

SynAV: Integrated Grounding and Captioning for Fine-grained Omni Video Understanding via Balanced Audio-Visual Synergy

Bolun Zhang^{1,2*}, Guangming Yao^{2†}, Canjun Yang^{1†}, Yi Yuan²

¹ Zhejiang University ² Ant Group

xiaotudou@zju.edu.cn, yaoguangming.ygm@antgroup.com

ycj@zju.edu.cn, yy464236@antgroup.com

Abstract

Video understanding is rapidly evolving toward omni-modal fusion, yet Multimodal Large Language Models (MLLMs) still struggle with complex audio-visual reasoning and semantic alignment. Current paradigms often rely on independent modal projectors to map features into the LLM space, an architecture that frequently overlooks intrinsic modal heterogeneity and inadvertently reinforces vision-centric bias by avoiding early, direct semantic interaction between native cross-modal representations. To address these challenges, we propose SynAV, a novel MLLM architecture designed for language-grounded multimodal reasoning. Our framework introduces a two-stage hierarchical cross-modal semantic interaction mechanism: it first performs alignment within the textual semantic space, and subsequently executes a visual-to-audio refinement that bridges semantic gaps and captures temporal dynamics. Furthermore, we employ a frame-by-frame interleaved audio-visual token arrangement enriched with explicit textual timestamps. This strategy acts as a temporal-semantic anchoring mechanism, allowing the LLM to better perceive the synchrony between modalities and seamlessly conduct long-sequence logical reasoning. Our model jointly optimizes temporal grounding and dense captioning tasks in an autoregressive manner. Extensive experiments across 11 benchmarks demonstrate that SynAV significantly outperforms state-of-the-art methods, achieving a superior performance in complex audio-visual language reasoning.

1 Introduction

The rapid advancement of Multimodal Large Language Models (MLLMs) has significantly pushed the boundaries of complex multimodal scene understanding and natural language interaction (Tang

et al., 2026; Wu et al., 2026). Integrating temporal grounding and dense captioning within Audio-Visual Question Answering (AVQA) evaluates a model’s ability to precisely localize events and reason over highly heterogeneous modalities. While existing methods have achieved notable progress on these tasks individually, few frameworks effectively unify them to achieve coherent language-grounded multimodal reasoning (Geng et al., 2025; Liu et al., 2025). The core challenges fundamentally lie in the semantic heterogeneity between auditory and visual streams, alongside a pervasive vision-centric bias where audio is treated as a secondary signal (Chen et al., 2024a; Park et al., 2025; Sun et al., 2024). Such disparities ultimately lead to poor cross-modal semantic alignment and hallucinated, temporally ambiguous descriptions when the LLM attempts to generate coherent natural language responses.

We argue that these fundamental hurdles are exacerbated by the architectural designs of current MLLM paradigms, which typically rely on independent modality projectors or isolated cross-attention modules to map features into a unified textual space (Huang et al., 2024; Geng et al., 2025; Liu et al., 2025). Consequently, the LLM fails to capture the fine-grained semantic nuances required to accurately reason over and describe synchronized multimodal events. We attribute these limitations to the lack of early semantic alignment and the absence of precise, text-guided temporal anchoring (Chen et al., 2025; Wang et al., 2025; Li et al., 2025c).

To address these limitations and facilitate language-instructed audio-visual alignment, we propose **SynAV**, an MLLM architecture designed for balanced **Audio-Visual Synergy**. We introduce a two-stage cross-modal semantic interaction mechanism: the first stage utilizes text instructions to establish a coherent baseline context, while the second leverages temporally discriminative audio

* Work done during an internship at Ant Group.

† Corresponding authors.

cues to direct visual attention for precise temporal grounding. Furthermore, the refined tokens are chronologically arranged in a frame-by-frame interleaved sequence with explicit textual timestamps ($\langle \text{time}_k \rangle$).

As illustrated in fig. 2, our framework achieves balanced improvements across grounding and captioning tasks on the LongVALE and VALOR datasets, alongside robust zero-shot generalization across 9 additional multimodal benchmarks. Furthermore, our ablation studies emphasize that while the visual stream constructs the basic factual context, deeply aligning and leveraging the auditory modality is crucial for robust temporal reasoning. By addressing the lack of early direct interaction mechanisms in current MLLMs, our unified model successfully enables simultaneous high-precision temporal grounding and richly detailed generative reasoning.

The main contributions of our work are summarized as follows:

- We present SynAV¹, a simple yet effective MLLM architecture that addresses semantic misalignment and vision-centric bias via a two-stage hierarchical semantic interaction mechanism, utilizing language instructions to deeply align discriminative audio cues and rich visual context.
- We introduce an interleaved token arrangement strategy coupled with explicit textual timestamps, functioning as a temporal-semantic anchoring mechanism that significantly improves the LLM’s capacity to logically reason over synchronized multimodal sequences.
- We demonstrate that our unified model achieves state-of-the-art performance on the challenging LongVALE and VALOR benchmarks for complex multi-task audio-visual reasoning, alongside robust zero-shot generalization across diverse multimodal benchmarks.

2 Related Work

2.1 Audio-Visual Video Understanding

While Multimodal Large Language Models (MLLMs) have transitioned from early vision-language models (Alayrac et al., 2022; Li et al.,

2023; Dai et al., 2023; Zhu et al., 2023; Liu et al., 2023; Huang et al., 2023) to omni-modal video systems (Lin et al., 2024; Maaz et al., 2024; Zhang et al., 2023; Li et al., 2025b; Song et al., 2024), they remain predominantly vision-centric. This impedes balanced multimodal fusion for crucial tasks like fine-grained captioning (Geng et al., 2025; Liu et al., 2025) and temporal grounding (Geng et al., 2025; Huang et al., 2024). Existing approaches typically tackle grounding (Huang et al., 2024; Pramanick et al., 2025; Qian et al., 2024; Lin et al., 2023) or captioning (Tang et al., 2025a; Li et al., 2024b; Zhang et al., 2025) in isolation, as joint training often incurs a performance trade-off due to inherent modality bias (Chen et al., 2023; Elizalde et al., 2023). Furthermore, contemporary models still struggle with this “feature space gap”: Video-SALMONN (Tang et al., 2025a) lacks explicit timestamps for long-form precision, Qwen-Omni (Xu et al., 2025) connects modalities without fine-grained refinement, and TimeMarker (Chen et al., 2024b) omits audio entirely. To overcome the vision-centric bias, our framework explicitly uses audio temporal dynamics as primary queries to guide cross-modal fusion, ensuring balanced and fine-grained modality integration.

2.2 Cross-Modal Alignment for Fine-grained Video Understanding

Unlike methods relying on simple linear projectors (LongVALE-LLM (Geng et al., 2025)), symmetric bidirectional cross-attention (Dolphin (Guo et al., 2025)), or complex dual pathways (AV-Master (Zhang et al., 2025)), our design leverages audio’s temporal sensitivity as queries for task-aware cross-modal alignment. While AVicuna (Tang et al., 2025b) also interleaves audio-visual tokens, it does so without deep cross-modal fusion or explicit timestamp anchoring; our framework uniquely combines frame-level interleaving, explicit textual timestamps, and deeply fused semantic anchors to actively resolve multimodal heterogeneity and guide complex sequential alignment. Furthermore, while works like TimeChat (Ren et al., 2024) and ARC-Chapter (Pu et al., 2025) explore textual timestamps within vision-centric pipelines, our framework uniquely interleaves time tokens as semantic anchors with deeply fused features to actively resolve multimodal heterogeneity and guide complex sequential alignment.

¹Code is available at <https://github.com/FlyingPTT/emnlp2026-find-synav>.

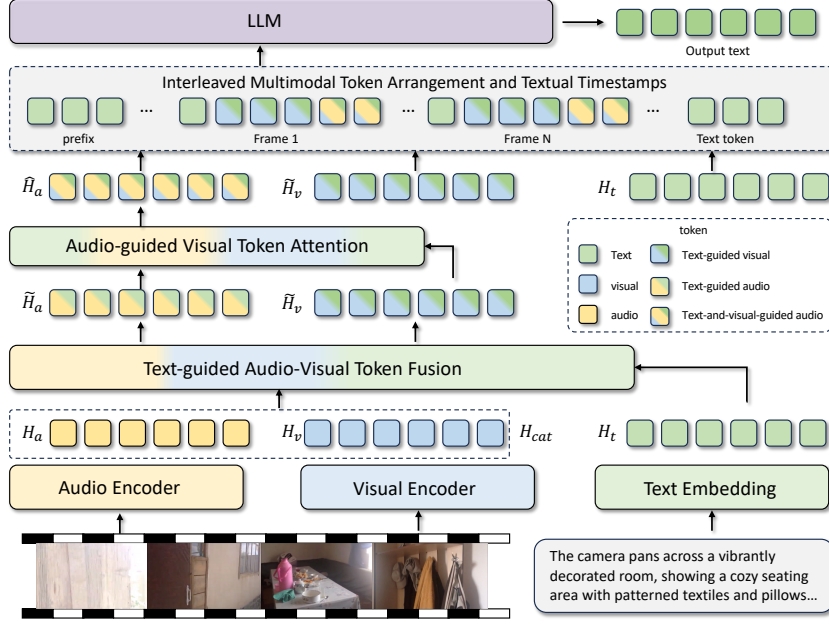


Figure 1: Overview of our proposed multimodal understanding framework. Different colors of tokens indicate the fusion of information from different modalities.

3 Method

In this section, we detail our proposed framework for multimodal understanding. As illustrated in Figure 1, our approach consists of three main stages: (1) modality-specific feature extraction, (2) hierarchical multimodal token fusion, and (3) interleaved token arrangement for Large Language Model (LLM) processing.

