15 s, 42–60 s), while the audio has him warning about concentrated wealth and a tech-industrial complex (48–60 s, 128–145 s). Rather than engaging with the substance, the hosts pivot immediately to mockery: Steve Doocy jokes about "teleprompterarchy" (130–147 s), and the ghostly overlay of Biden on his black-clad family (152–157 s) visually associates his warnings with funereal gloom rather than civic urgency. The CNN poll graphic showing 36% approval and 61% "failure" (374–388 s) is then presented as the electorate's verdict on those same warnings. The contrast is completed by hosts praising tech CEOs for the economic growth Biden's presidency benefited from, inverting his oligarchy framing into ingratitude. [. . .]

Continued on next page

Table 8 – continued from previous page

[YouTubeID] Caption	Question	GT Response
[youtube.com/watch?v=-0YyxReol5E] A close-up of a South African rugby player in a green, black, and yellow jersey with a white headband, eyes closed, as rain falls. The camera moves to his back showing the "SOUTH AFRICA RUGBY" crest and "RUGBY WORLD CUP FRANCE 2023" text. A second player sings with eyes open; the camera pans a line of players singing with serious expressions in a packed stadium. Spectators in a box seat watch and photograph the ceremony. Close-ups of a light-skinned and a bald dark-skinned player, both sweating and singing with eyes closed. Fans in green and yellow jerseys sing looking upward; a player with long blonde hair sings with eyes closed for an extended shot. All 15 players stand arms-around-shoulders, joined by a young boy in a black-and-white kit, before turning and walking away as the crowd erupts.	How do the lyrics and musical build-up of the choral anthem combine with the behavior and arrangement of the South African players, officials, and supporters to show that this moment is not just a routine pre-match ritual but a deeply spiritual and nationally unifying ceremony?	The moment becomes clearly more than a routine pre-match ritual because the religious, sacrificial language and swelling structure of the anthem are mirrored by the intensity, posture, and inclusiveness of everyone shown in the stadium. The choir sings of a people beloved by God—lines such as "Para que seas su pueblo, y su nación" frame the singers as a chosen community under divine care (6–16 s). Repeated invocations of "*Oh, Señor*" and references to "Su sangre" (37–57 s) give the anthem a liturgical character, while the music grows into a full orchestral and choral crescendo (33–84 s). Visually, the players' closed eyes, sweat, and upward tilted heads mirror the posture of private prayer (0–84 s). The inclusion of officials and fans singing in the stands (17–33 s) extends the ceremony into a national act. Most powerfully, the young boy joining the players' line (77–83 s) signals intergenerational transmission of identity, and the crowd's eruption the instant the anthem ends (84–86 s) confirms that the entire stadium participated in a shared spiritual moment. [. . .]

G. Prompt Templates

For AV-Skills-Long datasets, we curate synthetic QA pairs and options for questions from various categories using GPT. The category-specific AV-Skills-Long prompt templates are shown in Figures 6–19. Figure 20 shows the full prompt used to construct the question–answer–reasoning triplets in AV-Think.

H. Assets and Licenses

Table 9 lists the datasets and model checkpoints used in this paper, along with their associated licenses or distribution terms. For web-video datasets distributed as metadata, captions, or URLs, the listed terms refer to the released dataset artifacts; underlying source media remain subject to the original platform and copyright-holder terms. The Aidatatang corpus was retracted by its distributor after our training runs were completed; we retain it in Table 9 for transparency about the data used to train the released checkpoints, and it is not redistributed as part of any AVF release artifact.

I. Broader Impacts

AV-Flamingo advances open audio-visual intelligence, with positive societal applications in accessibility (e.g., automated audio description for the visually impaired), education (e.g., lecture and documentary understanding), and media analysis (e.g., automated captioning and content moderation). By releasing the model, training code, and AV-Skills dataset openly, we also lower the barrier for researchers without access to large-scale proprietary resources, supporting a more inclusive research ecosystem.

However, models capable of joint audio-visual understanding carry potential risks. They could be misused to automate surveillance, assist in the generation or verification of deepfakes, or enable multimodal disinformation at scale by jointly analyzing and synthesizing audio and visual content. To mitigate these risks, we release AV-Flamingo and AV-Skills for non-commercial research use only, under licenses that explicitly prohibit harmful applications. In addition, we include AV-Safety QA (Section A) during training to encourage safe refusal behavior on harmful audio-visual requests. We encourage the community to continue developing safeguards, evaluation protocols, and governance frameworks for open audio-visual systems.

Table 3: Examples of AV-Skills-Short training data

Category	**Prompt**	**Response**
Causal Reasoning	Why did the person cut the cake roll after attempting to roll it?	Because when rolling the cake, the cream overflowed and since others disagreed with eating the overflowed portion, the person decided to cut the cake instead of continuing to roll it.
Spatial Perception	During the scene when room with wooden floor, large TV, person on floor with dog, where is the small fluffy dog named Lala located relative to the person wearing star-patterned pajamas playing with dog? A) Lala is in front of the person B) Lala is under the person's legs C) Lala is on the person's lap D) Lala is standing beside the person	C) Lala is on the person's lap
Emotion Change	What is emotion / mood change of the woman with long dark hair, glasses, and a sweater who sits at the right side of the panel during the panel discussion on child-abuse reporting when the question appears and a comic-style bubble with 'wae?' (why?) pops up? A) surprised B) confused C) amused D) curious	A) surprised
Hallucination	Does Mickey Mouse wear a red hat in the video? A. I am not sure. B. No, Mickey does not wear a red hat. C. None of the above. D. Yes, Mickey is wearing a red hat.	B. No, Mickey does not wear a red hat.
Audio Counting	How many times does a sniffle occur in the audio? A. 0 B. 2 C. 3 D. 1	D. 1
Video Counting	How many times does the hand adjust a pressure switch? A. Five B. Six C. Four D. Three	C. Four
Temporal Reasoning	Arrange the below list of audio-visual events in the order of their occurrence in the video. Use arrow '→' to indicate the order of the events. A) Speaker discussing surface preparation while gloved hands open a white tube of 'CHASER', placed next to an open can of Bondo filler. B) Speaker mentions applying filler like Bondo while a person applies purple filler to a hairdryer. C) Speaker mentions '400 grit sandpaper' as text appears on screen while a person holds a sanding block. D) A person wearing gloves mixes Bondo and hardener as on-screen text explains the process.	A → D → B → C
Relation Reasoning	During the scene when room with aquarium and instrumental electronic music, where is the man near the aquarium positioned relative to the man in the hoodie? A) The man in the hoodie is behind the man in the t-shirt B) The man in the hoodie is in front of the man in the t-shirt C) The man in the hoodie is to the right of the man in the t-shirt D) The man in the hoodie is to the left of the man in the t-shirt	C) The man in the hoodie is to the right of the man in the t-shirt

Table 4: Training settings across stages.

