LatentOmni: Rethinking Omni-Modal Understanding via Unified Audio-Visual Latent Reasoning

Yifan Dai[1,4], **Zhenhua Wu**[2], **Bohan Zeng**[3], **Daili Hua**[3], **Jialing Liu**[6], **Bozhou Li**[3], **Yuran Wang**[3], **Chengzhuo Tong**[3], **Hao Liang**[3], **Xiaochen Ma**[7], **Junbo Niu**[3], **Tianyu Guo**[3], **Yang Shi**[3], **Yue Ding**[8], **Yiyan Ji**[9], **Bingyin Mei**[10], **Yushuo Guan**[2], **Yuanxing Zhang**[2], **Pengfei Wan**[2], **Fangcheng Fu**[1], **Wentao Zhang**[3,4,5 *]
[1]School of Artificial Intelligence, Shanghai Jiao Tong University,
[2]KlingAI Research, [3]Peking University, [4]Zhongguancun Academy,
[5]Beijing Key Laboratory of Data Intelligence and Security (Peking University)
[6]Renmin University of China, [7]HKUST, [8]CASIA, [9]Nanjing University, [10]Tsinghua University

Abstract

While joint audio-visual understanding is fundamental to advancing machine cognition, current multimodal large language models (MLLMs) still struggle with complex cross-modal reasoning. Existing text-based chain-of-thought (CoT) compresses rich multi-modal features into discrete text, incurring information loss and inducing a language-bound phenomenon that diminishes attention to audio-visual signals. In contrast, a continuous latent space inherently preserves dense representations, serving as an ideal carrier for audio-visual information. Motivated by this, we propose **LatentOmni**, a novel cross-modal reasoning framework. By introducing a feature-level supervision mechanism to directly reconstruct raw sensory inputs within the latent space, LatentOmni leverages native latent features to bridge audio-visual modalities and text, ensuring sustained attention on original audio-visual inputs throughout reasoning. Furthermore, to maintain temporal consistency across modalities in latent space, we design Omni-Sync Position Embedding (OSPE), which generalizes multimodal rotary position encodings to drive audio-visual synchrony. To supervise this reasoning process, we construct LatentOmni-Instruct-35K, a dataset interleaving text with audio-visual segments that serve as dense evidence for latent reconstruction. Comprehensive evaluation across multiple audio-visual reasoning benchmarks demonstrates that LatentOmni substantially outperforms strong explicit-CoT baselines, validating latent space joint reasoning as a promising path toward genuine omnimodal understanding.

1 Introduction

Information in the real world is inherently multimodal (1; 2), with vision and audio serving as the two most fundamental channels through which humans perceive and understand their surroundings (3; 4; 5). The joint comprehension of audio-visual information therefore constitutes a prerequisite for artificial agents to achieve genuine cognition of the external world (6). Central to this process are the capacities for perception and reasoning (7; 8): the former extracts meaningful features from visual and auditory signals, while the latter builds upon these representations to perform cross-modal semantic integration and higher-level cognition.

Although current multimodal large language models (MLLMs) have shown considerable progress in perceptual tasks such as captioning and grounding (9; 10; 11; 12; 13; 14; 15; 16), their capacity for complex reasoning remains constrained (17; 18), especially in cross-modal audio-visual scenarios. A

*Corresponding author: wentao.zhang@pku.edu.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

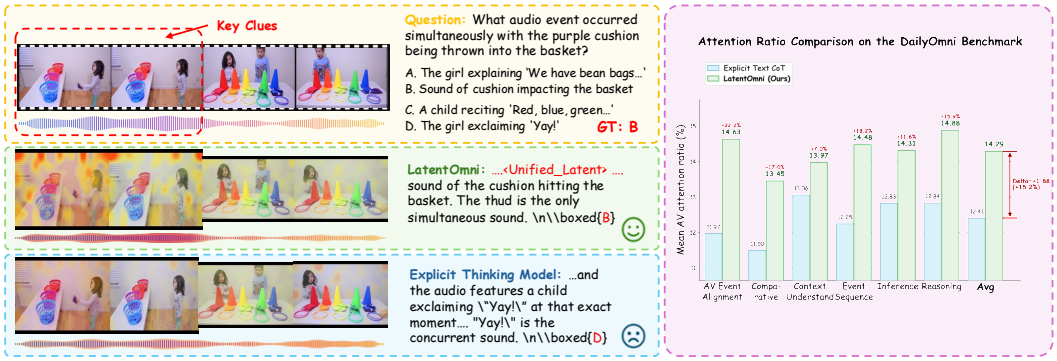


Figure 1: Comparison between LatentOmni and the Explicit Text CoT baseline (detailed in 4.1). (Left) Qualitatively, unlike the baseline, LatentOmni accurately anchors on key audio-visual (AV) clues (indicated by heatmaps) to answer correctly. (Right) Quantitatively, it maintains a significantly higher AV token attention ratio across tasks on the Daily-Omni benchmark, ensuring robust grounding of original modalities.

critical bottleneck lies in the transition from perception to reasoning. While existing research usually introduces explicit or structured chain-of-thought to enhance omni-reasoning capabilities (19; 20; 21; 22), these mechanisms inherently compress high-dimensional, continuous perceptual information into discrete text tokens during the reasoning phase. This forced mapping inevitably causes feature loss and information decay. Consequently, as illustrated in Figure 1, pure-text explicit CoT tends to over-rely on textual inputs and exhibits insufficient attention to native audio-visual signals during cross-modal reasoning. This hinders the effective utilization of fine-grained inter-modal details (e.g., temporal synchronization), thereby restricting the model's comprehensive reasoning capabilities.

Unlike discrete text tokens, the latent space remains continuous and information-rich. This makes latent embeddings an optimal medium for retaining multimodal signals. Building on this intuition, we propose LatentOmni, a novel framework that equips MLLMs with joint audio-visual reasoning capabilities within a unified latent space. We design a dedicated post-training pipeline to develop this latent reasoning capability. To ensure the model attends to raw audio-visual signals and preserves their fine-grained information within the latent space during reasoning, we introduce a feature-level supervision mechanism. This mechanism aligns pre-annotated dense representations of reasoning-relevant audio-visual segments with the intermediate latent embeddings. It forces the model to actively revisit the source modalities and autoregressively generate reasoning-relevant audio-visual features within the latent space. To further ensure temporal coherence between auditory and visual signals, we propose the Omni-Sync Position Embedding (OSPE). By extending the time-aligned multimodal RoPE (23) to the unified latent space, OSPE explicitly enforces cross-modal temporal alignment. As illustrated in Fig. 1, compared to purely explicit CoT reasoning methods, LatentOmni substantially improves attention to the original audio-visual (AV) modalities, particularly on AV alignment tasks. While anchoring key information from the raw AV modalities, these latent embeddings serve as a robust semantic bridge connecting visual, auditory, and textual information, thereby enabling advanced cross-modal joint reasoning.

Implementing feature-level supervision within the latent space necessitates CoT data with pre-annotated, reasoning-relevant audio-visual segments, which are largely absent from existing datasets. To fill this gap, we develop a scalable data curation pipeline featuring audio-video interleaved reasoning trajectory and construct LatentOmni-Instruct-35K, a high-quality dataset specifically tailored for cross-modal reasoning tasks. Extensive experiments demonstrate that LatentOmni significantly outperforms a number of models using explicit cot across multiple audiovisual reasoning benchmarks. In brief, our contributions are summarized as follows:

- We propose **LatentOmni**, an novel audio-visual reasoning framework that equips MLLMs with a tailored post-training pipeline to conduct joint reasoning in a unified latent space.
- We introduce explicit feature-level supervision in latent space and Omni-Sync Position Embedding (OSPE) to facilitate cross-modal temporal alignment, which efficiently preserves attention to audio-visual modalities and bridges audio-visual with textual semantics.

- We develop a novel audio-visual interleaved CoT data synthesis pipeline, and construct **LatentOmni-Instruct-35K**, a high-quality dataset filling the gap in tailored training data for complex cross-modal latent reasoning.
- Our extensive experiments show that LatentOmni substantially outperforms numerous models using explicit CoT and achieves competitive performance on challenging benchmarks, confirming its substantial promise.