3.1 Multimodal Feature Extraction

Given an input sample comprising audio A , visual frames V , and text T , we first employ modality-specific encoders to extract feature representations. Unlike previous frameworks using dual audio encoders for distinct tasks (e.g., LongVALE-LLM (Geng et al., 2025)), we utilize a single, unified multi-task audio encoder to simplify cross-modal alignment. Let E_{audio} , E_{visual} , and E_{text} denote the audio encoder, visual encoder, and text embedding layer, respectively. The initial token sequences are obtained as:

$$\begin{aligned} \mathbf{H}_a &= E_{audio}(A), \\ \mathbf{H}_v &= E_{visual}(V), \\ \mathbf{H}_t &= E_{text}(T) \end{aligned} \quad (1)$$

where $\mathbf{H}_a \in \mathbb{R}^{N_a \times d}$, $\mathbf{H}_v \in \mathbb{R}^{N_v \times d}$, and $\mathbf{H}_t \in \mathbb{R}^{N_t \times d}$ represent the audio, visual, and text token embeddings, with d being the hidden dimension.

3.2 Cross-modal Fusion Module vs. Standard Cross-Attention

We introduce a unified cross-attention module, $\mathcal{F}_{fusion}$, to align variable-length multimodal sequences. Since standard cross-attention often causes training instability and fails to effectively leverage the complementary strengths of both modalities, we propose a stabilized fusion module:

$$\tilde{\mathbf{Q}} = \text{RMSNorm}(\mathbf{Q}) \quad (2)$$

$$\mathcal{F}_{fusion}(\mathbf{Q}, \mathbf{C}) = \mathbf{Q} + \text{Dropout}(\text{MultiHeadAttn}(\tilde{\mathbf{Q}}, \mathbf{C}, \mathbf{C})) \quad (3)$$

where query $\mathbf{Q}$ is the primary signal and context $\mathbf{C}$ provides keys and values. To ensure numerical stability during joint training, we integrate Pre-Layer Normalization (RMSNorm) and residual dropout. This bounds the operator norm, stabilizes gradients, and prevents exploding activations between modalities.

3.3 Hierarchical Multimodal Fusion

We employ a decoupled semantic-spatiotemporal alignment path. In the first stage, we use textual semantic guides to bridge the conceptual gap between modalities, directing audio-visual encoder tokens to focus on text query elements relevant to

temporal descriptions. In the second stage, we execute a visual-to-audio refinement that leverages the time-sensitive characteristics of audio tokens to enhance the temporal boundary representation of audio-visual events, effectively adapting to both synchronous and asynchronous or synergistic and complementary states inherent within each event.

Both stages utilize the same $\mathcal{F}_{fusion}(Q, C)$ module defined previously. While employing identical architectures, they maintain independent learnable parameters to capture distinct cross-modal correlations.

In the first stage, we align audio and visual features with textual semantics. We concatenate the audio and visual tokens to form the query sequence $\mathbf{H}_{cat} = [\mathbf{H}_a; \mathbf{H}_v]$, and use text tokens $\mathbf{H}_t$ as the context. Let $\mathcal{F}_{fusion}(\mathbf{Q}, \mathbf{C})$ denote the Multi-Head Cross-Attention (MHCA) based fusion block, where $\mathbf{Q}$ acts as the query and $\mathbf{C}$ serves as the key and value context. The aligned representation $\tilde{\mathbf{H}}_{cat}$ is obtained as:

$$\tilde{\mathbf{H}}_{cat} = \mathcal{F}_{fusion}(\mathbf{H}_{cat}, \mathbf{H}_t) \quad (4)$$

Subsequently, $\tilde{\mathbf{H}}_{cat}$ is split back into updated audio and visual tokens, $\tilde{\mathbf{H}}_a$ and $\tilde{\mathbf{H}}_v$, according to the original concatenation order: $[\tilde{\mathbf{H}}_a; \tilde{\mathbf{H}}_v] = \tilde{\mathbf{H}}_{cat}$.

In the second stage, we further refine the audio representations by leveraging the visual context. Specifically, the updated audio tokens $\tilde{\mathbf{H}}_a$ serve as queries ($\mathbf{Q}$), while the updated visual tokens $\tilde{\mathbf{H}}_v$ act as the grounding context ($\mathbf{C}$). This yields the final refined audio tokens $\hat{\mathbf{H}}_a$:

$$\hat{\mathbf{H}}_a = \mathcal{F}_{fusion}(\tilde{\mathbf{H}}_a, \tilde{\mathbf{H}}_v) \quad (5)$$

This mechanism allows the model to leverage visual cues to refine the auditory stream, thereby simultaneously enhancing the representation capabilities for both fine-grained captioning and temporal grounding tasks.

In summary, the feature sequences prepared for the subsequent interleaved arrangement include the twice-fused audio tokens $\hat{\mathbf{H}}_a$, the once-fused visual tokens $\tilde{\mathbf{H}}_v$, and the original textual embeddings $\mathbf{H}_t$.

3.3.1 Multi-turn Inference Efficiency

A natural concern regarding our text-guided cross-modal fusion is the potential computational overhead during multi-turn dialogues, as changing text instructions might imply recomputing representations. However, during inference, the heavy

modality-specific encoders (E_{audio} and E_{visual}) are evaluated only once, and the resulting deep features ($\mathbf{H}_a$ and $\mathbf{H}_v$) are cached. For each new dialogue turn, only the highly lightweight cross-modal fusion module $\mathcal{F}_{fusion}$ is re-executed with the new text queries. This decoupled architectural design intuitively ensures that multi-turn interactions remain highly efficient without redundant feature extraction.

3.4 Interleaved Arrangement with Textual Timestamps

To better leverage the refined features obtained from our hierarchical fusion module, we employ a frame-by-frame interleaved arrangement for audio-visual tokens, with each pair preceded by an explicit textual timestamp $\langle \text{time}_k \rangle$ to achieve precise temporal anchoring.

The relationship between the fused feature sequences and their segment-wise components is defined as:

$$\begin{aligned} \hat{\mathbf{H}}_a &= [\hat{\mathbf{h}}_a^{(1)}, \dots, \hat{\mathbf{h}}_a^{(N)}], \quad \tilde{\mathbf{H}}_v = [\tilde{\mathbf{h}}_v^{(1)}, \dots, \tilde{\mathbf{h}}_v^{(N)}] \\ \mathbf{H}_t &= [\mathbf{H}_t^{prefix}, \mathbf{H}_t^{(t_1)}, \dots, \mathbf{H}_t^{(t_N)}, \mathbf{H}_t^{suffix}] \end{aligned} \quad (6)$$

where $\tilde{\mathbf{h}}_v^{(k)}$ represents the spatial patches of the k -th visual frame, $\hat{\mathbf{h}}_a^{(k)}$ denotes the auditory tokens partitioned along the temporal dimension, and $\mathbf{H}_t^{(t_k)} = E_{text}(\langle \text{time}_k \rangle)$ is the embedding of the k -th timestamp.

The final input sequence $\mathbf{S}$ fed into the LLM is:

$$\mathbf{S} = [\mathbf{H}_t^{prefix}, \langle \text{time}_1 \rangle, \tilde{\mathbf{h}}_v^{(1)}, \hat{\mathbf{h}}_a^{(1)}, \dots, \langle \text{time}_N \rangle, \tilde{\mathbf{h}}_v^{(N)}, \hat{\mathbf{h}}_a^{(N)}, \mathbf{H}_t^{suffix}] \quad (7)$$

This arrangement better allows the LLM to reason about the temporal dynamics of the audio-visual content in conjunction with the textual instructions.

4 Experiments

4.1 Experimental Setup

We evaluate our method on fine-grained understanding benchmarks VALOR-32K (Liu et al., 2025) and LongVALE (Geng et al., 2025), reporting Recall@1 at various IoUs and mIoU for grounding, and BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SODA_c (Fujita et al., 2020) for segment and dense

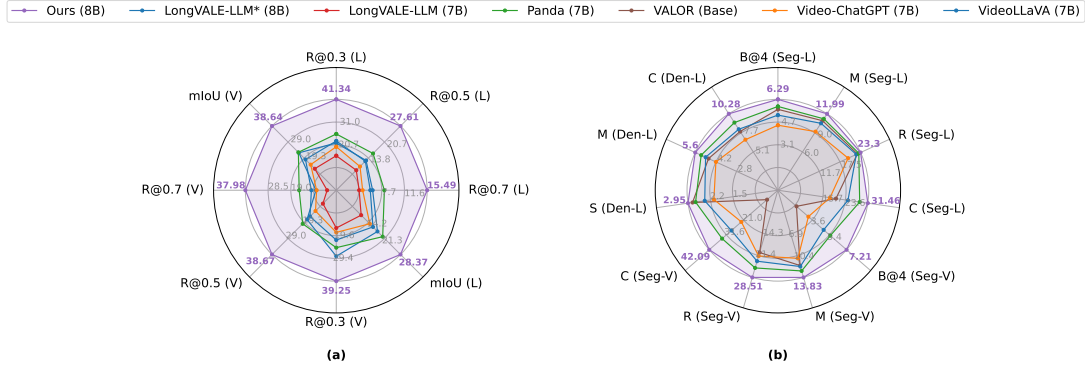


Figure 2: Comparison of (a) temporal grounding and (b) captioning performance on LongVALE (L) and VALOR-32K (V) datasets. R: Recall at different IoU thresholds, mIoU: mean IoU. B@4: BLEU-4, M: METEOR, R: ROUGE-L, C: CIDER, S: SODA_c. Seg: segment captioning, Den: dense captioning. Note that LongVALE-LLM* represents a variant of the LongVALE-LLM baseline modified to use our unified foundational models (Qwen3-8B) for an identical structural comparison, isolating architectural performance gains.

Table 1: Comparison of zero-shot accuracy (%) on audio-visual benchmarks.