Settings	**Pre-training**	**Mid-Training**	**Post-Training**
global batch size	128	128	64
learning rate	1e-5	1e-5	2e-5
learning schedule		Cosine decay	
warm up ratio		0.03	
weight decay		0.0	
epoch	1	1	2
bf16	✓	✓	✓
grad accumulate		8	
Parallelism	Zero-3	Zero-3 + SP	Zero-3 + SP
GPUs		512×H100	

Table 5: Compact training recipe for AV-Flamingo. Dataset-level mixing ratios are provided in Table 2.

Stage	**Main data mixture**	**Max input / context**	**Epochs**	**Batch**	**LR**	**Compute / parallelism**
Short-context SFT	AV-Skills-Short + AF3/AudioSkills/MF + image/video/speech	5 min / 16K	1	128	1e-5	512×H100; ZeRO-3
Long-context SFT	AV-Skills-Long + downsampled short/audio/music/speech + think-time data	15 min / 32K	1	128	1e-5	512×H100; ZeRO-3 + SP
CoT post-training	AV-Think reasoning triplets with SFT + GRPO	15 min / 32K	2	64	2e-5	512×H100; ZeRO-3 + SP

Table 6: Ablation study on the effect of AV-Skills-Short and AV-Skills-Long. AVF-Stage1 isolates the contribution of AV-Skills-Short during pre-training, while AVF-Instruct adds AV-Skills-Long during mid-training. Higher is better.

Model	**DailyOmni**	**WorldSense**	**VideoMME**
OmniVinci	66.5	48.2	67.3
AVF-Stage1	69.5	48.5	68.5
AVF-Instruct	**72.4**	**50.3**	**70.7**

Table 7: Examples of Temporal Audio-Visual Interleaved Chain-of-Thought (TAVIT) data

Question [YouTube URL]	**Thinking Chain**	**Ground Truth Response**
How does the video use both the upbeat music, the brief spoken subscription message, and the evolving cooking visuals to guide the viewer from learning how to make mushroom tarts to being nudged toward engaging further with the channel, and what specific moments show this shift in purpose? [https://www.youtube.com/watch?v=2tPqC6quZQ]	I need to trace how the video's purpose evolves and how both sound and visuals participate in that change. The clip opens with a close-up of a finished mushroom tart and a graphic that labels it "MUSHROOM TART" (0–3s), then immediately shows raw puff pastry with a text overlay specifying "200 grams puff pastry dough" (3–5s). That tells me the initial goal is instructional. The visuals proceed step by step through the cooking process: rolling dough (5–9s), cutting rounds (9–15s), brushing with milk and baking (17–22s), sautéing garlic and mushrooms with seasonings (22–37s), filling baked pastry and topping with mozzarella (42–49s), and garnishing with thyme (49–51s). Upbeat funk/disco music plays throughout with no speech during cooking steps. The turning point is when "enjoy!" appears (51–53s). Immediately after, visuals change to an outro screen with video thumbnails, a YouTube Subscribe button, and social media links for Sanjeev Kapoor (53–63s). The male speaker's line—"If you like the videos that we share on YouTube, do subscribe to us on our YouTube channel"—occurs at (54–59s), explicitly promotional. The first phase combines cooking visuals with music to teach the recipe (0–51s); the second phase shifts to channel branding and a spoken call to subscribe (53–63s).	The video first focuses on teaching the viewer how to make mushroom tarts, then gradually pivots to encouraging ongoing engagement with the channel. During the cooking portion, there is no speech—only a steady, upbeat funk/disco track (0–40s). Visually, this phase walks step-by-step through the recipe: the finished tart and title overlay (0–3s), rolling and cutting puff pastry (3–15s), baking (17–22s), sautéing garlic and mushrooms (24–37s), and filling, topping, rebaking, and garnishing (42–51s). The purpose shifts once "enjoy!" appears (51–53s). The visuals change to an outro screen with thumbnails, a Subscribe button, and social media links for Sanjeev Kapoor (53–63s). A friendly male voice says, "If you like the videos that we share on YouTube, do subscribe to us on our YouTube channel" (54–59s). The cooking visuals plus music build trust, and once that value is delivered, the audio shifts to a call to action while the visuals shift to subscribe prompts and social links.
Across the entire highlight, what is the main reason the sequence of play is presented as a showcase of rugby sevens excellence, rather than just a routine attacking move? (A) Commentators focus on crowd chants over players' actions (B) Clip shows scoreboard and time remaining (C) Sequence combines ecstatic live call and crowd with slow-motion replays and detailed analysis of quick taps, dummy passes, and footwork (D) Visuals center on referee decisions (E) Player shown arguing with teammates (F) Focus on scrums and mauls (G) Clip shows pre-match warm-ups (H) Video emphasizes sponsor logos (I) Commentators describe a defensive masterclass [https://www.youtube.com/watch?v=2hY4Ap7xkA]	I need the option that best captures why the whole clip frames this as a showcase of sevens excellence. From the audio: loud crowd cheering (4–24s), an excited commentator shouting about "some footwork", "out to Watton, he throws the dummy", and "That's why we love the game of seven so much" (10–24s). Then analytical commentary: "Quick tap from Lindsay, off one foot", "He's electric and then the one-handed pass", "The show and go, the English footwork paying off hours at the Lensbury Club" (25–46s). From the visuals: live aerial/close-up shots of a white-shirted player breaking through red defenders (4–17s), then slow-motion replays showing number 9 tapping quickly (28–30s), number 11 making a one-handed pass (32–35s), and number 1 repeatedly stepping and dodging (30–47s). This combination—live emotional high plus slow-motion technical breakdown—is exactly option (C). Other options eliminated: (A) crowd isn't the focus; (B) no scoreboard shown; (D) no refereeing debate; (E) hands-on-head is not anger; (F) this is open play; (G) Lensbury Club only mentioned verbally; (H) branding is just framing; (I) attack succeeds.	(C) Because the sequence combines an ecstatic live call and roaring crowd with slow-motion replays and detailed analysis of quick taps, dummy passes, and practiced footwork, highlighting both the excitement and the underlying skill.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and compose question-answer pairs of type ****comparative**** that require advanced understanding of the video (including the associated audio) and complex reasoning over it to answer correctly. Comparative questions should ask about **key differences or similarities between two distinct audio-visual segments or presentations**. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. Segments are chronological; missing info appears as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA must highlight a significant contrast or similarity observable only by integrating both modalities.
2. The question should be abstract and conceal obvious hints.
3. Neither question nor answer may exceed 20-30 words.
4. Do NOT use phrases like "according to the details"; write as if you observed and heard the video yourself.
5. Audio captions may be noisy; verify them with visuals.
6. Ensure both audio and visual evidence are needed to answer.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Example event-sequence QAs:
Question: "What is the primary distinction between Dr. Rajani's video analysis and Clare's presentation regarding the Revox product?"
Options: "**A.** One explains its mechanism as a Botox alternative, the other details cost and availability", "**B.** One demonstrates application steps, the other chemical composition", "**C.** One uses clinical before/after visuals, the third-party product endorsements", "**D.** One features medical professional testimonials, the other consumer demographics"
Answer:"**A.** One explains its mechanism as a Botox alternative, the other details cost and availability"