2 Related work

2.1 Multimodal Large Language Models Reasoning

Multimodal Large Language Models (MLLMs) originally aimed to equip LLMs with diverse perceptual capabilities (24; 25; 26; 27; 28); however, to tackle complex real-world tasks, research has progressively shifted toward enhancing their reasoning abilities. A prevailing paradigm to achieve this is leveraging explicit Chain-of-Thought (CoT) techniques (20; 21; 29; 30). By establishing text as the primary semantic bridge for cross-modal integration, these models can effectively decompose complex tasks via natural language rationales (31). This text-centric reasoning approach has demonstrated encouraging progress in individual visual and audio domains, and has now naturally extended to drive recent omnimodal frameworks like Gemini (32), Video-LLaMA series (33), and the Qwen-Omni series (23).

Despite its widespread adoption, recent research reveals that this discrete reasoning paradigm fundamentally constrains complex cross-modal inference (34; 35; 36; 37). Forcing high-dimensional audio-visual signals through a narrow textual bottleneck inevitably causes information loss. Furthermore, this text-centric abstraction results in insufficient attention to raw audio-visual signals. This imbalance leads to sensory detachment and multimodal hallucinations, where generated rationales decouple from the actual underlying evidence (38; 39). Although recent tool-augmented approaches (e.g., think with audio, image and video) (40; 41; 42) attempt to mitigate this, they fail to fundamentally resolve the inherent neglect of cross-modal inputs. Consequently, these limitations severely impede the scalability of explicit CoT reasoning (43).

2.2 Reasoning in Latent Space

To mitigate the constraints of discrete token generation, recent studies have explored conducting reasoning directly within continuous latent spaces (44; 45; 46). As a pioneering work in this direction, Coconut (45) bypasses the autoregressive generation of intermediate textual tokens by executing reasoning steps entirely within the model's hidden states. This continuous reasoning paradigm has subsequently been extended to the multimodal domain to better accommodate continuous real-world sensory signals (47). In this context, current research generally follows two mainstream methodologies: some works design specific training frameworks (48; 49; 50) to optimize reasoning trajectories within the latent space, while others develop training-free inference mechanisms (51) to elicit latent reasoning capabilities directly from pre-trained representations.

Despite these advances, existing latent reasoning methods predominantly focus on pure text or single-modality extensions, such as visual-textual integration (49; 48; 51; 52). The joint comprehension and reasoning of dynamic Audio-Visual (AV) signals within a unified continuous space remains underexplored. Recognizing this gap, our work introduces LatentOmni to extend continuous latent reasoning to omnimodal scenarios, explicitly addressing the temporal and semantic alignment of cross-modal AV integration.

3 LatentOmni

We propose LatentOmni, a novel architecture that performs audio-visual understanding within a unified latent space. To realize this capability, we carefully design an Audio-Video Interleaved Chain-of-Thought data synthesis pipeline alongside an efficient training strategy tailored for omnimodal latent reasoning. In this section, we elaborate on the reasoning process of LatentOmni, the synthesis of training data, and the corresponding training strategy.

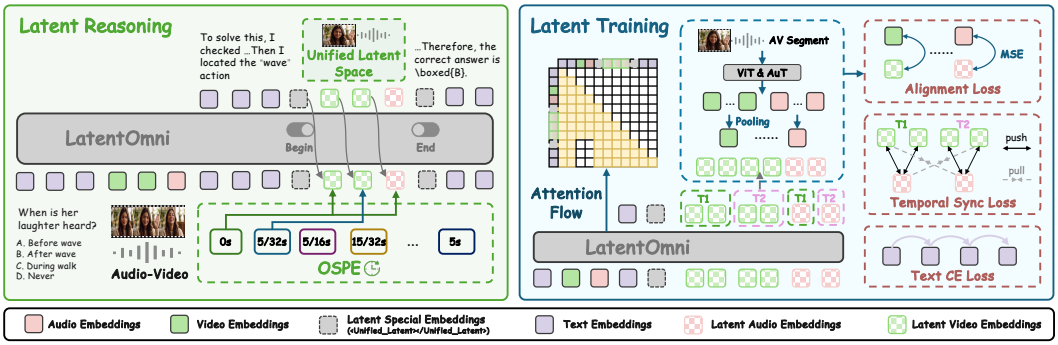


Figure 2: LatentOmni Overview. Left: A start embedding automatically triggers latent reasoning. The model autoregressively generates latent audio-visual embeddings, temporally synchronized via OSPE for precise cross-modal reasoning. Right: Training employs three specialized losses (Sec. 3.4) and a controlled attention flow, explicitly guiding the model to leverage latent representations for downstream reasoning.

3.1 Audio-Visual Latent Reasoning

Existing MLLMs usually rely on explicit, text-only chain-of-thought for reasoning, leading to inadequate attention allocation to audio-visual modalities as the reasoning chain lengthens. To overcome this bottleneck without losing the rich logical priors inherent in language, LatentOmni introduces a text-latent reasoning paradigm. Instead of a purely discrete or purely continuous trajectory, our model auto-regressively generates a hybrid sequence. Given the encoded visual features H^v, auditory features H^a, and the textual query H^q, the reasoning process seamlessly switches between explicit textual deduction and implicit cross-modal synthesis. Specifically, when the reasoning process requires revisiting the audio-visual information, the model emits a special token $<$ `Unified_Latent` $>$. This token acts as a trigger, transitioning the decoding process from the discrete vocabulary space $\mathcal{V}$ into a continuous, high-dimensional latent space $\mathbb{R}^d$. After generating a predefined number K of latent embeddings, we explicitly insert a stop token $<$ /`Unified_Latent` $>$ to terminate the continuous reasoning phase and revert to explicit textual generation. The resulting reasoning trajectory can be formulated as an interleaved sequence S, as illustrated in Fig. 2 left:

$$S = [w_{1:i}, u, z_{1:K}, u', w_{i+1:j}, u, z_{K+1:2K}, u', \ldots, a], \tag{1}$$

where w denotes standard text, u is the $<$ `Unified_Latent` $>$ trigger, u' represents the inserted $<$ /`Unified_Latent` $>$ stop token, z represents the continuous latent reasoning states, and a is the final definitive answer. This mechanism allows LatentOmni to utilize text for structural logic while leveraging the latent space for dense audio-visual joint reasoning, effectively bridging vision, audio, and language. We provide a comprehensive analysis of the effect of K in Section 4.3.

3.2 Unified Latent Representation and Temporal Alignment

During the latent reasoning phases triggered by u, the model generates a sequence of continuous states auto-regressively. At the k-th latent step, the latent representation $z_k \in \mathbb{R}^d$ is instantiated as the last-layer hidden state of the transformer backbone prior to the language modeling head (Fig. 2 left):

$$z_k = \mathrm{LM}_\theta^{(L)}\left(H^v, H^a, H^q, S_{<k}\right), \tag{2}$$

where L denotes the total number of transformer layers, and $S_{<k}$ represents the historical mixed context of previous text tokens and latent states. Each generated latent representation z_k is directly propagated as an input embedding to subsequent positions, thereby forming a continuous reasoning trajectory. This trajectory spans K steps, which is explicitly structured as a modality-ordered sequence. Specifically, the first K_v latent steps prioritize fine-grained visual reasoning, followed by K_a steps dedicated to auditory semantic extraction. Crucially, rather than mapping back to the discrete space $\mathcal{V}$, all generated z_k are inherently preserved within a shared, high-dimensional continuous space $\mathbb{R}^d$.