Method	Size	Music-AVQA	AVSD	VGGSound
PandaGPT (Su et al., 2023)	13B	33.7	26.1	32.7
Macaw-LLM (Lyu et al., 2023)	7B	31.8	34.3	36.1
VideoLLaMA (Zhang et al., 2023)	7B	36.6	36.7	40.8
ChatBridge (Zhao et al., 2023)	13B	43.0	-	-
X-InstructBLIP (Panagopoulou et al., 2024)	13B	44.5	-	-
AV-LLM (Shu et al., 2025a)	13B	45.2	52.6	47.6
OneLLM (Han et al., 2024)	7B	47.6	-	-
AVicuna (Tang et al., 2025b)	7B	49.6	53.1	-
LongVALE-LLM (Geng et al., 2025)	7B	49.4	54.8	-
Ours	8B	58.05	59.8	50.1

captioning. For zero-shot evaluation, we assess top-1 accuracy on audio-visual benchmarks (Music-AVQA (Li et al., 2022), AVSD (Alamri et al., 2019), VGGSound (Chen et al., 2020), AVQA (Yang et al., 2022)) and general video benchmarks (Video-MME (Fu et al., 2025), MVBench (Li et al., 2024b), MLVU (Zhou et al., 2025), NextQA (Xiao et al., 2021), EgoSchema (Mangalam et al., 2023)).

4.2 Main Results

4.2.1 Performance on Fine-grained Grounding and Captioning

To evaluate the fine-grained understanding capabilities of our model, we conduct experiments on audio-video temporal grounding and captioning (including segment and dense captioning) across the LongVALE and VALOR-32K benchmarks. As illustrated in fig. 2, our approach achieves a substantial and balanced performance boost compared to the state-of-the-art baselines. On the temporal grounding task (fig. 2a), our model demonstrates

superior localization precision, significantly outperforming the baseline in terms of Recall@IoU and mean IoU (mIoU) metrics. This improvement is particularly evident on the VALOR-32K dataset, where our mIoU increases significantly from 12.93% to 38.64%, highlighting the effectiveness of our proposed interleaved arrangement and two-stage fusion. For captioning tasks (fig. 2b), our model consistently yields higher scores across both fragment-level segment captioning and global dense captioning metrics. This dual-track improvement validates that our balanced audio-visual synergy architecture significantly enhances both temporal grounding precision and captioning accuracy.

To explicitly rule out the confounding factor of utilizing a more advanced foundational model, we re-implement the state-of-the-art baseline, LongVALE-LLM, substituting its legacy encoders with our exact setup (Qwen3-VL and Qwen3-Omni). We denote this reproduced baseline as LongVALE-LLM*. Note that LongVALE-

Table 2: Comparison of zero-shot accuracy (%) on video understanding benchmarks.

	Size	Video-MME	MVBench	MLVU	NextQA	EgoSchema
GPT-4V (OpenAI, 2023)	-	59.9	43.7	-	-	-
VTimeLLM (Huang et al., 2024)	7B	-	-	41.9	-	-
LLaVA-NeXT-Interleave (Li et al., 2025a)	7B	-	53.1	-	78.2	-
LLaVA-NeXT-Video (Zhang et al., 2024)	7B	-	46.5	-	-	43.9
LLaVA-OneVision (Li et al., 2024a)	7B	58.2	56.7	64.7	79.4	-
Video-XL (Shu et al., 2025b)	7B	55.5	55.3	64.9	-	-
Video-CCAM-v1.1 (Fei et al., 2024)	9B	-	64.6	-	-	-
Video-CCAM-v1.1 (Fei et al., 2024)	14B	53.2	-	63.1	-	-
TimeMarker (Chen et al., 2024b)	8B	57.3	67.4	63.9	-	-
Ours	8B	62.3	66.2	66.8	78.4	44.6

Table 3: Fair comparison and SOTA baselines. LongVALE-LLM* uses the same Qwen3-8B backbone as SynAV; its native fusion is a linear projector + concatenation, hence numerically equivalent to Naive Concat / Serial [Video;Audio] in Tables 4 and 6. Video-SALMONN 2 and Dolphin are evaluated under the same protocol. All SynAV gains over LongVALE-LLM* are significant ($p < 0.01$, paired t -test over three seeds); see Appendix D for full mean $\pm$ std. (G) and (C) denote grounding (mIoU) and captioning (CIDEr).

Method	Params	LongVALE		VALOR		Music-AVQA Acc.
		mIoU (G)	CIDEr (C)	mIoU (G)	CIDEr (C)	
LongVALE-LLM (Geng et al., 2025)	7B	11.0	16.5	12.9	13.2	49.4
LongVALE-LLM*	8B	18.2	24.1	22.5	18.9	-
Video-SALMONN 2 (Tang et al., 2025a)	-	19.4	30.1	24.8	39.5	54.6
Dolphin (Guo et al., 2025)	-	17.6	26.3	22.1	34.7	52.9
SynAV (Ours)	8B	26.4	28.5	35.1	38.2	58.1

LLM’s native cross-modal fusion is a linear projector followed by simple concatenation (Eq. 2 in the original paper); therefore our faithful re-implementation coincides numerically with the Naive Concat and Serial [Video;Audio] baselines reported in Tables 4 and 6. As shown in Table 3, while the stronger backbone naturally elevates the lower bound, it still falls significantly short of SynAV (e.g., improving from 11.0 to 18.2 in mIoU on LongVALE, whereas SynAV reaches 26.4). This fair comparison directly confirms that the core performance gain inherently stems from our proposed two-stage cross-modal fusion and interleaved arrangement mechanisms, which bridge the modal heterogeneity better than traditional projection-based pipelines. All reported gains are statistically significant at $p < 0.01$ (paired t -test over three random seeds); detailed mean $\pm$ std values are provided in Appendix D.

We further compare SynAV against recent audio-visual MLLMs not included in the original LongVALE benchmark, namely Video-SALMONN 2 (Tang et al., 2025a) and Dolphin (Guo et al., 2025). As shown in Table 3, SynAV substantially outperforms both on temporal grounding (e.g., +6.9 mIoU on LongVALE

and +10.3 on VALOR over Video-SALMONN 2) and Music-AVQA accuracy. On captioning, Video-SALMONN 2 is competitive (CIDEr 30.1 vs. our 28.5 on LongVALE; 39.5 vs. 38.2 on VALOR), consistent with its strength in dense description generation. Our primary advantage lies in temporal grounding and balanced audio-visual synergy rather than pure captioning.

4.2.2 Audio-Visual and General Video Understanding

We further evaluate the proposed framework on a suite of zero-shot benchmarks to assess whether our single-stage SFT preserves core omni-modal capabilities and demonstrates robust generalization in general video understanding. Firstly, within the audio-visual domain, our model demonstrates remarkable zero-shot performance on Music-AVQA, AVSD, and VGGSound (Table 1), surpassing the LongVALE-LLM baseline across the board. Notably, the accuracy on Music-AVQA reaches 58.1%, reflecting a significant margin over previous best-performing methods. Secondly, to test general video perception, we evaluate on 5 comprehensive benchmarks including Video-MME and MVBench. As summarized in Table 2, our 8B model achieves superior performance, notably reaching 62.3% ac-

Table 4: Modality contribution analysis (Δ_{audio}). We compare our proposed hierarchical synergy mechanism against a standard naive concatenation baseline using the same strong backbones. Δ_{audio} represents the absolute performance gain achieved by adding the audio modality compared to the Video-Only baseline. (G) and (C) denote the grounding and captioning tasks, respectively.

Architecture	Modality	LongVALE		VALOR		Zero-shot Acc.	
		mIoU (G)	CIDEr (C)	mIoU (G)	CIDEr (C)	AVQA	Music-AVQA
Naive Concat Baseline	Video Only	17.1	24.5	22.4	26.8	68.4	28.2
	Audio Only	4.2	8.6	5.1	10.4	50.1	31.5
	Video + Audio	18.2	24.1	22.5	18.9	70.1	35.4
	Δ_{audio} (Concat)	+1.1	-0.4	+0.1	-7.9	+1.7	+7.2
SynAV (Ours)	Video + Audio	26.4	28.5	35.1	38.2	85.0	58.1
	Δ_{audio} (SynAV)	+9.3	+4.0	+12.7	+11.4	+16.6	+29.9

curacy on Video-MME and 66.2% on MVBench. On MVBench, SynAV is competitive with Video-CCAM (67.4%) but not universally leading, which is expected given that our model is specialized for audio-visual synergy rather than pure visual understanding. These results indicate that despite being specialized on dense, time-aware supervision (SFT), our architecture maintains and even leverages its pre-trained knowledge to excel in broader zero-shot reasoning tasks. This indicates that our balanced synergy and hierarchical fusion mechanism provide a more robust foundational representation for general video understanding.

4.3 Modality Synergy Analysis

4.3.1 Modality Contribution Analysis

To directly address whether the performance improvements of our model stem from the proposed architecture rather than merely the adoption of stronger backbones, we conduct a quantitative modality contribution analysis. We define Δ_{audio} as the absolute performance gain achieved by combining audio and video compared to a solely visual baseline (i.e., Video Only).

As shown in Table 4, we compare SynAV against a naive concatenation baseline that employs identical pre-trained backbones but fuses modalities via simple sequence concatenation. The naive baseline exhibits a severe vision-centric bias: its Δ_{audio} is marginal to negative across temporal grounding and fine-grained captioning tasks (e.g., -7.9 CIDEr on VALOR), indicating that the LLM space fails to effectively utilize native auditory features without explicit cross-modal interaction. In contrast, SynAV achieves a profound positive Δ_{audio} across all benchmarks. Most notably, on Music-AVQA, the inclusion of audio within our framework yields an unprecedented +29.9% boost in zero-shot accuracy, compared to a mere +7.2% in the base-

line. While Music-AVQA is audio-dominant by construction, the positive Δ_{audio} on non-audio-dominant benchmarks (LongVALE: +9.3 mIoU, +4.0 CIDEr; VALOR: +12.7 mIoU, +11.4 CIDEr) confirms that the balanced synergy generalizes beyond audio-heavy tasks. These results explicitly prove that our two-stage hierarchical fusion and interleaved arrangement mechanisms are the true drivers for successfully mitigating the vision-centric bias and activating deep audio-visual synergy.