Generate up to 10 options.
Only return a JSON with the keys like this: {"Question": "...","Choices": { "A": "...", "B": "...", "C": "...", "D": "..." },"Answer": "X"}. If this question type is impossible, output "None". Focus on quality. Here is the input information:

Figure 6: **Prompt for Comparative Reasoning QA generation.** Instructs the LLM to produce a multiple-choice question (with up to 10 options) that hinges on a key contrast or similarity between two distinct audio-visual segments in a long video, requiring both modalities to answer.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of type **"context understanding"** that require advanced understanding of the video (including the associated audio) and complex reasoning over it to answer correctly. Context-understanding questions should ask about **broader setting, background elements, or situational context that emerge only by integrating audio and visuals**. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. Segments are chronological; missing info appears as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA must rely on holistic scene context rather than isolated actions.
2. The question should be abstract and conceal obvious hints.
3. Neither question nor answer may exceed 20-30 words.
4. Do NOT use phrases like "according to the details"; write as if you observed the video yourself.
5. Audio captions may be noisy; verify them with visuals.
6. Ensure both audio and visual evidence are needed to answer.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Example event-sequence QAs:
Question: "What specific visual elements were present when Sara mentioned 'email Anderson'?"
Options: "A. Indoor office with binder and whiteboard", "B. Dirt field with construction equipment and trees", "C. Excavator actively lifting dirt", "D. Yellow bulldozer parked in foreground"
Answer: "B. Dirt field with construction equipment and trees"

Generate up to 10 options.
Only return a JSON with the keys like this: {"Question": "...","Choices": { "A": "...", "B": "...", "C": "...", "D": "..." },"Answer": "X"}. If this question type is impossible, output "None". Focus on quality. Here is the input information:

Figure 7: **Prompt for Context Understanding QA generation.** Instructs the LLM to produce a multiple-choice question that depends on the broader situational setting, background, or scene conditions inferable only by jointly reasoning over visuals, sounds, and speech.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of type **counting** that requires understanding of the video (including the associated audio) and the skill to count. Counting questions should ask about the count of a particular type of audio-visual event across the video -- where one requires to watch the entire video and understand its content. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA should either require understanding an important part in the audio-visual scene **or** require an understanding of the entire video to answer.
2. The question should be abstract and not reveal too much about the video or hints that can help answer the question.
3. Neither the prompt nor the response should be more than 20-30 words.
4. Do NOT use phrases like "according to the details" in both the questions and answers; you should ask and answer as if you were seeing the video and hearing the corresponding audio by yourself.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair.**
6. Most importantly, generate the questions and answers from an audio-visual perspective, i.e., an agent should be able to answer seeing both visuals and sounds. Do not ask questions that require only one.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

An example question-answer pair is:
Question: "Across the video, how many separate times do suited presenters speak beside giant Prysm screens?",
Answer: "Six times." - use your reasoning abilities to generate better questions about audio-visual events (one that has both audio and visual cues).

Only return a JSON with the Question and the Answer with the keys Question and the Answer respectively and nothing else. If a question type is not possible, don't output and only output "None". Focus on quality. Here is the input information:

Figure 8: **Prompt for Counting QA generation.** Instructs the LLM to produce a question requiring the model to count occurrences of a specific audio-visual event across the full video, verifying noisy audio captions against visual cues.

You are a helpful AI assistant. You need to act as an audio-visual caption generator for long videos that requires advanced understanding of the entire video (including the associated audio). My final objective is to train an audio-visual understanding agent with these captions to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate 1 caption with the following requirements:
1. The caption should cover the entire video.
2. The caption should include as much details as possible but not include redundant ones.
3. The caption should not be more than 1000 words.
4. Do NOT use phrases like "according to the details" in the caption.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair.**
6. Most importantly, generate the captions from an audio-visual perspective, i.e., an agent should be able to generate seeing both visuals and sounds. Do not generate captions that focus on only one.
7. **The caption should have all transcripts in-line**.
8. Do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments as one full video and your output should be such that the model I am training has no access to the segment-wise captions and is considering the the video as a whole. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that audio captions and whisper transcripts may have hallucinations. For audio captions, always use visual to cross-check, if it's not a background sound. For whisper transcripts, transcripts that are extra small and a few words (e.g., "Oh", "uh", "Okay", etc) may be a hallucination. You can also safely ignore such small utterances in the caption.

An example caption is as follows:
A blue title card flashes: "Standard Process — Transforming Lives Through Prysm." A calm male voice over light strings explains, "Standard Process is a clinical-nutrition company that's been in business for 90 years." A tractor tills lush rows; the narrator continues as cameras sweep the Wisconsin HQ, water tower, and test plots: "… headquartered in Wisconsin and has its research and innovation center in North Carolina. Our whole philosophy is whole food and effective nutrition and it really starts with the soil." Inside the plant, amber bottles rattle along conveyors while he asserts, "We're making sure that we can control from seed to supplement and crop to clinic and plant to patient." Gears fade; in a dark auditorium soft ambient strings rise around a towering 190-inch Prysm wall. A suited presenter gestures through razor-sharp full-body anatomy and a rotating 3-D brain, noting, "Scientific information and scientific insight… we spent a lot of money on innovative new research findings, but all that would mean nothing if we couldn't be effective at educating people and create calls to action and follow through." As panoramic lakes, maps, and harvest footage glide across the screen, he challenges practitioners: "How do you get them comfortable using that information sooner… and make it so that it's engaging and fun?" He recalls, "Two years ago at a trade show, we looked at audiovisual technologies and Prysm was one that frankly wowed us." Classical D-minor strings swell while attendees tap the display; he adds, "It allows the participants to be part of the dialogue… sitting in front of a 190-inch screen and seeing a field—you don't have to get your boots dirty." Cutaways show sparkling lenses and collaborators smiling. Voices praise, "The second behavior that's consistently seen is they wanna go up and touch it…