However, since visual and audio latent embeddings are generated sequentially, their natural temporal correspondence could be compromised. To guarantee temporal consistency between visual and

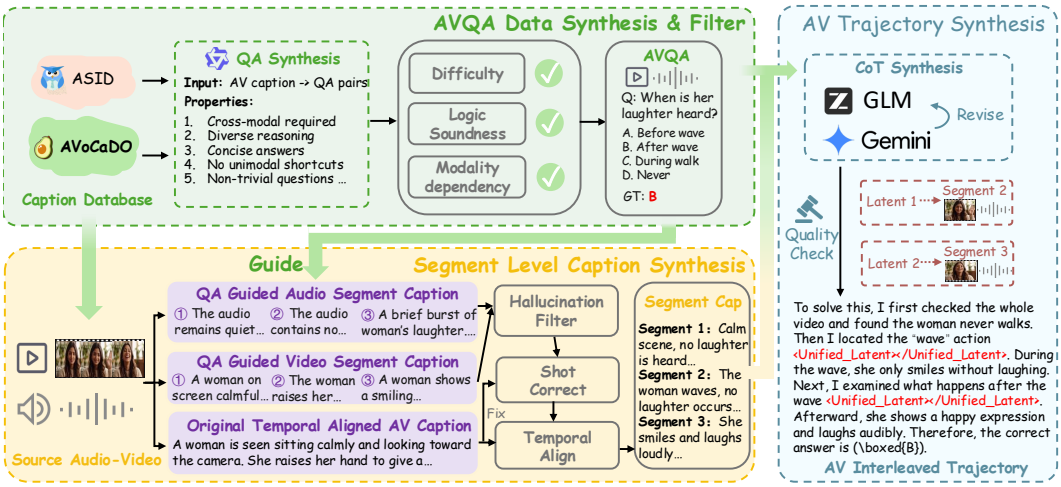


Figure 3: LatentOmni-Instruct-35K Dataset Construction Pipeline, including AVQA Data Synthesis & Filtering, Segment-Level Caption Synthesis, and AV-Interleaved Reasoning Trajectory Synthesis.

audio information within this unified latent space, we introduce the Omni-Sync Position Embedding (OSPE) by extending the Time-aligned Multimodal RoPE from Qwen2.5-Omni (23). OSPE assigns a continuous, shared physical timestamp t to temporally corresponding visual frames and audio segments. For a given unified latent feature $h \in \{h^v, h^a\}$ at timestamp t, OSPE applies a temporal rotary transformation:

$$\mathrm{OSPE}(h, t) = h \odot \cos(t\Theta) + \mathcal{R}(h) \odot \sin(t\Theta),$$

where h^v and h^a represent the latent visual feature and latent audio feature respectively, $\Theta = \{\theta_i\}_{i=1}^{d/2}$ represents the predefined base frequency vector, $\odot$ denotes the Hadamard product, and $\mathcal{R}(\cdot)$ is the block-diagonal rotation matrix operating on adjacent feature dimensions. By injecting this synchronized relative positional prior, sequentially generated latent features representing identical timestamps are rigorously aligned. This guarantees that the generation of z_k attends to strictly synchronized cross-modal evidence without temporal misalignment.

3.3 LatentOmni-Instruct-35K Dataset Construction

To effectively supervise LatentOmni's latent reasoning capabilities, specialized training data is required. Therefore, constructing these audio-video interleaved reasoning trajectories demands careful data preparation, synthesis, and filtering. As shown in Fig.3, our data synthesis pipeline comprises three stages: AVQA Data Synthesis & Filtering, Segment-Level Caption Synthesis, and AV-Interleaved Reasoning Trajectory Synthesis.

AVQA Data Synthesis & Filtering. This stage we collect raw data from two high-quality audio-visual caption datasets with strict temporal alignment, ASID (53) and AVoCaDO (9), and employs the Qwen3-235B-A22B (54) model to transform cross-modal captions into preliminary question-answer pairs. During generation, the model is explicitly instructed to produce questions that satisfy cross-modal information dependency, typological diversity, and answer accuracy. The GLM4.7 (55) model then annotates each pair by category and scores it along three dimensions, namely question difficulty, logic soundness, and modality dependency, each on a 1-to-5 scale, with samples scoring below 13 discarded. To ensure category balance, we additionally constrain the ratio between any two adjacent categories to be within $3\times$. Through this pipeline, a high-quality, unbiased, and logically rigorous AVQA dataset is ultimately constructed. The prompts used for data synthesis, quality scoring, and category annotation are provided in Appendices B.1 and Appendices B.2, respectively.

Segment Level Captions Synthesis. To precisely locate the audio-visual segments for interleaving reasoning trajectory, we first segment the raw audio and visual streams from the source data according to timestamps. Since directly generating joint audio-visual descriptions leads to information omissions from either the audio or visual modality (9), we introduce the Qwen3-Omni-30B-A3B-Captioner (54) to produce separate audio and video segment captions that are strongly correlated with specific QA

pairs. Using the original temporally aligned captions from the data source as a reference, we then apply GLM-4.7 to perform three tasks: filtering out AVQA pairs with high hallucination rates, fixing shot fragmentation caused by segmentation, and temporally aligning the audio and video captions. Through this multi-step pipeline, we construct high-quality AV Segment-Level Captions.

AV Interleaved Reasoning Trajectory Synthesis. Finally, leveraging the AVQA pairs and segment-level captions generated in the preceding stages, the GLM-4.7 model constructs the reasoning trajectories. When the reasoning process requires referencing a specific segment, the model inserts a corresponding marker as a placeholder for the audio-visual content. Subsequently, we employ Gemini2.5-Flash to audit and refine these generated trajectories, focusing on correcting citation errors and pruning redundant logical branches to improve the overall quality of the reasoning. Entries that exhibit logical paradoxes or significant hallucinations are excluded to preserve the rigor of the dataset. Ultimately, by substituting the markers with their respective audio-visual segments, a high-quality dataset comprising 35K AV interleaved reasoning trajectories is constructed.

3.4 LatentOmni Training

In this phase, we perform supervised fine-tuning (SFT) on LatentOmni utilizing the Audio-Video interleaved Chain-of-Thought (CoT) dataset constructed in Section 3.3. This stage aims to provide explicit supervision for latent reasoning, enabling LatentOmni to jointly model textual CoT steps and audio-visual latent embeddings within an autoregressive trajectory. To achieve this efficiently, we design a post-training pipeline driven by a joint optimization strategy.

To establish a temporally cohesive foundation prior to deep joint reasoning, we first explicitly enforce correspondence between the visual and auditory information in latent space using a **temporal synchronization objective ($\mathcal{L}_{\text{sync}}$)**. Given the synchronized sequence of latent visual features h_t^v and auditory features h_t^a at matching timestamps $t \in \mathcal{T}$, we optimize a symmetric InfoNCE contrastive loss:

$$\mathcal{L}_{\text{sync}} = -\frac{1}{2|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(\log \frac{\exp\left(\text{sim}(h_t^v, h_t^a)/\tau\right)}{\sum_{t'} \exp\left(\text{sim}(h_t^v, h_{t'}^a)/\tau\right)} + \log \frac{\exp\left(\text{sim}(h_t^a, h_t^v)/\tau\right)}{\sum_{t'} \exp\left(\text{sim}(h_t^a, h_{t'}^v)/\tau\right)} \right), \quad (3)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a learnable temperature. This objective encourages temporally co-occurring audio-visual token pairs to align tightly in the continuous space, while pushing apart asynchronous pairs.

Building upon this temporally synchronized space, our primary objective is to mitigate the language-bound phenomenon—a tendency where models over-rely on linguistic priors while neglecting native audio-visual signals during reasoning. To ensure sustained attention to the original AV modalities, we explicitly ground each auto-regressively generated latent embedding z_k in the raw perception space.