4.3.2 Modality Robustness under Perturbation

To further evaluate how our model mitigates vision-centric bias in challenging, modality-imbalanced scenarios—a critical concern raised in existing omni-modal literature—we design a controlled perturbation study. Specifically, we apply two types of degradation to the LongVALE validation set and the Music-AVQA benchmark: 1) Visual Masking (V-Mask), where 50% of the spatial patches in the video frames are replaced with Gaussian noise to simulate severe visual impairment, and 2) Audio Masking (A-Mask), where the audio stream is entirely muted.

As illustrated in Appendix E (Table 21), the Naive Concat Baseline exhibits catastrophic performance drops under visual masking (e.g., from 35.4% to 14.2% on Music-AVQA) while remaining largely insensitive to audio masking. This confirms its deeply rooted vision-centric bias, treating audio merely as a dispensable auxiliary signal. Conversely, SynAV maintains robust performance when visual cues are compromised, relying on the intact audio stream to anchor the events. Furthermore, SynAV demonstrates a substantial performance dependency on the audio stream (evidenced by the steepest drop under A-Mask on

Music-AVQA, -28.6%), directly proving that our architecture achieves a balanced modality reliance and successfully integrates both streams into the core reasoning process.

4.3.3 Audio-Dependent Reasoning

As illustrated in Appendix H.2 (Figure 4, top), we compare our model with the vision-centric baseline Video-LLaVA (Lin et al., 2024). In scenarios where visual information is ambiguous or remains relatively static over a period of time, our model correctly leverages auditory cues to reason about the scene. For instance, in a house tour video, our model accurately identifies the Kyrgyz tradition of hospitality based on narrated explanations within the audio, whereas the baseline fails to perceive this non-visual cultural concept. This confirms that our two-stage fusion mechanism successfully grants audio an equal representational status.

To further investigate how our model bridges the feature space gap, we visualize the temporal distribution of attention weights in Appendix H.2 (Figure 4, bottom). Whereas the vision-centric baseline tends to rely excessively on visual cues, leading to misaligned or incomplete event understanding when visual information is insufficient, our model exhibits a strong correspondence between the attention peaks and the predicted temporal boundaries. This demonstrates our model’s ability to identify complementary states between auditory and visual streams to construct a cohesive omni-modal event.

4.4 Architecture Ablation Studies

In this section, we conduct extensive ablation studies to evaluate the effectiveness of our proposed components and training strategies.

4.4.1 Component Effectiveness

Table 5: Ablation study on different components. Comp.: Components, TT: Textual Timestamps, IA: Interleaved Arrangement, TSF: Two-stage Fusion. LV: LongVALE, VA: VALOR. G: mIoU for temporal grounding, C: CIDEr for captioning.

TT	Comp.		LV		VA	
	IA	TSF	G	C	G	C
			7.8	20.5	9.4	6.8
✓			11.5	22.1	14.2	12.4
✓	✓		16.2	25.4	21.5	25.6
✓	✓	✓	26.4	28.5	35.1	38.2

We investigate the contribution of the Textual Timestamps (TT), Interleaved Arrangement (IA),

and Two-stage Fusion (TSF). As shown in Table 5, each component provides consistent improvements. The combined model achieves the best performance, especially in temporal grounding, where the integration of TSF yields a significant boost ($+10.2\%$ mIoU on LongVALE and $+13.6\%$ on VALOR compared to the IA-only baseline), highlighting its effectiveness in bridging modal heterogeneity and providing precise temporal cues.

Anomaly Analysis of the Base Variant. An intriguing observation from Table 5 is that the base variant (lacking TT, IA, and TSF) yields an mIoU of merely 7.8 on LongVALE, which is lower than the 11.0 mIoU achieved by the legacy LongVALE-LLM (Table 3), despite utilizing a substantially stronger LLM backbone. We attribute this anomaly to *modality interference and hallucination bias*. When raw, unaligned heterogeneous modalities are directly concatenated and fed into a highly capable LLM without explicit synergy mechanisms, the model tends to experience severe training instability. The powerful LLM “shortcuts” the learning process by over-relying on internal text priors (hallucinating) rather than faithfully grounding the misaligned multimodal inputs. Quantitatively, the base variant exhibits a timestamp hallucination rate of 34.6% (predictions falling outside the video duration or yielding zero IoU), compared to only 8.7% for the full SynAV model. This directly corroborates that scaling up the foundational backbone is not a panacea; it necessitates our meticulously designed cross-modal synergy architecture to stably harness its full multimodal potential.

4.4.2 Token Arrangement Strategy

Table 6: Comparison of different token arrangement strategies. LV: LongVALE, VA: VALOR. G: mIoU for temporal grounding, C: CIDEr for captioning.

Strategy	LV		VA	
	G	C	G	C
Serial [Video; Audio]	18.2	24.1	22.5	18.9
Serial [Audio; Video]	17.5	23.8	21.0	17.4
Interleaved [V, A, V, A...]	26.4	28.5	35.1	38.2

We compare our interleaved strategy with traditional concatenation methods (Table 6). The results demonstrate that the two serial [Video; Audio] and [Audio; Video] baselines yield similar performance, as the LLM primarily treats them as independent, long-context sequences without deep cross-modal awareness. In contrast, our interleaved

arrangement allows the LLM to better perceive the synchrony between modalities, leading to a significant boost in temporal grounding tasks.

4.4.3 Fusion Direction Analysis

To isolate the contributions of the two fusion stages and their directionality, we conduct a systematic ablation varying the cross-modal fusion design. As shown in Table 7, we compare seven configurations: (i) Base (no TT/IA/TSF), (ii) Text-only (text-guided fusion only, no AV cross-attention), (iii) Stage1-only (text-to-AV alignment), (iv) Stage2-only (audio-to-visual cross-attention), (v) A→V (audio as query, visual as context), (vi) Bidirectional (symmetric cross-attention), and (vii) our V→A design (visual-to-audio refinement, i.e., audio as query).

Table 7: Fusion direction ablation. LV: LongVALE, VA: VALOR. G: mIoU, C: CIDEr. Base = no TT/IA/TSF; M-AVQA = Music-AVQA zero-shot accuracy.

Variant	LV		VA		M-AVQA Acc
	G	C	G	C	
Base	7.8	12.1	10.9	11.3	49.2
Text-only	13.2	18.6	16.5	15.7	51.4
Stage1-only	15.7	21.3	19.8	18.4	53.0
Stage2-only	18.5	24.1	23.4	21.6	54.3
A→V	20.8	25.1	26.7	27.9	55.9
Bidirectional	25.6	27.8	34.0	37.0	57.6
V→A (Ours)	26.4	28.5	35.1	38.2	58.1

Three key findings emerge. First, each fusion stage contributes independently: Stage1-only (15.7) and Stage2-only (18.5) both exceed the Text-only baseline (13.2) on LongVALE mIoU, and their combination (V→A) yields a synergistic gain (26.4), confirming that the two stages are complementary rather than redundant. Second, the direction matters: V→A outperforms A→V by +5.6 mIoU on LongVALE and +8.4 on VALOR, indicating that using audio as the query to attend to visual context is more effective for temporal grounding. Third, bidirectional fusion (25.6) is slightly below V→A (26.4), suggesting that a single, well-chosen direction suffices and avoids unnecessary parameters. A detailed analysis of temporal anchoring variants is provided in Appendix C.4.

For detailed analyses of normalization and training strategies, please refer to Appendix C.5 and Appendix C.6.

5 Conclusion

In this paper, we propose SynAV, a novel MLLM architecture that achieves state-of-the-art performance in fine-grained audio-visual grounding and captioning by addressing vision-centric bias through a hierarchical two-stage fusion mechanism and an interleaved token arrangement with textual timestamps. We hope our preliminary exploration of balanced synergy provides helpful insights for the community, encouraging further investigation into how intrinsic modality properties and task-specific adaptability can inform more effective architectural designs. Looking forward, we aim to extend this unified approach to explore Mixture-of-Experts (MoE) structures, reasoning-enhanced “thinking” models, and the integration of a broader range of heterogeneous multimodal tasks.

Limitations

Although SynAV achieves state-of-the-art performance in audio-visual temporal grounding and fine-grained captioning, it has several limitations:

- **Computational Overhead & Context Limit:** While our cross-modal fusion reduces heavy visual token redundancy, interleaving dense audio-visual tokens with explicit textual timestamps for very long-form videos still demands substantial context windows and memory from the LLM backbone.
- **Dependency on Audio Quality:** As demonstrated in our mask perturbation experiments, SynAV heavily relies on audio signals to anchor temporal boundaries. In real-world scenarios with extreme background noise or corrupted audio tracks, the performance may degrade.
- **Language Bias:** Current experiments and instruction tuning are predominantly conducted in English. The model’s cross-modal semantic alignment capabilities in multilingual or low-resource language settings remain to be fully evaluated.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful comments and constructive suggestions, which have significantly improved the quality of this work.