Figure 9: **Prompt for Detailed Captioning with Transcripts. (Part 1 of 2)**

```
they actively learn faster and apply it faster." Another speaker lists tools: "You see
all the collaboration tools and the ability to use three-dimensional imaging and bring
things together—data live on a bigger screen." He concludes, "Prysm has transformed our
ability to recoup a higher rate of return on the information we've invested in…
ultimately our business can achieve double-digit growth." A sunset tractor harvest rolls
while a brief whoosh ushers in the final Prysm logo, and reflective piano fades to
silence."

 Only return a JSON with the key "caption" and the corresponding caption as the value,
and nothing else. Output "None" if you do not have enough details for it. Focus on
quality. Here is the input information:
```

Figure 9: **Prompt for Detailed Captioning with Transcripts. (Part 2 of 2)** Instructs the LLM to produce a single comprehensive audio-visual caption (up to 1000 words) that integrates speech transcripts in-line with descriptions of visuals, sounds, and scene dynamics across the entire video.

Table 9: Licenses for all datasets and model checkpoints used in AV-Flamingo.

Asset	Type	License	Reference
Model Checkpoints			
OmniVinci	Init. checkpoint	NVIDIA OneWay Noncommercial License	Ye et al. (2025)
AF-Whisper	Audio encoder	NVIDIA OneWay Noncommercial License	Goel et al. (2025)
Qwen2.5-7B	Base LLM	Apache 2.0	Team. (2025)
Training Datasets			
YouTube-8M	Video	CC BY 4.0	Abu-El-Haija et al. (2016)
HD-VILA	Video	Microsoft Research License (non-commercial)	Xue et al. (2022)
InternVid	Video	CC BY-NC-SA 4.0	Wang et al. (2024b)
VidChapters-7M	Video	MIT License; YouTube/API terms for source media	Yang et al. (2023)
HarmonySet	Video	Not specified; source-video terms apply	Zhou et al. (2025)
LSMDC	Video	Research use only	Rohrbach et al. (2016)
MMTrail	Video	R-UDA 1.0 (non-commercial research)	Chi et al. (2024)
MovieCLIP	Video	Not specified; source-media terms apply	Bose et al. (2022)
MiraData	Video	GPL-3.0 metadata; source-video terms apply	Ju et al. (2024)
AF3 training mix	Audio	NVIDIA OneWay Noncommercial License	Goel et al. (2025)
AudioSkills-XL	Audio	NVIDIA OneWay Noncommercial License	Goel et al. (2025)
MF-Skills	Audio	NVIDIA OneWay Noncommercial License	Ghosh et al. (2025a)
OmniVinci image-text mix	Image-text	Mixed public sources; see source licenses	Ye et al. (2025)
OmniVinci video-text mix	Video-text	Mixed public sources; see source licenses	Ye et al. (2025)
LibriSpeech	Speech	CC BY 4.0	Panayotov et al. (2015)
CHiME	Speech	Research use/source corpus terms	Vincent et al. (2013)
AliMeeting	Speech	Research/challenge use terms	Yu et al. (2022)
CoVoST	Speech	CC0 (Common Voice-based)	Wang et al. (2020)
GigaSpeech	Speech	Apache 2.0	Chen et al. (2021)
EMILIA	Speech	CC BY-NC 4.0	He et al. (2024)
Aishell-1	Speech	Apache 2.0	Bu et al. (2017)
Aidatatang	Speech	CC BY-NC-ND 4.0 (retracted by distributor)	Beijing DataTang Technology Co., Ltd. (2018)
Multi-talker Switchboard	Speech	LDC User Agreement	Godfrey and Holliman (1993)
MuST-C	Speech	CC BY-NC-ND 4.0	Cattoni et al. (2021)
AV-Safety QA	AV safety	AVF release terms	Ours
Audio-Think-Time	Audio reasoning	AVF release terms	Ours
Video-Think-Time	Video reasoning	AVF release terms	Ours
AV-Think	AV reasoning	AVF release terms	Ours
Baseline Models (Evaluation Only)			
Gemini 1.5/2.0/2.5	AV Model	Proprietary (API Terms of Service)	Team et al. (2024); Team (2025)
GPT-4o	LLM	Proprietary (API Terms of Service)	Hurst et al. (2024)
Qwen2.5-Omni	AV Model	Apache 2.0	Xu et al. (2025a)
Qwen3-Omni	AV Model	Apache 2.0	Xu et al. (2025b)
Audio Flamingo 3	Audio Model	NVIDIA OneWay Noncommercial License	Goel et al. (2025)
Phi-4-mm	AV Model	MIT License	Abouelenin et al. (2025)
NVILA	VLM	CC BY-NC-SA 4.0 for weights; Apache 2.0 for code	Liu et al. (2025a)

You are a helpful AI assistant. You need to act as an audio-visual caption generator for long videos that requires advanced understanding of the entire video (including the associated audio). My final objective is to train an audio-visual understanding agent with these captions to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate 1 caption with the following requirements:
1. The caption should cover the entire video.
2. The caption should include as much details as possible but not include redundant ones.
3. The caption should not be more than 500 words.
4. Do NOT use phrases like "according to the details" in the caption.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair.**
6. Most importantly, generate the captions from an audio-visual perspective, i.e., an agent should be able to generate seeing both visuals and sounds. Do not generate captions that focus on only one.
7. Do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments as one full video and your output should be such that the model I am training has no access to the segment-wise captions and is considering the the video as a whole. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
8. Lastly, you should be aware that audio captions and whisper transcripts may have hallucinations. For audio captions, always use visual to cross-check, if it's not a background sound. For whisper transcripts, transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", etc) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

An example caption is as follows (to get an example of style):
"A blue title card reading "Standard Process — Transforming Lives Through Prysm" dissolves to a tractor plowing lush rows, while a calm male narrator (light instrumental bed) introduces the 90-year nutrition company; visuals pan the Wisconsin headquarters, water-tower skyline, and test fields as he outlines a soil-first, whole-food philosophy and North-Carolina R&D hub in dry, music-free audio. Inside, amber supplement bottles rattle along conveyors behind the same speaker describing control "from seed to supplement, crop to clinic, plant to patient." The scene shifts to a 190-inch Prysm wall where suited presenters, backed by soft ambient strings, swipe through razor-sharp full-body anatomy, rotating 3-D brain models, panoramic lakes-and-mountains, maps, and harvest footage, emphasizing costly research, practitioner education, and faster insight adoption. Classical D-minor strings underscore interviews praising Prysm's "wow" factor, tactile engagement, collaboration tools, and live data ROI, while pride-tinged voices cite double-digit growth and eliminating "muddy-boot" farm visits. A sunset tractor harvest bridges to the final Prysm logo over a brief whoosh and reflective instrumental fade-out."