Specifically, each audio-visual segment in our interleaved dataset dictates a latent reasoning span of K steps. To construct the corresponding dense anchor sequence $A = [a_1, \ldots, a_K]$, we first extract audio-visual features from the segment using model's visual and audio encoders. We then compress these features into K_v visual and K_a auditory tokens ($K = K_v + K_a$) via a train-free L2-norm weighted pooling. This strategy preserves salient transient actions and acoustic events. As the reasoning unfolds autoregressively, we align each generated embedding z_k with its target anchor a_k via a latent alignment loss:

$$\mathcal{L}_{\text{latent}} = \frac{1}{K} \sum_{k=1}^{K} \|z_k - a_k\|_2^2 . \quad (4)$$

To ensure the model maintains its inherent linguistic priors while navigating the text-latent reasoning paradigm, we apply a standard **next-token prediction loss ($\mathcal{L}_{\text{text}}$)** over all discrete tokens within the hybrid sequence. Given the reasoning sequence $S = \{s_1, s_2, \ldots, s_L\}$ containing both discrete text tokens and continuous latent states, we compute the auto-regressive cross-entropy loss exclusively on the elements belonging to the vocabulary $\mathcal{V}$:

$$\mathcal{L}_{\text{text}} = -\frac{1}{N_{\text{text}}} \sum_{t=1}^{L} \mathbb{I}(s_t \in \mathcal{V}) \log p(s_t \mid S_{<t}, H^v, H^a, H^q), \quad (5)$$

where $\mathbb{I}(\cdot)$ is the indicator function, N_{text} is the total number of discrete tokens (including the text reasoning steps w, the trigger token u, and the final answer a) in the sequence, and $S_{<t}$ represents the

Table 1: Performance comparison of proprietary and open-source models on Daily-Omni, WorldSense, OmniVideoBench, and LVOmniBench benchmarks. The **best** is highlighted

Method	Daily-Omni	WorldSense	OmniVideoBench	LVOmniBench
Proprietary Models				
GPT-4o	56.5	42.6	-	-
Gemini-2.0-Flash	67.8	56.2	41.5	42.9
Gemini-2.5-Pro	81.4	64.6	58.9	-
Audio-Visual Understanding Models				
MiniCPM-o-7B	53.1	29.7	29.7	34.8
VITA-1.5-7B	44.7	36.9	30.5	-
HumanOmniV2-7B	58.5	47.1	30.5	32.3
Baichuan-Omni-1.5	50.0	43.3	30.7	32.8
OmniVinci	66.5	48.2	32.1	-
video-SALMONN 2+	67.1	48.0	34.3	33.1
video-SALMONN-o1	61.9	45.0	34.4	33.8
Different Base Models(LatentOmni)				
VideoLLaMA2-7B	35.2	25.4	29.2	27.0
+ Explicit Text CoT	40.1	31.3	28.9	27.5
+ Vanilla SFT	38.9	30.3	29.5	26.2
+ **LatentOmni**	42.0	33.8	30.7	28.9
Qwen2.5-Omni-7B	62.9	45.4	29.3	32.0
+ Explicit Text CoT	65.6	46.6	33.2	32.1
+ Vanilla SFT	62.0	47.5	30.5	33.2
+ **LatentOmni**	**67.4**	**48.9**	**35.4**	**35.1**

preceding hybrid context. This formulation ensures that each predicted text token s_t is dynamically conditioned on the rich, interleaved history of both explicit textual deductions and implicit cross-modal latent states.

The model is optimized end-to-end with the combined objective function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \lambda_1 \mathcal{L}_{\text{latent}} + \lambda_2 \mathcal{L}_{\text{sync}}, \tag{6}$$

where λ_1 and λ_2 are balancing hyperparameters. This joint optimization carefully weighs the implicit reasoning and synchronization terms against the primary text generation task, ensuring a robust alignment across modalities.

4 Experiment

4.1 Experiment Setup

Training. The entire training process of LatentOmni follows the pipeline described in Section 3.4. During implementation, we iteratively optimize the model for 750 steps (2 epochs), using the LatentOmni-Instruct-35K dataset described in Section 3.3. Prior work (48) indicates that fixed-token strategies outperform dynamic adjustments in practical scenarios. To determine the optimal fixed count, we conducted ablation studies on audio-visual allocation ratios and token quantities. Results show that 40 latent tokens yield the best performance. Therefore, we strictly maintain this configuration throughout training and testing.

Benchmarks. We evaluate audio-visual joint reasoning on four omnimodal benchmarks that stress complementary capabilities: everyday scenario reasoning (Daily-Omni (2)), physical and spatial-temporal commonsense (WorldSense (1)), cross-modal alignment and question answering (OmniVideoBench (17)), and long-form multi-sensory understanding (LVOmniBench (56)). This benchmark suite is intended to test whether latent reasoning helps beyond a single data regime: Daily-Omni emphasizes common event understanding, WorldSense tests structured commonsense over time and space, OmniVideoBench contains fine-grained audio-type and video-duration splits, and LVOmniBench stresses sustained reasoning over longer inputs.

Table 2: Accuracy comparison of LatentOmni and other methods on OmniVideoBench. The **best** is highlighted and the second-best is underlined. The performance gain of our method over the base model is indicated in red parentheses.

Method	Audio Type			Video Duration				Avg.
	Music	Sound	Speech	(0,1] min	(1,5] min	(5,10] min	(10,30] min	
Closed-source Models								
Gemini-2.0-Flash	29.7	40.3	43.2	49.4	43.2	41.1	34.9	41.5
Gemini-2.5-Pro	38.5	57.7	61.7	57.8	64.4	55.0	55.9	58.9
Gemini-3-Pro	56.2	54.1	55.7	61.0	56.4	52.9	52.5	55.5
Open-source Models								
VideoLLaMA2-7B	26.4	30.7	29.3	32.0	28.2	29.6	28.3	29.2
VITA-1.5-7B	25.3	28.6	31.5	31.3	27.4	30.6	34.0	30.5
HumanOmniV2-7B	20.9	31.1	31.6	36.6	29.4	29.6	29.3	30.5
Baichuan-Omni-1.5-7B	24.2	31.3	31.4	28.9	31.8	28.4	32.4	30.7
Qwen2.5-Omni-7B	23.1	25.3	30.7	41.6	27.4	25.3	26.7	29.3
+ Explicit Text CoT	30.0	**32.0**	33.9	39.4	32.7	31.0	30.7	33.2
+ LatentOmni	**33.3**	30.2	**36.7**	**45.2**	**33.2**	**33.3**	**34.0**	**35.4** (+6.1pp)

Table 3: Ablation of the components of LatentOmni. The **best** is highlighted.

Method	Daily-Omni	WorldSense	OmniVideoBench	LVOmniBench
w/o Audio in Latent Space	65.9	47.8	33.6	31.6
w/o Visual in Latent Space	63.5	47.2	33.5	32.1
w/o $\mathcal{L}_{\text{latent}}$	61.0	45.2	31.8	30.2
w/o $\mathcal{L}_{\text{sync}}$	65.9	47.1	34.0	33.1
full	**67.4**	**48.9**	**35.4**	**35.1**

Table 4: Ablation of OSPE. The **best** is highlighted.

Positional Scheme	Daily-Omni	WorldSense	OmniVideoBench	LVOmniBench
Conventional RoPE	62.1	45.8	33.1	32.3
Direct TMRoPE	64.4	46.7	34.0	32.9
No Latent PE (w/o OSPE)	66.0	47.8	34.9	33.1
OSPE	**67.4**	**48.9**	**35.4**	**35.1**

Baselines. Open-source AV Understanding Models comparisons include VideoLLaMA2-7B (57), MiniCPM-o-7B (58), VITA-1.5-7B (59), HumanOmniV2-7B (60), Baichuan-Omni-1.5 (61), OmniVinci (62), and Qwen2.5-Omni-7B (23). **Explicit Text CoT** removes the interleaved segment entries from the supervision trajectories and trains on the remaining textual reasoning. **Vanilla SFT** retains the same evidence-segment links and reasoning structure as LatentOmni, but represents each linked segment through its textual description in the intermediate supervision, without continuous latent rollout. Vanilla SFT and LatentOmni use the same original audio-visual inputs, QA pairs, training examples, and update budget. At inference, Vanilla SFT generates its intermediate text from the input rather than receiving reference segment descriptions. This controls for localized textual evidence supervision. We also compare with LVR (48) and Monet (49) under a vision-only protocol. GPT-4o (63) and Gemini (64; 65) provide proprietary reference points.

4.2 Main Results

Table 1 summarizes the main results on four omnimodal benchmarks. We report proprietary systems for context, but focus the controlled comparison on open-source models, text-only reasoning variants, and latent reasoning baselines. Overall, LatentOmni achieves the best performance among the evaluated open-source methods on all four benchmarks, supporting the effectiveness of unified latent-space reasoning for audio-visual tasks.