References

- Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. [Vggsound: A large-scale audio-visual dataset](#). In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725.
- Liangyu Chen, Zihao Yue, Boshen Xu, and Qin Jin. 2024a. [Unveiling visual biases in audio-visual localization benchmarks](#). *Preprint*, arXiv:2409.06709.
- Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xi-angzhan Yu, and Furu Wei. 2023. [BEATs: Audio pre-training with acoustic tokenizers](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.
- Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024b. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*.
- Yijing Chen, Yihan Wu, Kaisi Guan, Yuchen Ren, Yuyue Wang, Ruihua Song, and Liyun Ru. 2025. [Chronusomni: Improving time awareness of omni large language models](#). *Preprint*, arXiv:2512.09841.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. [Clap learning audio concepts from natural language supervision](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. 2024. [Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos](#). *Preprint*, arXiv:2408.14023.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*.
- Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Soda: Story oriented dense video captioning evaluation framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. 2025. Long-vale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18959–18969.
- Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. 2025. [Aligned better, listen better for audio-visual large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang

- Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. [Language is not all you need: Aligning perception with language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. [Llava-onevision: Easy visual task transfer](#). *arXiv preprint arXiv:2408.03326*.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. 2025a. [LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models](#). In *The Thirteenth International Conference on Learning Representations*.
- Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19108–19118.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025b. [VideoChat: chat-centric video understanding](#). *Science China Information Sciences*, 68(10):200102.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024b. [Mvbench: A comprehensive multi-modal video understanding benchmark](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.
- Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. 2025c. [Universal video temporal grounding with generative multi-modal large language models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. [Video-LLaVA: Learning united visual representation by alignment before projection](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.
- Chin-Yew Lin and Franz Josef Och. 2004. [Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics](#). In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.
- Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univt: Towards unified video-language temporal grounding.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2025. [Valor: Vision-audio-language omni-perception pre-training model and dataset](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):708–724.
- Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. [Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration](#). *Preprint*, arXiv:2306.09093.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. [Video-ChatGPT: Towards detailed video understanding via large vision and language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. [Egoschema: A diagnostic benchmark for very long-form video language understanding](#). *Preprint*, arXiv:2308.09126.
- OpenAI. 2023. [Gpt-4v\(ision\) system card](#). https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2024. [X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. 2025. [Assessing modality bias in video question answering benchmarks with multi-modal large language models](#).

- Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. 2025. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24297–24308.
- Junfu Pu, Teng Wang, Yixiao Ge, Yuying Ge, Chen Li, and Ying Shan. 2025. [Arc-chapter: Structuring hour-long videos into navigable chapters and hierarchical summaries](#). *Preprint*, arXiv:2511.14349.
- Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning.
- Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323.
- Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2025a. Audio-visual llm for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4246–4255.
- Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025b. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26160–26169.
- Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232.
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.
- Peiwen Sun, Honggang Zhang, and Di Hu. 2024. Unveiling and mitigating bias in audio visual segmentation.
- Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025a. [video-salmonn 2: Caption-enhanced audio-visual large language models](#). *Preprint*, arXiv:2506.15220.
- Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2026. [Video understanding with large language models: A survey](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376.
- Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. 2025b. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. [Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features](#). *Preprint*, arXiv:2502.14786.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [Cider: Consensus-based image description evaluation](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.
- Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. 2025. [Grouned-VideoLLM: Sharpening fine-grained temporal grounding in video large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 959–975, Suzhou, China. Association for Computational Linguistics.
- Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. 2026. [A survey on video temporal grounding with multimodal large language model](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1521–1541.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. [Qwen3-omni technical report](#). *Preprint*, arXiv:2509.17765.
- Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. [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, Singapore. Association for Computational Linguistics.

Jiayu Zhang, Qilang Ye, Shuo Ye, Xun Lin, Zihan Song, and Zitong Yu. 2025. [Av-master: Dual-path comprehensive perception makes better audio-visual question answering](#). *Preprint*, arXiv:2510.18346.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. [Llava-next: A strong zero-shot video understanding model](#).

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13691–13701.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Potential Risks and Ethical Considerations

While SynAV significantly advances fine-grained audio-visual understanding through balanced modality synergy, we acknowledge several potential risks and ethical considerations associated with the development and deployment of such Multi-modal Large Language Models (MLLMs):

- **Privacy and Surveillance:** The core strength of SynAV lies in its precise temporal grounding and dense captioning of audio-visual events. If deployed in public or private monitoring systems without proper consent, these fine-grained localization capabilities could be misused for invasive automated surveillance and unjust profiling.
- **Bias and Fairness:** Like many foundation models trained on web-scale datasets, our framework inherits the risk of amplifying demographic, cultural, or gender biases present in the training data. As noted in our limitations regarding language bias, the model may underperform or produce culturally misaligned descriptions when processing content from underrepresented communities.
- **Misinformation and Malicious Use:** The robust ability to align and reason over complex audio-visual sequences could theoretically be exploited by malicious actors. For instance, the detailed semantic understanding could be adapted to filter, segment, and manipulate media to generate highly convincing deceptive content (e.g., deepfakes), thereby contributing to the spread of misinformation.

We strongly advocate for the responsible use of multimodal understanding systems, emphasizing the necessity of human-in-the-loop oversight and the development of robust bias-mitigation strategies before any real-world deployment.

B System Efficiency Analysis

To address the computational overhead inherent in processing long-form videos with multiple modalities, we compare our model with standard early-fusion baselines, Video-SALMONN (Tang et al., 2025a), and standard late-fusion methods, LongVALE-LLM (Geng et al., 2025). As shown in Table 8, our proposed mechanism significantly

reduces memory consumption and inference latency. The core driver of this efficiency is our cross-modal alignment phase: by utilizing temporal audio features as primary queries, we apply $8\times$ mean-pooling to compress the audio token sequence before fusion, while the visual token count remains unchanged. Combined with the backbone’s native 2×2 spatial merge and the multi-turn caching of modality encoder outputs, this design prevents the sequence length from exploding, thereby mitigating the $O(N^2)$ complexity of the LLM’s self-attention mechanism. Consequently, this bottleneck-alleviating compression transforms a multimodal fusion necessity into an architectural advantage, enabling significantly faster processing throughput with lower memory footprints.

C Extended Ablation Study Results

This section provides comprehensive experimental data and analysis to complement the Ablation Studies presented in the main paper. While the main paper focuses on the most critical results, we here report extended metrics across multiple datasets (LongVALE (Geng et al., 2025) and VALOR-32K (Liu et al., 2025)) and detailed performance breakdowns for each module and strategy of our SynAV framework. For each ablation study, we conduct comprehensive evaluations on both datasets, covering two key fine-grained tasks: temporal grounding and segment captioning.

C.1 Component Effectiveness

We first investigate the individual contributions of our proposed components: Textual Timestamps (TT), Interleaved Arrangement (IA), and Two-stage Fusion (TSF). As shown in tables 9 and 10, the integration of each component leads to consistent improvements across all fine-grained metrics on both LongVALE (Geng et al., 2025) and VALOR-32K (Liu et al., 2025) datasets.

Specifically, the use of TT and IA provides a baseline for structured multimodal understanding, but adding our TSF module yields a significant boost in performance. In the temporal grounding task (table 9), TSF results in an improvement of +17.1% Recall@0.3 on LongVALE (Geng et al., 2025) and +7.9% on VALOR-32K (Liu et al., 2025), highlighting its effectiveness in bridging the heterogeneity between audio and visual modalities. Similarly, for segment captioning (table 10), the synergy between the interleaved sequence and the

Table 8: System efficiency comparison in terms of training memory per batch, inference latency (Speed), and inference GPU memory. Evaluated on an H800 GPU utilizing an identical sequence of video inputs.

Method	Train Mem (GB)	Speed (FPS)	Inf Mem (GB)
Video-SALMONN	42.5	15.2	18.4
LongVALE-LLM	38.6	21.0	16.5
SynAV (Ours)	28.4	36.5	12.8

fusion layer allows for more precise semantic generation, particularly reflected in the substantial increase in CIDEr scores (+3.1 on LongVALE (Geng et al., 2025) and +12.6 on VALOR-32K (Liu et al., 2025)). These results confirm that our proposed architecture successfully captures fine-grained temporal and semantic dependencies.

C.2 Token Arrangement Strategy

We evaluate our interleaved token arrangement strategy against traditional serial concatenation methods (tables 11 and 12). The results show that serial arrangements ([Video; Audio] or [Audio; Video]) perform similarly but significantly worse than our interleaved approach.

As discussed in the main paper, such serial strategies treat the modalities as independent long-context sequences, failing to establish deep cross-modal awareness. In contrast, our interleaved arrangement allows the model to perceive temporal synchrony directly within the sequence. This is evident in the grounding results (table 11), where the interleaved strategy achieves a +14.5% improvement in Recall@0.3 on LongVALE (Geng et al., 2025) compared to the [Video; Audio] baseline. For segment captioning (table 12), we observe that synchrony helps the model better align semantic content with specific timestamps, resulting in higher BLEU-4 and CIDEr scores.

C.3 Modality Analysis

To investigate the synergy between audio and visual information, we conduct an ablation study on modality combinations (tables 14 and 15). The results demonstrate that the multimodal “Video + Audio” setting significantly outperforms both “Video Only” and “Audio Only” configurations across all metrics.

While visual information remains dominant in many grounding and captioning scenarios, auditory cues provide critical supplementary information, especially in cases of high visual ambiguity or off-screen events. In the VALOR-32K (Liu et al., 2025) grounding task (table 14), adding audio leads

to a substantial performance leap compared to the Video Only baseline (+6.9% R@0.3). For captioning (table 15), the inclusion of audio enhances the richness and accuracy of descriptions. These findings further support our hypothesis that SynAV effectively mitigates vision-centric bias by leveraging audio-visual synergy for robust multimodal reasoning.

C.4 Temporal Anchoring Mechanism

To verify that our interleaved arrangement with textual timestamps is both necessary and sufficient for temporal alignment, we compare four variants in Table 13: (i) None (pure interleaving without textual timestamps), (ii) Plain (our default), (iii) + Time-local (restricting cross-modal attention to a ± 5 -second window), and (iv) + Cross-modal (shared sinusoidal positional encoding indexed by wall-clock time across modalities).