Only return a JSON with the key "caption" and the corresponding caption as the value, and nothing else. Output "None" if you do not have enough details for it. Focus on quality. Here is the input information:

Figure 10: **Prompt for Detailed Captioning.** Instructs the LLM to produce a single audio-visual caption (up to 500 words) summarizing the full video from a joint perceptual perspective, without segment-wise structure or explicit timestamps.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and create question-answer pairs of type **audio-visual event alignment** that require advanced understanding of the video (including the associated audio) and reasoning over it to answer correctly. AV-alignment questions should ask about **the exact audio event that coincides with a specific visual frame—or vice-versa**. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. Segments are chronological; missing info appears as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA must identify the **synchronous audio cue or visual cue** linked to a particular frame.
2. The question should be abstract and conceal obvious hints.
3. Neither question nor answer may exceed 20-30 words.
4. Do NOT use phrases like "according to the details"; write as if you observed and heard the video yourself.
5. Audio captions may be noisy; verify them with visuals or something similar.
6. Ensure both audio and visual evidence are needed to answer.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Example event-sequence QAs:
Question: "At the moment the static shot of Obsidian Fury first appeared, which specific audio event was synchronized with it?"
Options: "A. Speaker's self-introduction and figure presentation", "B. Reference to avoiding plot spoilers", "C. Logo transition sound effect", "D. Analysis of combat sequences"
Answer: "A. Speaker's self-introduction and figure presentation"

Generate up to 10 options.
Only return a JSON with the keys like this: {"Question": "...","Choices": { "A": "...", "B": "...", "C": "...", "D": "..." },"Answer": "X"}. If this question type is impossible, output "None". Focus on quality. Here is the input information:

Figure 11: **Prompt for Audio-Visual Event Alignment QA generation.** Instructs the LLM to produce a multiple-choice question identifying the exact audio cue that coincides with a specific visual frame, or vice versa, requiring tight cross-modal synchronization.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of type **event sequence** that require advanced understanding of the video (including the associated audio) and reasoning over it to answer correctly. Event-sequence questions should ask about **the order in which key audio-visual events occur across the timeline**. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. The segments are in chronological order. If any information is not available, it is given as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA must rely on recognizing the **relative order** of two or more significant audio-visual events, not a single clip.
2. The question should be abstract and conceal obvious hints.
3. Neither question nor answer may exceed 20-30 words.
4. Do NOT use phrases like "according to the details"; write as if you observed and heard the video yourself.
5. Audio captions may be noisy; verify them with visuals.
6. Ensure both audio and visual evidence are needed to answer.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Example event-sequence QAs:
Question: "What visual elements were displayed immediately after Dr. Rajani's 'BOTOX WITHOUT THE BOTOX' video concluded?"
Options: "A. Still product bottle → Price text overlay", "B. Facial treatment demonstration → Presenter holding product while explaining", "C. Presenter's torso shot → Secondary screen activation", "D. Bookshelf backdrop → Close-up of string lights"
Answer: "B. Facial treatment demonstration → Presenter holding product while explaining"

Generate up to 10 options.
Only return a JSON with the keys like this: {"Question": "...","Choices": { "A": "...", "B": "...", "C": "...", "D": "..." },"Answer": "X"}. If this question type is impossible, output "None". Focus on quality. Here is the input information:

Figure 12: **Prompt for Event Sequence QA generation.** Instructs the LLM to produce a multiple-choice question about the relative temporal ordering of two or more significant audio-visual events spanning the video timeline.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos with both audio and visual cues that can assess the model's understanding of both audio and visual cues in the video and relationships between audio-visual events, including its ability to reason over them. The questions should require a complete understanding of all audio and visual events throughout the video and deep thinking to respond. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA **should not** be for a particular event in the audio-visual scene but should require an understanding of the entire or a significant part of the video to answer.
2. The question should be abstract and not reveal too much about the video or hints that can help answer the question.
3. Neither the prompt nor the response should be more than 15-20 words.
4. Do NOT use phrases like "according to the details" in both the prompts or responses; you should ask and answer as if you were hearing the audio by yourself.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair.**
6. Most importantly, generate the questions and answers from an audio-visual perspective, i.e., an agent should be able to answer seeing both visuals and sounds. Do not ask questions that require only one.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

An example question for a movie would be:
"Why did the male protagonist arrange for his friend to rob the female protagonist?"

Only return a JSON with the Question and the Answer with these 2 keys and nothing else. If a question type is not possible, don't output and only output "None". Focus on quality. Here is the input information:

Figure 13: **Prompt for Holistic Reasoning QA generation.** Instructs the LLM to produce an open-ended question that requires integrating audio-visual evidence distributed across a substantial portion of the video, rather than any single localized moment.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and create question-answer pairs of type **inference** that require advanced understanding of the video (including the associated audio) and complex reasoning over it to answer correctly. Inference questions should ask about **unstated purposes, intentions, or outcomes that must be deduced from multiple audio-visual clues**. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. Segments are chronological; missing info appears as "Not Available".
Generate 1 question-answer pair with the following requirements:
1. The QA must be answerable only through logical deduction integrating clues spread across several segments.
2. The question should be abstract and conceal obvious hints.
3. Neither question nor answer may exceed 20-30 words.
4. Do NOT use phrases like "according to the details"; write as if you observed and heard the video yourself.
5. Audio captions may be noisy; verify them with visuals or something similar.
6. Ensure both audio and visual evidence are needed to answer.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Example event-sequence QAs:
Question: "What is the primary purpose of the white car maneuvering through the water-filled obstacle course with orange cones?"
Options: "**A.** Testing acceleration performance on wet surfaces", "**B.** Conducting routine vehicle safety inspections", "**C.** Demonstrating vehicle water resistance capabilities", "**D.** Simulating emergency response driving scenarios"
Answer: "**D.** Simulating emergency response driving scenarios" (the answer here is hidden in the audio and spoken content)