Comparison with Open-Source Models. LatentOmni consistently improves over existing open-source audio-visual models. Compared with its base model, Qwen2.5-Omni-7B, LatentOmni obtains absolute gains of 4.5% on Daily-Omni, 3.5% on WorldSense, 6.1% on OmniVideoBench, and

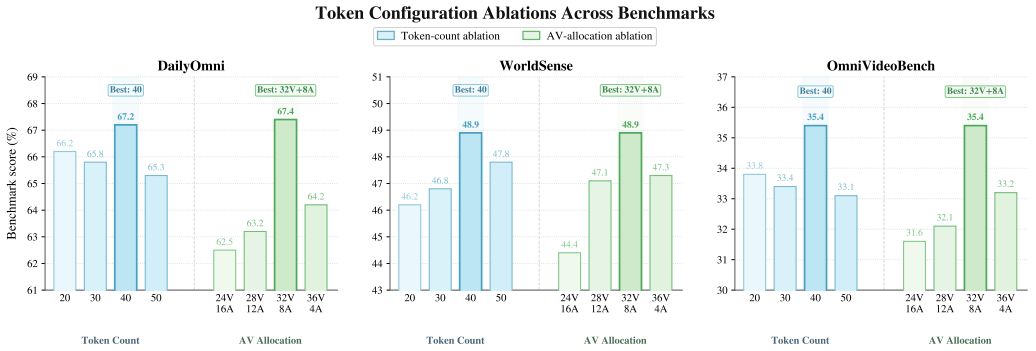


Figure 4: Ablation studies on latent configurations across three benchmarks, specifically evaluating the impact of latent token counts and the allocation ratio of audio and visual latents.

3.1% on LVOmniBench. It also outperforms strong open-source competitors such as OmniVinci and HumanOmniV2-7B on the benchmarks where they report results. The improvement is especially clear on OmniVideoBench, where LatentOmni reaches 35.4% and surpasses all evaluated 7B open-source models, indicating stronger cross-modal alignment and reasoning.

Comparison with Textual Supervision. On Qwen2.5-Omni-7B, LatentOmni improves over Explicit Text CoT by 1.8, 2.3, 2.2, and 3.0 percentage points on Daily-Omni, WorldSense, OmniVideoBench, and LVOmniBench, respectively. Against Vanilla SFT, which retains localized segment descriptions and the reasoning structure, the corresponding gains are 5.4, 1.4, 4.9, and 1.9 points. These gains support the latent-supervised framework beyond textual evidence supervision, without isolating representation from alignment objectives. The input-masking diagnostic (Appendix E) reports a 12.3-point Selective Drop for LatentOmni versus 2.0 for Explicit Text CoT on 300 Daily-Omni samples, indicating selective input dependence under that protocol rather than latent-trace faithfulness.

Fine-Grained Results Analysis To further evaluate cross-modal reasoning, we present a fine-grained analysis on OmniVideoBench. As shown in Table 2, LatentOmni achieves the highest average accuracy (35.4%) among the evaluated open-source models, improving upon the base model by 6.1pp. Furthermore, it consistently outperforms the Explicit Text CoT baseline across most audio categories and all video durations. Notably, on the longest video subset ((10,30] min), our method obtains an accuracy of 34.0%, compared to 30.7% for the text-only CoT variant. These findings indicate that reasoning within a synchronized continuous latent space benefits both joint cross-modal understanding and long-form video processing.

4.3 Ablation Study

We ablate the main design choices of LatentOmni to identify where the gains come from. Specifically, we examine the modality composition of the unified latent space, the role of OSPE, the latent sequence configuration, and the individual contributions of $\mathcal{L}_{\text{latent}}$ and $\mathcal{L}_{\text{sync}}$. Unless otherwise noted, ablations follow the same evaluation protocol as the main experiments.

Component Analysis. Table 3 shows that removing either audio or visual features from the latent space consistently degrades performance, confirming that both modalities contribute to the final reasoning trajectory. Among the training objectives, $\mathcal{L}_{\text{latent}}$ is the most influential: without it, performance drops sharply to 61.0 on Daily-Omni and 31.8 on OmniVideoBench. Ablating $\mathcal{L}_{\text{sync}}$ yields smaller but consistent losses, indicating that temporal synchronization complements latent grounding rather than replacing it.

Position Embeddings Ablation. As shown in Table 4, conventional RoPE and Direct TMRoPE still index latent states by modality-blocked generation order, which differs from physical time and assigns synchronized audio-visual latents different rotary phases. As shown below, removing latent PE avoids this mismatch and can outperform both alternatives, but provides no explicit synchronization signal. OSPE instead combines the time-aware rotary operator with latent-specific re-indexing onto shared physical timestamps, while causality remains controlled by the attention mask. It thus preserves temporal structure without conflating generation order with evidence time.

Impact of Latent Token Configuration. To justify our architectural choices, we ablate the sequence length and modality allocation of latent reasoning trajectories (Figure 4). First, scaling the total token count reveals an empirical optimum at 40 tokens; shorter sequences limit representational capacity, whereas longer sequences introduce noise without yielding further performance gains. Second, by fixing the sequence length to 40 tokens and varying the audio-visual ratio in latent space, we observe that a configuration of 32 visual latent tokens and 8 audio latent tokens consistently achieves the highest accuracy across benchmarks. These results validate our default setup, indicating that this specific token count and modality distribution provides an efficient balance for processing multi-modal information in our model.

5 Conclusion

In this paper, we presented **LatentOmni**, a unified audio-visual reasoning framework that interleaves explicit textual reasoning with continuous latent states, enabling MLLMs to maintain sustained attention to native audio-visual evidence beyond purely language-based CoT. We introduced feature-level supervision to ground latent states in task-relevant audio-visual features, designed Omni-Sync Position Embedding (OSPE) for cross-modal temporal alignment, and constructed **LatentOmni-Instruct-35K** to support latent supervision. Experiments on four omnimodal benchmarks showed that LatentOmni consistently improves over Qwen2.5-Omni and explicit-text CoT, achieving the best performance among evaluated open-source models. These results demonstrate the promise of latent-space joint reasoning as a practical and effective path toward more faithful omnimodal understanding.

Acknowledgments and Disclosure of Funding

This work is supported by the Zhongguancun Academy (Grant Nos. C20250204, C20250602), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), Beijing Municipal Natural Science Foundation (Grant No. L2603008), National Natural Science Foundation of China (92470121, 62402016), National Key R&D Program of China (2024YFA1014003), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

Competing Interests. The authors declare no competing interests.

References

[1] J. Hong, S. Yan, J. Cai, X. Jiang, Y. Hu, and W. Xie, "Worldsense: Evaluating real-world omnimodal understanding for multimodal llms," *arXiv preprint arXiv:2502.04326*, 2025.

[2] Z. Zhou, R. Wang, Z. Wu, and Y.-G. Jiang, "Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities," *arXiv preprint arXiv:2505.17862*, 2025.

[3] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 351–376, 2021.

[4] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[5] Y. Zhang, Y. Shi, W. Yu, Q. Wen, X. Wang, W. Yang, Z. Zhang, L. Wang, and R. Jin, "Debiasing multimodal large language models via penalization of language priors," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 4232–4241.

[6] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 2024.

[7] Y. Li, Z. Liu, Z. Li, X. Zhang, Z. Xu, X. Chen, H. Shi, S. Jiang, X. Wang, J. Wang *et al.*, "Perception, reason, think, and plan: A survey on large multimodal reasoning models," *arXiv preprint arXiv:2505.04921*, 2025.

[8] L. Chen, Y. Zhang, S. Ren, H. Zhao, Z. Cai, Y. Wang, P. Wang, X. Meng, T. Liu, and B. Chang, "Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 1086–1104.

[9] X. Chen, Y. Ding, W. Lin, J. Hua, L. Yao, Y. Shi, B. Li, Y. Zhang, Q. Liu, P. Wan *et al.*, "Avocado: An audiovisual video captioner driven by temporal orchestration," *arXiv preprint arXiv:2510.10395*, 2025.