Removing timestamps causes a substantial drop (-8.5 mIoU on LongVALE), confirming that explicit textual timestamps are necessary for precise temporal grounding. Adding more sophisticated mechanisms (time-local attention or cross-modal positional encoding) yields only marginal improvements (≤ 0.4 mIoU), all within the standard deviation of the plain timestamp variant. This indicates that plain textual timestamps combined with interleaving are both necessary and sufficient for temporal anchoring, and that the LLM’s inherent attention mechanism already captures cross-modal temporal correspondence effectively when given explicit timestamp anchors.

C.5 Normalization Analysis

We explore various normalization strategies within the fusion module to stabilize the joint training process (tables 16 and 17). Our empirical findings suggest that applying RMSNorm exclusively to the query (Pre-Q) provides the optimal balance of training stability and representation power.

Compared to a vanilla setup without normalization, the Pre-Q scheme significantly stabilizes training gradients, which is crucial for high IoU

Table 9: Extended ablation study on component effectiveness for the temporal grounding task. Recall@1 at different IoU thresholds and mean IoU (mIoU) are reported. TT, IA, and TSF denote Textual Timestamps, Interleaved Arrangement, and Two-stage Fusion, respectively.

Configuration	LongVALE				VALOR-32K			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Baseline	12.8	7.2	3.1	7.8	14.5	6.8	2.9	9.4
+ TT	15.6	9.4	4.2	11.5	19.8	10.2	5.1	14.2
+ TT + IA	24.2	14.8	6.5	16.2	31.4	18.5	9.4	21.5
+ TT + IA + TSF (Ours)	41.3	27.6	15.5	26.4	39.3	38.7	38.0	35.1

Table 10: Extended ablation study on component effectiveness for the segment captioning task. TT, IA, and TSF denote Textual Timestamps, Interleaved Arrangement, and Two-stage Fusion, respectively.

Configuration	LongVALE				VALOR-32K			
	B@4	R-L	M	C	B@4	R-L	M	C
Baseline	4.1	18.2	8.5	20.5	1.2	15.4	8.2	6.8
+ TT	4.8	19.8	9.2	22.1	2.4	18.6	10.1	12.4
+ TT + IA	5.4	21.5	10.4	25.4	5.2	23.5	12.4	25.6
+ TT + IA + TSF (Ours)	6.3	23.3	12.0	28.5	7.2	28.5	13.8	38.2

grounding results (R@0.7). Interestingly, adding normalization to the context (Pre-Q + Pre-C) or applying post-attention normalization leads to a slight performance degradation. We hypothesize that normalizing context tokens might over-compress the rich modality-specific features, whereas our selective Pre-Q approach ensures precise query-key interaction without losing nuanced cross-modal information.

C.6 Training Strategy and Dataset Statistics

Finally, we investigate the impact of our single-stage joint training strategy by comparing it against two-stage sequential SFT alternatives (tables 18 and 19). The results clearly show that our single-stage joint approach achieves superior results across both tasks and datasets.

Dataset Statistics To ensure reproducibility and transparency, we summarize the statistics of the primary datasets utilized in our training and evaluation phases. For the **LongVALE** dataset, our data split adheres to the standard configuration comprising approximately 120,000 temporally grounded training examples, 1,500 validation examples, and 1,000 test examples. For the **VALOR-32K** dataset, the split consists of 25,000 training videos, 3,500 validation videos, and 3,500 test videos. For the zero-shot evaluation benchmarks (e.g., **Music-AVQA**, **AVSD**, **Video-MME**), we strictly utilized their official standard test sets, containing between 1,000 and 5,000 evaluation instances per dataset. No

samples from these zero-shot evaluation sets were exposed during the training phase.

This reinforces the concept of mutual reinforcement between grounding and captioning: precise temporal localization provides more accurate context for semantic descriptions, while rich semantic reasoning helps the model better define fine-grained temporal boundaries. Training on both tasks simultaneously allows the model to leverage this synergy, leading to significant gains—for example, a +4.9% increase in mIoU on VALOR-32K (Liu et al., 2025) compared to the best sequential strategy.

D Statistical Significance Analysis

Table 20 reports the mean and standard deviation over three independent runs with different random seeds for the key comparisons between SynAV and the Naive Concat baseline under the same Qwen3-8B backbone. All gains are statistically significant at $p < 0.01$ (paired t -test).

E Modality Perturbation Robustness

To further evaluate how our model mitigates vision-centric bias in challenging, modality-imbalanced scenarios, we design a controlled perturbation study. Specifically, we apply two types of degradation to the LongVALE validation set and the Music-AVQA benchmark: (1) **Visual Masking (V-Mask)**, where 50% of the spatial patches in the video frames are replaced with Gaussian noise

Table 11: Extended comparison of different token arrangement strategies for the temporal grounding task.

Strategy	LongVALE				VALOR-32K			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Serial [Video; Audio]	26.8	16.4	7.2	18.2	32.5	19.4	10.2	22.5
Serial [Audio; Video]	25.4	15.2	6.8	17.5	30.8	18.2	9.4	21.0
Interleaved (Ours)	41.3	27.6	15.5	26.4	39.3	38.7	38.0	35.1

Table 12: Extended comparison of different token arrangement strategies for the segment captioning task.

Strategy	LongVALE				VALOR-32K			
	B@4	R-L	M	C	B@4	R-L	M	C
Serial [Video; Audio]	5.1	20.8	10.2	24.1	4.2	20.5	11.2	18.9
Serial [Audio; Video]	4.9	20.2	9.8	23.8	3.8	19.6	10.8	17.4
Interleaved (Ours)	6.3	23.3	12.0	28.5	7.2	28.5	13.8	38.2

Table 13: Temporal anchoring variants. LV = LongVALE, VA = VALOR. Time-local = time-local cross-attention; Cross-modal = cross-modal real-time positional encoding. All use the same backbone and fusion configuration; only the temporal alignment mechanism differs.

Variant	LV		VA	
	mIoU	CIDEr	mIoU	CIDEr
None	17.9	20.8	23.1	26.4
Plain (Ours)	26.4	28.5	35.1	38.2
+ Time-local	26.6	28.7	35.0	38.5
+ Cross-modal	26.5	28.3	35.3	38.1

to simulate severe visual impairment, and (2) **Audio Masking (A-Mask)**, where the audio stream is entirely muted. As shown in Table 21, the Naive Concat Baseline exhibits catastrophic performance drops under visual masking while remaining largely insensitive to audio masking, confirming its deeply rooted vision-centric bias. Conversely, SynAV maintains robust performance when visual cues are compromised, relying on the intact audio stream to anchor the events, and demonstrates a substantial performance dependency on the audio stream, directly proving that our architecture achieves a balanced modality reliance.

F Training Implementation Details

This section provides comprehensive details on the training configurations, models, and datasets.

For our model components, we employ SigLIP-2 (Tschanen et al., 2025) as the visual encoder and the Audio Transformer (AuT) from Qwen3-Omni (Xu et al., 2025) as the audio encoder. The text backbone is based on Qwen3-VL (Bai

et al., 2025). Our training follows a single-stage end-to-end approach, optimized using a unified autoregressive Cross-Entropy (CE) loss: $\mathcal{L} = -\sum \log P(s_i | s_{<i}, \mathbf{H}_v, \mathbf{H}_a)$, which covers both temporal grounding and fine-grained captioning. We jointly train the model on a composite dataset including LongVALE (Geng et al., 2025), VTimeLLM (Huang et al., 2024), and VALOR-32K (Liu et al., 2025).

During training, SynAV and all adapter layers (including fusion modules and projectors) are unfrozen, while all modality encoders and the text embedding layer remain frozen to preserve their pre-trained representational power. We use the AdamW optimizer with a cosine learning rate scheduler, starting from a peak learning rate of 2×10^{-5} and decaying to 1×10^{-6} . All experiments are conducted on 32 NVIDIA H800 (80GB) GPUs with a total batch size of 32 (1 sample per GPU), utilizing Fully Sharded Data Parallel (FSDP) for memory efficiency. To ensure robustness, all quantitative results reported in this paper are the average of three independent runs with different random seeds.

F.1 Details of Training Data Construction

As mentioned in the main paper, our model is jointly trained on a composite dataset including LongVALE (Geng et al., 2025), VTimeLLM (Huang et al., 2024), and VALOR-32K (Liu et al., 2025).

For LongVALE (Geng et al., 2025) and VTimeLLM (Huang et al., 2024), the original annotations are naturally aligned with our proposed fine-grained grounding and captioning training objective. Specifically, their structured event-level

Table 14: Impact of different modality combinations on the temporal grounding task.

Modality	LongVALE				VALOR-32K			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Video Only	25.2	14.8	6.4	17.1	32.4	19.8	10.4	22.4
Audio Only	4.8	2.5	1.2	4.2	6.2	3.8	1.8	5.1
Video + Audio (Ours)	41.3	27.6	15.5	26.4	39.3	38.7	38.0	35.1

Table 15: Impact of different modality combinations on the segment captioning task.

Modality	LongVALE				VALOR-32K			
	B@4	R-L	M	C	B@4	R-L	M	C
Video Only	5.2	21.2	10.8	24.5	5.8	24.2	12.1	26.8
Audio Only	1.4	9.2	4.5	8.6	1.8	10.5	5.2	10.4
Video + Audio (Ours)	6.3	23.3	12.0	28.5	7.2	28.5	13.8	38.2

descriptions and precise temporal boundaries are directly integrated into our interleaved audio-visual token arrangement without further semantic modification.

For VALOR-32K (Liu et al., 2025), although it provides rich fine-grained annotations, its original data format requires minor adjustments to fit our unified generative framework. We extract the segment-level captions and their corresponding timestamps, reformatting them into the explicit textual timestamp sequences ($\langle \text{time}_k \rangle$) used in our training pipeline.