Generate up to 10 options.
Only return a JSON with the keys like this: {"Question": "...","Choices": { "A": "...", "B": "...", "C": "...", "D": "..." },"Answer": "X"}. If this question type is impossible, output "None". Focus on quality. Here is the input information:

Figure 14: **Prompt for Inference QA generation.** Instructs the LLM to produce a multiple-choice question whose answer can only be deduced by combining multiple implicit audio-visual clues spread across the video, with the correct option deliberately not stated directly.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of type **needle-in-the-haystack** that require advanced understanding of the videos (including the associated audio) and complex reasoning over it to solve. Needle-in-the-haystack questions are defined as questions about a particular instance in a long video (including corresponding audio). My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. This is the most important information which you should focus on. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA should either be for a particular and important event in the audio-visual scene **or** require an understanding of the entire or a significant part of the video to answer.
2. The question should be abstract and not reveal too much about the video or hints that can help answer the question.
3. Neither the prompt nor the response should be more than 20-30 words.
4. Do NOT use phrases like "according to the details" in both the question and answer; you should ask and answer as if you were watching and hearing the video by yourself.
5. The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair. Additionally, use the other visual details to understand the main importance of the broader scene and make QA pairs accordingly.
6. Most importantly, generate the questions and answers from an audio-visual perspective, i.e., an agent should be able to answer seeing both visuals and sounds. Do not ask questions that require only one.
7. Questions may be from the start, middle or end of the video. Do not always make questions only from the start.
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

An example question and answer pair would be: Question: What does the man do right before the woman fires the gun? Output Answer: He glances at his reflection and hesitates saying "we are in trouble".

Only return a JSON with the Question and the Answer with the keys Question and the Answer respectively and nothing else. If a question type is not possible, don't output and only output "None". Focus on quality. Here is the input information:

Figure 15: **Prompt for Needle-in-the-Haystack QA generation.** Instructs the LLM to produce an open-ended question targeting a specific but important moment in a long video, where broader audio-visual context is still required to identify it correctly.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of type **audio-visual referring** that requires understanding of the video (including the associated audio) and complex reasoning over it to answer correctly. Audio-Visual referring questions should ask about the visuals referring to a particular event in the audio -- or vice-versa. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA should either be for a particular and important event in the audio-visual scene **or** require an understanding of the entire or a significant part of the video to answer.
2. The question should be abstract and not reveal too much about the video or hints that can help answer the question.
3. Neither the prompt nor the response should be more than 20-30 words.
4. Do NOT use phrases like "according to the details" in both the questions and answers; you should ask and answer as if you were seeing the video and hearing the corresponding audio by yourself.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair. Additionally, use the other visual details to understand the main importance of the broader scene and make QA pairs accordingly.**
6. Most importantly, generate the questions and answers from an audio-visual perspective, i.e., an agent should be able to answer seeing both visuals and audio. Do not ask questions that require only one.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Example question-answer pairs: for audio referred video question is:
Question: "What items is the child holding when the dog's bark interrupts the conversation?",
Answer: "A toothbrush and a tube of toothpaste."
or a video referred audio question is:
Question: "When the boy in the blue hoodie faces the camera, which unexpected sound can be heard?"
Answer: "A baby suddenly starts crying."

Only return a JSON with the Question-Answer pairs with the keys "Question 1", "Answer 1", "Question 2", "Answer 2" respectively and nothing else. If a question type is not possible, don't output and only output "None" -- or return just one question. Focus on quality. Here is the input information:

Figure 16: **Prompt for Audio-Visual Referring QA generation.** Instructs the LLM to produce up to two question–answer pairs that link a visual event to its corresponding sound, or a sound to its corresponding visual scene, exercising cross-modal grounding.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of tepe **subscene understanding** that requires advanced understanding of the video (including the associated audio) and perform complex reasoning over it to correctly answer. Sub-scene understanding questions are defined as questions that would ask the model to caption a relevant and important part of a long video (including audio) that is preceded and succeeded by a particular set of events - or ask a question that requires specifically to understand the context of that part. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available"

Generate 1 question-answer pair with the following requirements:
1. The QA should either be for a particular and important event in the audio-visual scene **or** require an understanding of the entire or a significant part of the video to answer.
2. The question should be abstract and not reveal too much about the video or hints that can help answer the question.
3. Neither the prompt nor the response should be more than 20-30 words.
4. Do NOT use phrases like "according to the details" in both the questions and answers; you should ask and answer as if you were seeing the video and hearing the corresponding audio by yourself.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair. Additionally, use the other visual details to understand the main importance of the broader scene and make QA pairs accordingly.**
6. Most importantly, generate the questions and answers from an audio-visual perspective, i.e., an agent should be able to answer seeing both visuals and sounds. Do not ask questions that require only one.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Some example question-answer pairs are as follows:
Question: What happens between the man staring at his reflection and the woman raising her gun?;
Output Answer: The man turns away from the mirror, his expression shifting to fear as distant gunfire echoes.

Only return a JSON with the Question and the Answer with the keys Question and the Answer respectively and nothing else. If a question type is not possible, don't output and only output "None". Focus on quality. Here is the input information:

Figure 17: **Prompt for Sub-scene Understanding QA generation.** Instructs the LLM to produce an open-ended question about a meaningful intermediate portion of a long video, defined by the events that precede and follow it.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of type **temporal reasoning** that require advanced understanding of the video (including the associated audio) and perform complex reasoning over it to correctly answer. Temporal reasoning questions are defined as questions that would ask the model about the order of particular audio-visual events in the video (including audio) or require an understanding of the order of audio-visual events to answer correctly. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate **4 question-answer pairs (one of each type) with the following requirements:**
1. The QAs should either be for a particular and important event in the audio-visual scene **or** require an understanding of the entire or a significant part of the video to answer.
2. The questions should be abstract and not reveal too much about the video or hints that can help answer the question.
3. Neither the questions nor the answers should be more than 20-30 words.
4. Do NOT use phrases like "according to the details" in both the questions and answers; you should ask and answer as if you were seeing the video and hearing the corresponding audio by yourself.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair. Additionally, use the other visual details to understand the main importance of the broader scene and make QA pairs accordingly.**
6. Most importantly, generate the questions and answers from an audio-visual perspective, i.e., an agent should be able to answer seeing both visuals and sounds. Do not ask questions that require only one.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Some example question-answer pairs are as follows:
Question (Temporal Order): What is the correct order of events in the video?;
Options (Temporal Order): (A) A man hesitates while looking at his reflection (B) A man looks at a spoon (C) Gunfire erupts (D) A woman says "Try and bend the spoon";
Answer (Temporal Order): (C)(A)(D)(B)
Question (Temporal Referring type): What happens just before the gunshots are heard?;
Options (Temporal Referring type): (A) A man sees his reflection and hesitates (B) A child says, "Ben, it is only yourself" (C) A man puts a spoon in his mouth (D) A woman whispers in English;
Answer (Temporal Referring type): (A) A man sees his reflection and hesitates
Question (Attribute type): How does the intensity of both the visuals and sounds change throughout the video?
Options (Attribute type): (A) The scene becomes darker, and the sounds get more intense (B) The lighting remains the same, and the audio stays calm (C) The visuals become brighter, and the sounds fade away;
Answer (Attribute type): (A) The scene becomes darker, and the sounds get more intense.