[10] Y.-F. Zhang, T. Yu, H. Tian, C. Fu, P. Li, J. Zeng, W. Xie, Y. Shi, H. Zhang, J. Wu *et al.*, "Mm-rlhf: The next step forward in multimodal llm alignment," *arXiv preprint arXiv:2502.10391*, 2025.

[11] X. Chen, W. Lin, J. Hua, L. Yao, Y. Ding, B. Li, B. Zeng, Y. Shi, Q. Liu, Y. Zhang *et al.*, "Diadem: Advancing dialogue descriptions in audiovisual video captioning for multimodal large language models," *arXiv preprint arXiv:2601.19267*, 2026.

[12] Y. Shi, H. Wang, W. Xie, H. Zhang, L. Zhao, Y.-F. Zhang, X. Li, C. Fu, Z. Wen, W. Liu *et al.*, "Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios," *arXiv preprint arXiv:2505.21333*, 2025.

[13] D. Team, B. Zeng, D. Hua, K. Zhu, Y. Dai, B. Li, Y. Wang, C. Tong, Y. Yang, M. Chang *et al.*, "Openworldlib: A unified codebase and definition of advanced world models," *arXiv preprint arXiv:2604.04707*, 2026.

[14] B. Zeng, K. Zhu, D. Hua, B. Li, C. Tong, Y. Wang, X. Huang, Y. Dai, Z. Zhang, Y. Yang *et al.*, "Research on world models is not merely injecting world knowledge into specific tasks," *arXiv preprint arXiv:2602.01630*, 2026.

[15] T. Bai, H. Liang, B. Wan, Y. Xu, X. Li, S. Li, L. Yang, B. Li, Y. Wang, B. Cui *et al.*, "A survey of multimodal large language model from a data-centric perspective," *arXiv preprint arXiv:2405.16640*, 2024.

[16] K. Team, D. Chen, J. Chen, X. Chen, Y. Dai, H. Dong, B. Geng, Y. Guan, M. Guo, G. He, G. He, J. Hua, Y. Huang, Y. Ji, X. Jia, Y. Jia, B. Li, J. Li, M. Li, S. Li, S. Li, S. Liao, W. Lin, S. Liu, X. Liu, S. Mao, X. Qu, D. Sheng, Y. Shi, J. Tang, L. Tang, Y. Tang, C. Tong, K. Wang, Q. Wang, R. Wang, T. Wang, Y. Wang, Z. Wang, Z. Wang, B. Xia, Y. Xu, L. Yang, Z. Yuan, B. Zeng, M. Zhang, Y. Zhang, Z. Zhang, Z. Zhang, Y. Zheng, K. Gai, C. Song, and P. Wan, "Behind video generation: Observations and insights on multimodal understanding from kling," 2026. [Online]. Available: https://arxiv.org/abs/2611.04916

[17] C. Li, Y. Chen, Y. Ji, J. Xu, Z. Cui, S. Li, Y. Zhang, W. Wang, Z. Song, D. Zhang *et al.*, "Omnivideobench: Towards audio-visual understanding evaluation for omni mllms," *arXiv preprint arXiv:2510.10689*, 2025.

[18] Z. Xing, X. Hu, C.-W. Fu, W. Wang, J. Dai, and P.-A. Heng, "Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning," *arXiv preprint arXiv:2505.04623*, 2025.

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[20] Y. Wang, S. Wu, Y. Zhang, S. Yan, Z. Liu, J. Luo, and H. Fei, "Multimodal chain-of-thought reasoning: A comprehensive survey," *arXiv preprint arXiv:2503.12605*, 2025.

[21] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li, "Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 8612–8642, 2024.

[22] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[23] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[24] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 180–15 190.

[25] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[26] Y. Shi, J. Liu, Y. Guan, Z. Wu, Y. Zhang, Z. Wang, W. Lin, J. Hua, Z. Wang, X. Chen *et al.*, "Mavors: Multi-granularity video representation for multimodal large language model," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 10 994–11 003.

[27] B. Li, Y. Guan, H. Li, B. Zeng, Y. Ji, Y. Ding, P. Wan, K. Gai, Y. Zhang, and W. Zhang, "Semantic routing: Exploring multi-layer llm feature weighting for diffusion transformers," *arXiv preprint arXiv:2602.03510*, 2026.

[28] B. Li, S. Yang, Y. Guan, R. An, X. Chen, Y. Shi, P. Wan, W. Zhang *et al.*, "Gran-ted: Generating robust, aligned, and nuanced text embedding for diffusion models," *arXiv preprint arXiv:2512.15560*, 2025.

[29] Z. Xie, M. Lin, Z. Liu, P. Wu, S. Yan, and C. Miao, "Audio-reasoner: Improving reasoning capability in large audio language models," *arXiv preprint arXiv:2503.02318*, 2025.

[30] Z. Ma, Z. Chen, Y. Wang, E.-S. Chng, and X. Chen, "Audio-cot: Exploring chain-of-thought reasoning in large audio language model," in *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2025, pp. 1–6.

[31] Y. Dong, Z. Liu, H.-L. Sun, J. Yang, W. Hu, Y. Rao, and Z. Liu, "Insight-v: Exploring long-chain visual reasoning with multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9062–9072.

[32] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[33] H. Zhang, X. Li, and L. 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*, 2023, pp. 543–553.

[34] T.-H. Pham and C. Ngo, "Multimodal chain of continuous thought for latent-space reasoning in vision-language models," *arXiv preprint arXiv:2508.12587*, 2025.

[35] Z. Zhang, T. Wang, X. Gong, Y. Shi, H. Wang, D. Wang, and L. Hu, "When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms," *arXiv preprint arXiv:2511.02243*, 2025.

[36] B. Li, X. Xue, S. Yang, Y. Shi, X. Chen, Y. Guan, Y. Zhang, and W. Zhang, "The unseen bias: How norm discrepancy in pre-norm mllms leads to visual information loss," *ArXiv*, vol. abs/2512.08374, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:283712146

[37] R. Zhang, D. Ji, L. Zhu, X. Liu, Y. Meng, R. Chu, and Y. Yang, "Video-zero: Self-evolution video understanding," *arXiv preprint arXiv:2605.14733*, 2026.

[38] Z. Qian, Y. Ma, Z. Ouyang, Z. Wang, Z. Xu, F. Luo, X. Liu, Z. Ge, Y. Guo, and J. Han, "Cognitive pivot points and visual anchoring: Unveiling and rectifying hallucinations in multimodal reasoning models," *arXiv preprint arXiv:2604.10219*, 2026.

[39] H. Fang, J. Li, J. Kong, T. Zhuang, K. Gao, B. Chen, S.-T. Xia, and Y. Wang, "Seeing through the chain: Mitigate hallucination in multimodal reasoning models via cot compression and contrastive preference optimization," *arXiv preprint arXiv:2602.03380*, 2026.

[40] Z. Xiong, Y. Cai, Z. Li, J. Yuan, and Y. Wang, "Thinking with sound: Audio chain-of-thought enables multimodal reasoning in large audio-language models," *arXiv preprint arXiv:2509.21749*, 2025.

[41] Z. Su, P. Xia, H. Guo, Z. Liu, Y. Ma, X. Qu, J. Liu, Y. Li, K. Zeng, Z. Yang *et al.*, "Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers," *arXiv preprint arXiv:2506.23918*, 2025.

[42] H. Zhang, X. Gu, J. Li, C. Ma, S. Bai, C. Zhang, B. Zhang, Z. Zhou, D. He, and Y. Tang, "Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning," *arXiv preprint arXiv:2508.04416*, 2025.

[43] S. S. Kancheti, A. S. Kanade, V. N. Balasubramanian, and T. Ganu, "Chain-of-thought degrades visual spatial reasoning capabilities of multimodal llms," *arXiv preprint arXiv:2604.16060*, 2026.

[44] S. Goyal, Z. Ji, A. S. Rawat, A. K. Menon, S. Kumar, and V. Nagarajan, "Think before you speak: Training language models with pause tokens," *arXiv preprint arXiv:2310.02226*, 2023.