Furthermore, we follow a rigorous data splitting protocol for VALOR-32K (Liu et al., 2025) to ensure fair evaluation. A specific subset is reserved exclusively for testing, while the remainder is used for training. All experimental results reported for VALOR-32K (Liu et al., 2025) in the main paper are based on this dedicated test set, which remains strictly mutually exclusive from the training subset. This ensures that the model’s reported performance reflects its genuine generalization capability on synchronized audio-visual understanding.

G Evaluation Setup

In this section, we provide the specific configuration details for evaluating our model on various tasks, including the prompts used and the decoding strategies applied.

G.1 Evaluation Prompts

We use specific prompts for evaluating our model on various tasks, including temporal grounding, segment captioning, and dense video captioning. To ensure a fair comparison, the prompt templates

for these tasks follow the exact settings used in LongVALE (Geng et al., 2025). The templates are shown in fig. 3.

G.2 Decoding Configurations

For all inference tasks, we employ a sampling-based decoding strategy to enhance the diversity and naturalness of the generated descriptions. Following the configuration of our underlying language model, we set the temperature to 0.7, top- p to 0.8, and top- k to 20. The repetition penalty is maintained at 1.0.

H Extended Qualitative Results

H.1 Calculation of Audio-Visual Attention Weights

To visualize the contribution of different modalities and provide a unified interpretation of the attention distributions presented in both the main paper and this supplementary material, we derive a one-dimensional temporal attention distribution from the original two-dimensional attention matrix of our proposed SynAV.

Specifically, we focus on the self-attention weights in the 24th layer of SynAV, which represents the high-level semantic integration stage for generating dense event descriptions and precise temporal boundaries.

Let $\mathcal{H}_{\text{suffix}} = \{\mathbf{h}_i^{\text{suffix}}\}_{i=1}^{N_s}$ be the set of text tokens in the suffix sequence (representing the generated captions, dense descriptions, or grounding results), and $\mathcal{S}_{\text{mm}}$ be the sequence of interleaved multimodal tokens. According to our arrangement strategy, the multimodal sequence is organized as:

$$\mathcal{S}_{\text{mm}} = [\dots, \langle \text{time}_k \rangle, \tilde{\mathbf{h}}_v^{(k)}, \hat{\mathbf{h}}_a^{(k)}, \dots] \quad (8)$$

Table 16: Ablation of normalization strategies within the fusion module for the temporal grounding task.

Normalization	LongVALE				VALOR-32K			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
None	34.2	21.5	10.8	22.8	31.5	24.8	18.2	28.6
Pre-Q + Pre-C	38.5	25.4	13.2	25.1	36.4	32.5	28.4	32.4
Post	32.8	20.4	9.5	21.5	29.8	22.5	15.4	26.9
Pre-Q + Pre-C + Post	37.4	24.8	12.8	24.8	35.8	31.8	27.2	31.9
Pre-Q (Ours)	41.3	27.6	15.5	26.4	39.3	38.7	38.0	35.1

Table 17: Ablation of normalization strategies within the fusion module for the segment captioning task.

Normalization	LongVALE				VALOR-32K			
	B@4	R-L	M	C	B@4	R-L	M	C
None	5.2	20.4	9.8	24.1	5.1	22.8	10.5	25.4
Pre-Q + Pre-C	5.8	22.1	11.2	27.2	6.4	26.5	12.4	34.1
Post	4.9	19.8	9.4	23.4	4.8	21.5	10.1	23.8
Pre-Q + Pre-C + Post	5.7	21.8	11.0	26.9	6.2	26.1	12.1	33.5
Pre-Q (Ours)	6.3	23.3	12.0	28.5	7.2	28.5	13.8	38.2

where $\tilde{\mathbf{h}}_v^{(k)}$ and $\hat{\mathbf{h}}_a^{(k)}$ are the visual patch tokens and audio segment tokens corresponding to the k -th timestamp.

For each suffix token $\mathbf{h}_i^{\text{suffix}}$, let $\mathbf{A}_{i,j}^{(L=24)}$ denote its attention weight towards the j -th multimodal token $\mathbf{m}_j \in \mathcal{S}_{\text{mm}}$ in the 24th layer. The raw attention map is a matrix $\mathbf{A} \in \mathbb{R}^{N_s \times N_{\text{mm}}}$. We collapse this into a temporal distribution by averaging over the suffix tokens:

$$\bar{a}_j = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{A}_{i,j}^{(24)} \quad (9)$$

To obtain continuous temporal distributions for both auditory and visual modalities, we aggregate these weights using a sliding window of 5 seconds. Let T be the total duration of the video. We define the modality-specific time-windowed attention weights $\mathcal{W}_a(t)$ and $\mathcal{W}_v(t)$ for a time interval $[t, t + \Delta t)$ where $\Delta t = 5s$:

$$\mathcal{W}_m(t) = \sum_{j: \text{time}(\mathbf{m}_j) \in [t, t+5s), \mathbf{m}_j \in \mathcal{M}_m} \bar{a}_j, \quad m \in \{a, v\} \quad (10)$$

where $\text{time}(\mathbf{m}_j)$ denotes the timestamp associated with the j -th multimodal token, and $\mathcal{M}_a, \mathcal{M}_v$ represent the sets of audio and visual tokens, respectively. This aggregation ensures that all audio and visual tokens within the same 5-second window contribute to their corresponding local attention scores, providing a clear indication of which temporal segments and modalities the model priori-

tizes during the generation of fine-grained captions, dense descriptions, and grounding outputs.

H.2 Qualitative Analysis

In this section, we provide qualitative examples including the one discussed in the main paper (fig. 4) and two additional cases (figs. 5 and 6). These results further demonstrate the effectiveness of our SynAV in handling complex multimodal scenarios.

The first case (fig. 5) highlights a scenario of audio-visual synergy required for fine-grained understanding. While the visual modality provides a general overview of a woman in a blue blazer or blouse conducting interviews, which persists across many segments, the model must integrate specific audio cues to correctly distinguish and ground the exact event among multiple visually similar segments. As illustrated, while the baseline model provides a fragmented description and suffers from “visual generalization” (leading to low IoU), SynAV accurately captures the nuanced audio-driven events and achieves precise grounding by anchoring the visual perception with synchronized audio peaks.

The second case (fig. 6) demonstrates another critical challenge where over-reliance on the visual modality leads to errors. In this video, there are adjacent and continuous segments that are visually very similar but correspond to different audio-visual events. A vision-centric approach tends to focus on these neighboring segments incorrectly, causing both temporal grounding and captioning

Table 18: Comparison of different training strategies for the temporal grounding task.

Strategy	LongVALE				VALOR-32K			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Two-stage SFT ($G \rightarrow C$)	35.4	22.8	11.5	23.2	32.8	25.4	18.5	27.4
Two-stage SFT ($C \rightarrow G$)	36.8	24.1	12.2	24.5	34.2	28.6	21.8	30.2
Single-stage Joint SFT (Ours)	41.3	27.6	15.5	26.4	39.3	38.7	38.0	35.1

Table 19: Comparison of different training strategies for the segment captioning task.

Strategy	LongVALE				VALOR-32K			
	B@4	R-L	M	C	B@4	R-L	M	C
Two-stage SFT ($G \rightarrow C$)	5.8	22.4	11.5	27.1	6.2	25.8	11.8	32.4
Two-stage SFT ($C \rightarrow G$)	5.5	21.2	10.8	25.8	5.8	24.5	11.2	30.1
Single-stage Joint SFT (Ours)	6.3	23.3	12.0	28.5	7.2	28.5	13.8	38.2

tasks to fail. By contrast, our method achieves robust audio-visual synergy, allowing it to correctly identify the target event and produce precise results where visual cues alone are ambiguous.

H.3 AI Assistance in Writing

During the preparation of this manuscript, we utilized Gemini-3.1-pro-preview to polish the manually written initial drafts and to correct grammatical errors. All authors reviewed and approved the final version, and take full responsibility for the content of this publication.

Table 20: Mean $\pm$ standard deviation over three random seeds. Δ denotes the absolute gain of SynAV over Naive Concat. All p -values are from a paired t -test.

Method	LongVALE mIoU	VALOR mIoU	Music-AVQA Acc.
Naive Concat	18.2 $\pm$ 0.3	22.5 $\pm$ 0.4	35.4 $\pm$ 0.5
SynAV (Ours)	26.4 $\pm$ 0.4	35.1 $\pm$ 0.5	58.1 $\pm$ 0.4
Δ (gain)	+8.2	+12.6	+22.7
Significance	$p < 0.01$	$p < 0.01$	$p < 0.01$

Table 21: Robustness under modality perturbation. V-Mask applies severe noise to the visual stream, and A-Mask mutes the audio stream. Unperturbed scores are 18.2 (mIoU) / 35.4% (Acc) for Naive Concat, and 26.4 / 58.1% for SynAV. Values in parentheses denote the absolute performance drop.

Model	LongVALE (mIoU $\downarrow$)		Music-AVQA (Acc. $\downarrow$)	
	V-Mask	A-Mask	V-Mask	A-Mask
Naive Concat	8.5 (−9.7)	17.1 (−1.1)	14.2 (−21.2)	28.2 (−7.2)
SynAV (Ours)	16.8 (−9.6)	17.1 (−9.3)	42.6 (−15.5)	29.5 (−28.6)

Task: Temporal Grounding

At which time interval can we find <enent> taking place in the video?
Give the timestamps in the fromat: From xx to xx.

Task: Dense Captioning

Could you please detail the events that took place during different time segments in the video? List the events in the format:
From xx to xx, event1.
From xx to xx, event2.
...

Task: Segment Captioning

Can you describe what occurred from <start> to <end> in the video?
Please give the event description directly.

Figure 3: Specific prompt templates used for different tasks: Temporal Grounding, Dense Captioning, and Segment Captioning.