Figure 18: **Prompt for Temporal Reasoning QA generation. (Part 1 of 2).**

Return a JSON in the following format:
{"Temporal Referring Question": Question, "Temporal Referring Options": Options, "Temporal Referring Answer": Answer, "Temporal Order Question": Question, "Temporal Order Options": Options, "Temporal Order Answer": Answer, "Temporal Attribute Question": Question, "Temporal Attribute Options": Options, "Temporal Attribute Answer": Answer}
Only return a JSON with the Questions and the Answers with these keys and nothing else. If a question type is not possible, don't output and only output "None". Focus on quality. Here is the input information:

Figure 18: **Prompt for Temporal Reasoning QA generation. (Part 2 of 2)** Instructs the LLM to produce four question–answer pairs covering Temporal Order, Temporal Referring, and Temporal Attribute reasoning over the chronological structure and evolving dynamics of the video.

You are a helpful AI assistant. You need to act as a question-answer generator for long videos and generate question-answer pairs of type **topic-level reasoning** that requires advanced understanding of the video (including the associated audio) and perform reasoning over it to correctly answer. Topic-level understanding questions should ask about the **main activity, objective, or theme** carried out across in the video, where one requires to watch the entire video and understand its content. My final objective is to train an audio-visual understanding agent with these question-answer pairs to endow it with long-video (including audio) understanding and QA abilities. You will be provided with a list of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like: [video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip]. This is the most important information which you should focus on. The 3 different captions provide you with different kinds of information -- you may use all to understand the video holistically, but may use only some from each segment to generate the question. The lists of segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA should either be for a particular and important event in the audio-visual scene **or** require an understanding of the entire or a significant part of the video to answer.
2. The question should be abstract and not reveal too much about the video or hints that can help answer the question.
3. Neither the prompt nor the response should be more than 20-30 words.
4. Do NOT use phrases like "according to the details" in both the questions and answers; you should ask and answer as if you were seeing the video and hearing the corresponding audio by yourself.
5. **The audio captions may be noisy and not correct. You need to verify using the visual captions and the other visual cues before framing your QA pair. Additionally, use the other visual details to understand the main importance of the broader scene and make QA pairs accordingly.**
6. Most importantly, generate the questions and answers from an audio-visual perspective, i.e., an agent should be able to answer seeing both visuals and sounds. Do not ask questions that require only one.
7. **Questions may be from the start, middle or end of the video. Do not always make questions only from the start.**
8. Lastly, do not consider the video segment-wise or do not have words like segment in your output. Understand the video holistically through the provided segments. Also ignore any explicit timestamps in the meta-data and do not include them in your output.
9. Lastly, you should be aware that whisper transcripts may have hallucinations. Transcripts that are extra small and only a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc.) may be a hallucination. You can also safely ignore such small utterances in your generated caption.

Only return a JSON with the Question and the Answer with the keys Question and the Answer respectively and nothing else. If a question type is not possible, don't output and only output "None". Focus on quality. Here is the input information:

Figure 19: **Prompt for Topic-Level Reasoning QA generation.** Instructs the LLM to produce an open-ended question about the main activity, objective, or overall theme of the video, requiring synthesis across the full timeline.

You are an Audio-Visual QA Generator. I will give you two JSON-like structures representing diarized and time-segmented captions of audio and visual streams of a video clip. Each segment contains timestamps. For audio, each segment is either a speech transcription or a non-speech event description. For visuals, they are visual captions describing all aspects of the visuals.
Your task is to generate two complete QA triplets and one caption from this audio-visual sequence. The two captions for audio and visual streams are provided to you separately, but you need to integrate and interleave them for all your generations.
Generation Requirements
- One Open-Ended Question-Answer-Thinking Process triplet
- One Multiple-Choice Question-Answer-Thinking Process triplet (8-10 options)
- One caption which describes the audio-visual sequence

Requirements for Question Generation
- Holistic Understanding: The question must require understanding the full audio-visual sequence, not just a small part of it. It should force the model to integrate information across the entire narration.
- Complex Reasoning: The question must demand heavy reasoning and multi-hop inference, going beyond surface-level transcription. Valid categories include (but are not limited to):
  - Needle-in-a-haystack: Identify a highly specific moment/event.
  - Subscene reasoning: Refer to scenes abstractly.
  - Other examples: causal chains, character-state inference, world understanding, implicit logic, etc.
- Abstraction: Avoid mentioning raw timestamps or verbatim phrases inside the question itself. The model must perform the mental work.
  - Thinking Chains:Include a <think> ... </think> reasoning process.
  - The reasoning must explicitly reference timestamps using <t>…</t> when grounding events.
  - The chain should sound natural, like a human listener reflecting on what they heard.
  - Do not mention "according to the input/caption/meta-data" or anything similar.
  - Multiple-Choice Construction:Provide 8-10 answer options.
  - Options should be plausible and confusing, not easily solvable via language priors.
  - Only one option should be correct.
- Most important: All generations should be interleaved with audio and visual thoughts with their corresponding timestamps. No question should be answerable through one modality alone.