[45] S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian, "Training large language models to reason in a continuous latent space," *arXiv preprint arXiv:2412.06769*, 2024.

[46] E. Zelikman, G. Harik, Y. Shao, V. Jayasiri, N. Haber, and N. D. Goodman, "Quiet-star: Language models can teach themselves to think before speaking," *arXiv preprint arXiv:2403.09629*, 2024.

[47] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," *arXiv preprint arXiv:2404.08471*, 2024.

[48] B. Li, X. Sun, J. Liu, Z. Wang, J. Wu, X. Yu, H. Chen, E. Barsoum, M. Chen, and Z. Liu, "Latent visual reasoning," *arXiv preprint arXiv:2509.24251*, 2025.

[49] Q. Wang, Y. Shi, Y. Wang, Y. Zhang, P. Wan, K. Gai, X. Ying, and Y. Wang, "Monet: Reasoning in latent visual space beyond images and language," *arXiv preprint arXiv:2511.21395*, 2025.

[50] C. Liu, Y. Yang, Y. Fan, Q. Wei, S. Liu, and X. E. Wang, "Reasoning within the mind: Dynamic multimodal interleaving in latent space," *arXiv preprint arXiv:2512.12623*, 2025.

[51] K. Li, C. Shang, L. Karlinsky, R. Feris, T. Darrell, and R. Herzig, "Latent implicit visual reasoning," *arXiv preprint arXiv:2512.21218*, 2025.

[52] Y. Qin, B. Wei, J. Ge, K. Kallidromitis, S. Fu, T. Darrell, and X. Wang, "Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens," *arXiv preprint arXiv:2511.19418*, 2025.

[53] Y. Li, H. Zhang, M.-H. Guo, W. Gao, S. Jia, S. Jiao, Q. Hou, and M.-M. Cheng, "Towards universal video mllms with attribute-structured and quality-verified instructions," *arXiv preprint arXiv:2602.13013*, 2026.

[54] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[55] A. Zeng, X. Lv, Q. Zheng, Z. Hou, B. Chen, C. Xie, C. Wang, D. Yin, H. Zeng, J. Zhang *et al.*, "Glm-4.5: Agentic, reasoning, and coding (arc) foundation models," *arXiv preprint arXiv:2508.06471*, 2025.

[56] K. Tao, Y. Zheng, J. Xu, W. Du, K. Shao, H. Wang, X. Chen, X. Jin, J. Zhu, B. Yu *et al.*, "Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms," *arXiv preprint arXiv:2603.19217*, 2026.

[57] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao *et al.*, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024.

[58] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[59] C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, H. Cao, Z. Long, H. Gao, K. Li *et al.*, "Vita-1.5: Towards gpt-4o level real-time vision and speech interaction," *arXiv preprint arXiv:2501.01957*, 2025.

[60] Q. Yang, S. Yao, W. Chen, S. Fu, D. Bai, J. Zhao, B. Sun, B. Yin, X. Wei, and J. Zhou, "Humanomniv2: From understanding to omni-modal reasoning with context," *arXiv preprint arXiv:2506.21277*, 2025.

[61] Y. Li, J. Liu, T. Zhang, S. Chen, T. Li, Z. Li, L. Liu, L. Ming, G. Dong, D. Pan *et al.*, "Baichuan-omni-1.5 technical report," *arXiv preprint arXiv:2501.15368*, 2025.

[62] H. Ye, C.-H. H. Yang, A. Goel, W. Huang, Z. Wan, J. Tian, A.-C. Cheng, L. Zhu, Y. Su, Y. Lou *et al.*, "Omnivinci: Enhancing architecture and data for omni-modal understanding llm," in *International Conference on Learning Representations*, vol. 2026, 2026, pp. 56 101–56 138.

[63] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[64] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[65] S. Pichai, D. Hassabis, and K. Kavukcuoglu, "A new era of intelligence with gemini 3," *Google. URL: https://blog.google/products-and-platforms/products/gemini/gemini-3/*, 2025.

Appendices

A Datasets Details

A.1 Caption Database

AvoCaDO is a newly curated dataset consisting of 107K high-quality, temporally-aligned audiovisual video captions. It emphasizes the temporal orchestration between visual and auditory events, offering semantically rich descriptions paired with precise temporal synchronization. This dataset is specifically designed to enhance temporal coherence, dialogue accuracy, and comprehensive multimodal alignment in audiovisual video captioning tasks.

ASID features a large-scale collection of one million structured, fine-grained audiovisual instruction annotations (ASID-1M). It provides single- and multi-attribute supervision—covering scenes, objects, actions, speech, camera movements, and narrative elements. Curated through an automated verification and refinement pipeline, this dataset is designed to mitigate hallucinations and facilitate highly controllable, reliable, and fine-grained video understanding.

A.2 Dataset Statistics and Data Quality

Dataset composition and supervision. LatentOmni-Instruct-35K contains 35K question-answer pairs with interleaved reasoning trajectories (Table 5). Its supervision links evidence-referencing reasoning steps to modality-tagged, timestamped audio-visual segments, from which the feature targets for latent alignment are extracted (Sec. 3.4). The reported average of 2.10 latent reasoning spans per trajectory counts segment-linked reasoning phases, not individual latent tokens; each phase contains K latent tokens.

The modality proportions describe the assigned question categories rather than experimentally verified modality necessity. In particular, 69.6% of samples are labeled AV-dependent, while the remaining 30.4% are assigned audio- or video-dominant categories. Thus, the dataset combines different modality requirements; these annotations alone do not establish that every question requires both modalities.

Table 5: Statistics of LatentOmni-Instruct-35K.

Dimension	Value
QA / reasoning trajectories	35K
Total / average duration	296.5 h / 30.5 s
Avg. latent reasoning spans per trajectory	2.10
Avg. linked evidence-segment duration	4.71 s
Audio-dominant / video-dominant / AV-dependent	6.2% / 24.2% / 69.6%
Quality-audit rejection rate	65%

Table 6: Comparison of audio-visual instruction data formats. Timestamped denotes localized evidence annotations; interleaved AV denotes explicit segment placeholders within a reasoning trajectory. These columns describe annotation formats, not comparative data quality.

Dataset	Training samples	Domain	Avg. duration	Timestamped	Interleaved AV	Evidence annotation
MUSIC-AVQA	32K	Music	60 s	×	×	×
AVQA	40K	Open	10 s	×	×	×
JavisInst-Und	110K	Open	10 s	×	×	×
OmniVideo-100K	100K	Open	103 s	✓	×	Clue level
LatentOmni-Instruct-35K	**35K**	**Open**	**30.5 s**	✓	✓	**Reasoning-step links**

Human audit. Two annotators independently reviewed 500 stratified trajectories from the final dataset, with disagreements adjudicated by a third annotator. Table 7 reports correctness and issue rates separately for QA/reasoning, evidence relevance, temporal alignment, and step-to-segment correspondence, together with the reported inter-annotator agreement. These dimensions assess complementary parts of the supervision: a correct answer need not imply that its evidence boundaries or reasoning-step links are correct.

Within this audited subset, dimension-wise correctness ranges from 96.2% to 98.2%, and major-error rates range from 1.0% to 1.2%. Step-to-segment correspondence has the lowest correctness (96.2%) and reported agreement ($\kappa = 0.69$), identifying evidence linkage as a remaining source of annotation uncertainty. These results characterize the audited subset; they do not establish error-free supervision or, without the sampling weights, an unbiased estimate of dataset-wide quality.

Table 7: Human audit of 500 stratified trajectories. Percentages are reported separately for each audit dimension; errors can overlap across dimensions, so the rows do not determine a trajectory-level error rate.