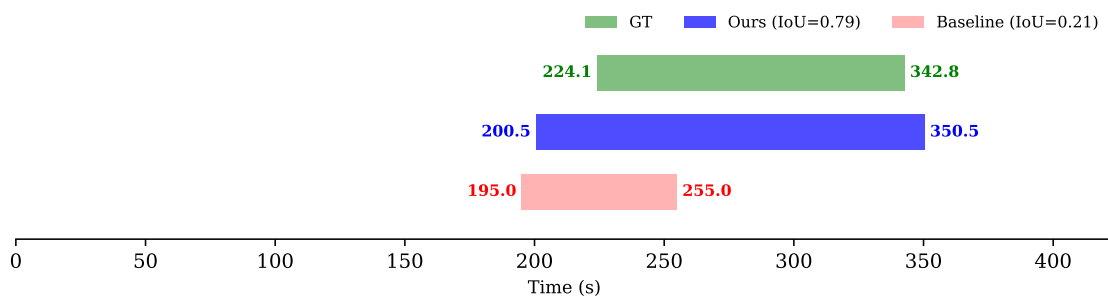


Question: Describe the video content in the [224.1, 342.8] time interval.

Baseline Answer: "The camera pans across a peaceful landscape of fields and trees, then focuses on a window with a colorful curtain before settling on a weathered wooden door. A woman speaks in a foreign language, describing the garden and a nearby building used for bathing."

Ours Answer: "The camera moves from a peaceful landscape into a house, showing a shoe rack and a cozy room. The narrator describes the kitchen and the guest room, highlighting the Kyrgyz tradition of hospitality."

Qualitative Result of Temporal Grounding



Cross-Modal Attention Temporal Distribution (Interval: 5s)

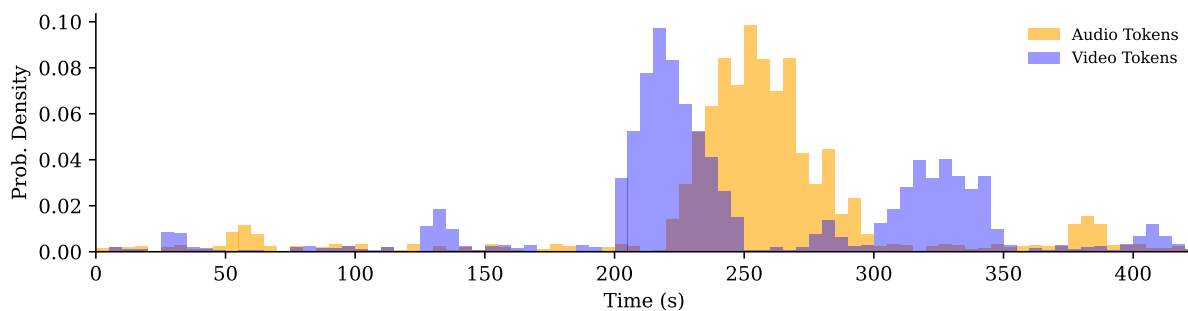


Figure 4: Qualitative results of our framework discussed in the main paper. **Top:** Comparison of audio-dependent QA results between Video-LLaVA (Lin et al., 2024) and our model. **Middle:** Visualization of predicted temporal grounding boundaries compared with the ground truth. **Bottom:** Temporal distribution of cross-modal attention weights during the reasoning process.

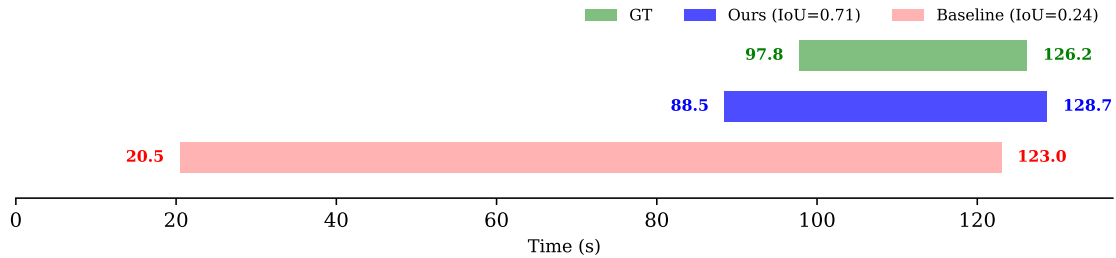


Question: Describe the video content in the [97.8, 126.2] time interval.

Baseline Answer: "A female host is presenting in the video, with scenes of a gathering crowd interspersed throughout. She is dressed in blue clothing, holding a microphone, and smiling warmly while interacting with the audience. In the background, other women in traditional attire can be seen clapping and greeting one another, creating a lively and joyful atmosphere."

Ours Answer: "The video shows a group of women in colorful clothes joining hands, followed by a woman in a blue blazer talking against a plain wall. She is then seen on a beach applying lotion and later standing in front of a beige wall holding a microphone. Finally, the scene returns to the group where she is holding a plate of food while people clap. The content appears to be a general lifestyle interview based on the changing backgrounds and gestures."

Qualitative Result of Temporal Grounding



Cross-Modal Attention Temporal Distribution (Interval: 5s)

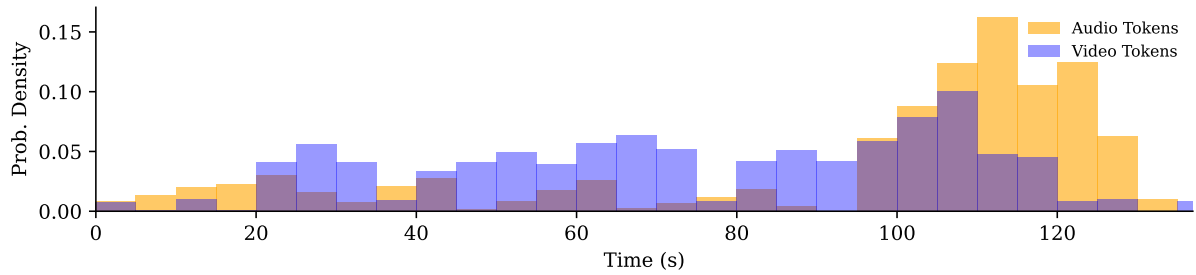


Figure 5: Qualitative results of our framework (the first additional case in the supplementary material), with the same layout as the case in the main paper. **Top:** Comparison of audio-dependent QA results between Video-LLaVA (Lin et al., 2024) and our model. **Middle:** Visualization of predicted temporal grounding boundaries compared with the ground truth. **Bottom:** Temporal distribution of cross-modal attention weights during the reasoning process.

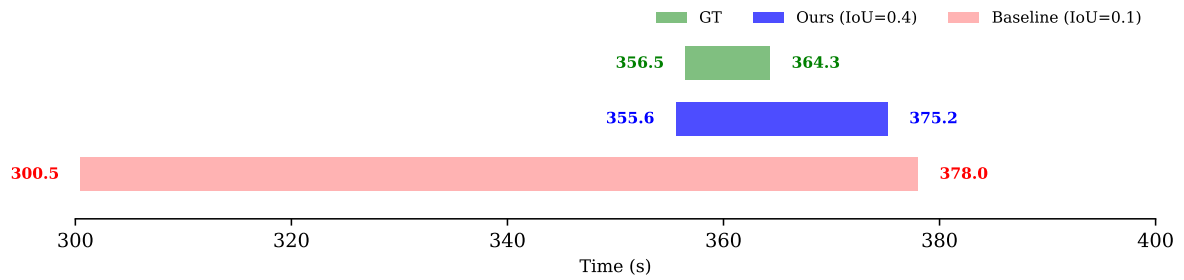


Question: Describe the video content in the [356.5, 364.3] time interval.

Baseline Answer: "The video showcases an amusement park featuring orange and blue roller coasters (including the Hysteria coaster) climbing steep inclines and twisting through lush greenery, a large yellow roller coaster, a sunlit Ferris wheel with passengers, a blue water slide with people splashing down, and a man on a boat in the water park, all set against clear blue skies and vibrant green surroundings."

Ours Answer: "A man on a whimsical boat smiles at the camera as upbeat music plays, the vessel gracefully gliding through the water with the lively atmosphere of the water park in the background contrasting with the calmer, almost magical scene in the foreground."

Qualitative Result of Temporal Grounding



Cross-Modal Attention Temporal Distribution (Sum=0.80)

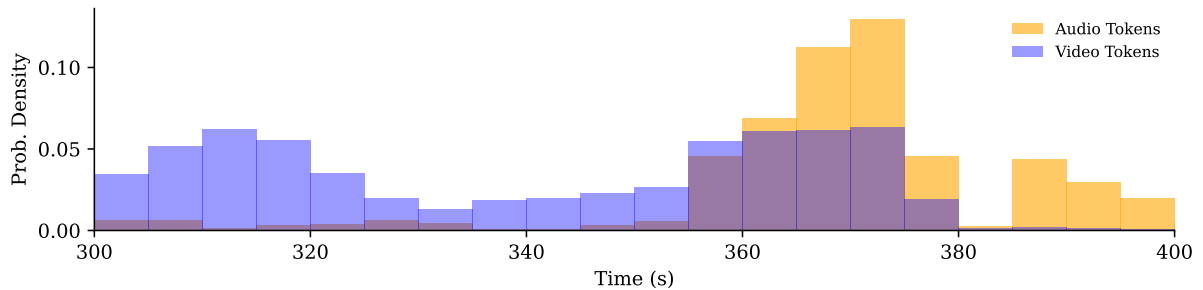


Figure 6: Qualitative results of our framework (the second additional case in the supplementary material), with the same layout as the case in the main paper. **Top:** Comparison of audio-dependent QA results between Video-LLaVA (Lin et al., 2024) and our model. **Middle:** Visualization of predicted temporal grounding boundaries compared with the ground truth. **Bottom:** Temporal distribution of cross-modal attention weights during the reasoning process.