Example Open-Ended Triplet
Question: How does the broadcast synergize the First Family's attire and specific editing techniques with the hosts' commentary to reframe the President's farewell address from a traditional goodbye into a narrative of somber, funereal defeat?
Answer: The broadcast constructs a narrative of a "political funeral" by tightly coupling the visual presentation of the First Family with commentary that explicitly recasts the event as mourning a failed presidency rather than celebrating a legacy.
Visually, the program cuts to shots of the First Family and the Vice President seated in a row, all wearing stark dark clothing (<t>147-152</t>). This imagery is immediately seized upon by the audio commentary, which explicitly points out that "his whole family was dressed in black" (<t>219-232</t>), drawing a direct parallel to a funeral aesthetic. This is reinforced by a specific editing choice: a semi-transparent "ghost" image of the President speaking is superimposed over the shot of his seated family (<t>152-157</t>), visually suggesting he is already a figure of the past or a spirit hovering over them, rather than a present, commanding leader.
The hosts amplify this somber mood by questioning the enthusiasm of the President himself, noting he "didn't seem like he was really into it" (<t>248-258</t>), and pivoting to his low approval ratings. This argument is solidified by on-screen graphics displaying a "FAILURE 61%" poll result (<t>383-388</t>) while the audio contrasts his decline with the rising popularity of his successor (<t>609-618</t>). The synergy of the "mourning" visual palette, the ghostly overlay, and the statistical graphics of failure transforms the farewell address into a grim eulogy for his administration.
Thinking Chain:
I need to trace how the broadcast weaves together visual symbolism and specific editing choices with the audio commentary to build a unified theme of "defeat" or "mourning."
First, the visual setup of the family: The narrative begins visually when the camera cuts to a shot of Jill Biden, Kamala Harris, and Doug Emhoff seated together (<t>147-152</t>). The key element here is their attire—they are all dressed in heavy, stark black clothing. This visual cue is immediately picked up by the audio commentary, where Steve Doocy explicitly points out, "his whole family was dressed in black" (<t>219-232</t>). By verbally flagging this visual detail, the host anchors the viewer's interpretation, suggesting the atmosphere is akin to a funeral without explicitly using the word yet, setting a tone of mourning rather than celebration.
Next, the "ghost" editing technique: The broadcast reinforces this "past tense" theme with a specific visual effect. The live feed of President Biden speaking is overlaid as a semi-transparent, ghostly image on top of the seated family (<t>152-157</t>). This editing choice visually separates him from the present reality of the room, framing him as a fading memory or a spirit hovering over his successors. The commentary supports this by speculating on the family's somber mindset, asking if they are thinking about "when they're going to be getting their pardons" (<t>235-239</t>), which couples the visual of the "ghost" and mourning clothes with the concept of the administration's political death.
Finally, the convergence of failure metrics: The narrative culminates by moving from mood to hard data. The audio argues that Donald Trump is now more powerful and popular than when he first took office (<t>609-618</t>). Simultaneously, this audio claim is synchronized with a harsh visual graphic on screen: "AMERICANS' VIEW OF JOE BIDEN'S PRESIDENCY - FAILURE 61%" (<t>383-388</t>). This alignment of audio and visual elements cements the argument: the "black dress" visual establishes the mood, the "ghost" overlay frames the President as a figure of the past, and the "Failure" chart provides the factual justification for the somber tone. Together, these elements transform the farewell address into a eulogy for a failed term.

Figure 20: Prompt used for synthesizing AV-Think reasoning triplets (Part 1 of 2).

Requirements for Caption Generation
- Timestamps are required for the caption but only for important and unique events. Add timestamps only where it makes most sense and is most attractive. Think like a human when captioning.
- Merging timestamps is required to refer to events.
- Do not mention "according to the input/caption/meta-data" or anything similar.

Example Caption
The segment opens with the Fox & Friends branding before cutting to President Biden delivering a farewell address from the Oval Office, where he is framed by the American flag and the Resolute Desk (<t>1-15</t>). The hosts describe the speech's tone as "grim" and "dark," a sentiment visually reinforced when the feed switches to correspondent Peter Doocy standing in the snow outside the White House (<t>19-42</t>). Doocy highlights the President's warnings, showing clips where Biden cautions against a rising "oligarchy" (<t>101-110</t>) and a "tech-industrial complex" (<t>134-145</t>), arguing that Biden only mistrusts these wealthy CEOs when they support his rival (<t>31-42</t>).

As the discussion moves to the studio, the visual narrative shifts to analyzing the demeanor of the First Family. The camera focuses on Jill Biden, Kamala Harris, and Doug Emhoff seated in a row, all dressed in stark black clothing (<t>147-152</t>). The hosts compare this aesthetic to a funeral (<t>336-342</t>) and speculate on the family's somber mood, asking if they are worrying about pardons (<t>235-239</t>). A distinctive visual overlay superimposes a semi-transparent, ghostly image of the speaking President over his seated family, enhancing the theme of a fading presidency (<t>152-157</t>).
The final act of the clip juxtaposes the President's warnings with his political standing. While the audio discusses his lack of defense from fellow Democrats (<t>422-434</t>), the visuals present stark polling graphics showing a 36% job approval and a "Failure" rating of 61% (<t>374-388</t>). The segment concludes with a discussion on the contrast between Biden's "bitter" tone and the rising excitement surrounding Donald Trump, framed by the hosts as a transition of power marked by relief rather than reverence (<t>609-632</t>).

Thinking Chain Guidelines
The thinking chains should ground to the time when referring to an event in the audio. Additionally, they should break down the thinking and do it step-by-step as humans would.

Output Format
Return a single JSON object:
{ "Question Open-Ended": "...", "Answer Open-Ended": "...", "Thinking Chain Open-Ended": "<think>...</think>", "Question MCQ": "...", "Options MCQ": ["(A) ...", "(B) ...", "..."], "Answer MCQ": "(X) ...", "Thinking Chain MCQ": "<think>...</think>", "Caption": "..." }
Output only the JSON. No explanation.
If a QA is not possible or only very simple ones are possible, return None for the corresponding keys.

Additional Notes
- Timestamps (in seconds as required) do not require decimal places even if present in the input JSON. In fact, round to the nearest second.
- Rarely, some timestamps in the input may be in minutes. You will only find this if the segment number is too long and the number in the timestamp is too small. That is for you to figure out. If so, please convert into seconds before caption/QA/reasoning generation.
- If the input information looks hallucinated (random/repeated/etc.), just return "None".

Here is the input information:

Figure 20: **Prompt used for synthesizing AV-Think reasoning triplets (Part 2 of 2).** Given timestamped audio and visual captions extracted from a long-form video, we prompt an LLM to generate a question–answer–reasoning triplet along with an integrated audio-visual caption. The prompt instructs the model to (i) require holistic understanding across both modalities, (ii) demand multi-hop temporal reasoning rather than surface-level transcription, and (iii) ground every intermediate thought to specific timestamps using `<t>...</t>` tags. We additionally enforce that no question be answerable from a single modality alone, ensuring that the resulting reasoning chains genuinely interleave audio and visual evidence.