Human-audited dimension	Correct (%)	Minor issue (%)	Major error (%)	Cohen's κ
QA and reasoning correctness	98.0	0.8	1.2	0.72
Evidence-segment relevance	98.2	0.8	1.0	0.75
Temporal boundary/alignment	97.0	2.0	1.0	0.79
Step-to-segment correspondence	96.2	2.8	1.0	0.69

B Implementation Details

We set the key hyperparameters for our training process as follows: the maximum number of frames per sample (FPS_MAX_FRAMES) is capped at 256. For optimization, the learning rate is set to 10^{-5}, with a warmup fraction of 0.05 to gradually ramp it up at the start of training. The weighting coefficients for the loss terms are configured as $\lambda_1 = 0.005$ and $\lambda_2 = 1.0$. Furthermore, due to limited computational resources, we restrict the batch size to 1, paired with 12 gradient accumulation steps to maintain an adequate effective batch size for stable optimization.

B.1 AVQA Synthesis Instruction

This section details the instructions for synthesizing AVQA data, covering both open-ended and multiple-choice question (MCQ) formats. The prompt is designed to ensure that the generated pairs satisfy strict cross-modal dependency and reflect the diversity of the source captions. The complete instruction template is provided in Fig. 5 and Fig. 6.

B.2 AVQA Classification and Filtering Instruction

To implement the category annotation and quality control described in Section 3.3, we design a joint instruction for GLM-4.7. This prompt directs the model to classify the reasoning type of each preliminary AVQA pair and assess its overall quality to facilitate our threshold-based filtering. The detailed prompt template is illustrated in Fig. 7.

B.3 Segment Level Caption Synthesis Instruction

To synthesize high-quality audio-visual segment-level captions, we design two distinct prompts tailored for the visual and audio modalities. These instructions direct the model to produce separate, concise, and objective descriptions that are strongly correlated with the QA pairs, thereby preventing information omission. The specific prompts for synthesizing observable visual elements and identifiable audio elements are illustrated in Fig. 8 and Fig. 9, respectively.

B.4 Segment Level Caption Fusion and Refinement Instruction

To synthesize comprehensive and cohesive audio-visual captions while resolving narrative fragmentation caused by shot transitions, we design two sequential instructions for caption fusion and refinement. The detailed prompts for the fusion and refinement processes are illustrated in Fig. 10 and Fig. 11, respectively.

B.5 AV Interleaved Reasoning Trajectory Synthesis Instruction

To construct the audio-visual interleaved reasoning trajectories, we prompt GLM-4.7 utilizing the previously generated AVQA pairs and aligned segment-level captions. The detailed prompt template for this synthesis is illustrated in Fig. 12.

Prompt 1: AV Open-Ended Question-Answer Synthesis

Role. *You are an expert multimodal dataset designer specializing in Audio-Visual Question Answering (AVQA).*
Input. *Temporally aligned audio and video captions describing synchronized events.*
Task Description.

1. ***Context Comprehension:*** *Thoroughly analyze the provided audio and video captions to understand the synchronized multimodal events and their temporal correlations.*
2. ***Open-Ended Generation:*** *Synthesize exactly* ***one*** *high-quality, open-ended question-answer pair that demands complex, multi-step cross-modal reasoning.*
3. ***Structured Output:*** *Format the final result strictly as a JSON object containing the question, concise answer, and the specific reasoning type employed.*

Hard Constraints.

- ***Cross-Modal Information Dependency:*** *The question must strictly rely on the synthesis of both visual and audio information. It must be logically impossible to deduce the answer using only a single modality.*
- ***Reasoning Typology:*** *The generated question must explicitly target a distinct complex reasoning category (e.g., causal reasoning, spatial relations, temporal sequencing, sound-action attribution, or object interactions).*
- ***Answer Accuracy:*** *The question and its concise answer (maximum 10 words) must be factually accurate, concrete, and grounded strictly within the provided caption content, with zero external hallucination.*
- ***Format and Style:*** *Avoid using object IDs, bounding box labels, timestamps, or raw XML tags in the generated question and answer.*
- ***Commonsense Integration:*** *Encourage structural world knowledge when appropriate, such as logically linking visible physical actions with expected environmental acoustics.*

Output Format. *Output raw JSON only. Do not wrap the output in Markdown blocks.*
Reference JSON Schema.

```
{
  "id": "OpenQA_01",
  "modality": "AV",
  "question": "...",
  "answer": "..."
}
```

Figure 5: Prompt used to synthesize a single complex, open-ended AVQA pair from temporally aligned audio-visual captions.

Prompt 2: AV Multiple-Choice QA Synthesis

Role. *You are an expert multimodal dataset designer specializing in Audio-Visual Question Answering (AVQA).*
Input. *Temporally aligned audio and video captions describing synchronized events.*
Task Description.

1. ***Context Comprehension:*** *Thoroughly analyze the provided audio and video captions to understand the synchronized multimodal events and their temporal correlations.*
2. ***Multiple-Choice Generation:*** *Synthesize exactly* ***one*** *high-quality multiple-choice question (MCQ) that demands complex, multi-step cross-modal reasoning.*
3. ***Structured Output:*** *Format the final result strictly as a JSON object containing the question, four distinct options, the correct answer, and the specific reasoning type employed.*

Hard Constraints.

- ***Cross-Modal Information Dependency:*** *The question must strictly rely on the synthesis of both visual and audio information. It must be logically impossible to deduce the answer using only a single modality.*
- ***Reasoning Typology:*** *The generated question must explicitly target a distinct complex reasoning category (e.g., causal reasoning, spatial relations, temporal sequencing, sound-action attribution, or object interactions).*
- ***Distractor Quality:*** *The MCQ must contain exactly four options. The three incorrect options (distractors) must be plausible, misleading, and reflect reasonable alternative interpretations, but must be definitively incorrect without any ambiguity or subjectivity.*
- ***Answer Accuracy:*** *The strictly correct answer must be factually accurate, concrete, and grounded entirely within the provided caption content, with zero external hallucination.*
- ***Format and Style:*** *Avoid using object IDs, bounding box labels, timestamps, or raw XML tags in the generated question and options.*

Output Format. *Output raw JSON only. Do not wrap the output in Markdown blocks.*
Reference JSON Schema.

```
{
  "id": "MCQ_01",
  "modality": "AV",
  "question": "...",
  "options": [
    "A. ...",
    "B. ...",
    "C. ...",
    "D. ..."
  ],
  "answer": "B",
  "answer_text": "B. ..."
}
```

Figure 6: Prompt used to synthesize a single complex, multiple-choice AVQA pair with grounded distractors and a unique correct answer from temporally aligned captions.

Prompt 3: AVQA Dataset Quality Evaluation and Classification

Role. *You are an expert evaluator and classifier specializing in Multimodal Large Language Models (MLLMs) and Audio-Visual Question Answering (AVQA) datasets.*
Objective. *Perform TWO tasks based on the provided inputs:*

1. *Objectively evaluate the quality, logical soundness, and modality dependency of the provided Question-Answer pair.*
2. *Classify the user's question into one specific AVQA category AND determine its primary modality dependency.*

Input Data.

- *`[Standard AV Caption]`: {AV_caption}*
- *`[Question]`: {question}*
- *`[Ground Truth Answer]`: {answer}*

Task 1: QA Quality Evaluation (1-5 Scale). *Evaluate the provided QA pair across the following 3 dimensions:*

1. *Question Difficulty (1-5): How inherently difficult is the question? (5 = requires complex, multi-step reasoning or nuanced integration of evidence; 1 = simple, shallow factual lookup or direct matching).*
2. *Logic Soundness (1-5): Is the question-answer pair logically consistent, unambiguous, and supported by the provided evidence? (5 = logically rigorous with a clearly supported answer; 1 = logically inconsistent, contradictory, or unsupported).*
3. *Modality Dependency (1-5): To what extent does answering the question require evidence from the audio and/or visual content rather than relying on general knowledge or linguistic shortcuts? (5 = strongly dependent on the provided sensory evidence; 1 = largely answerable without audio-visual evidence).*

Task 2: Question Classification & Modality. *First, classify the question into EXACTLY ONE of the following 10 categories: (1) Audio-Visual Joint Perception, (2) Action & Behavior Recognition, (3) Spatial Layout Understanding, (4) Temporal Sequence Understanding, (5) Attribute Comparison & Change, (6) Counting & Quantification, (7) Emotion & Atmosphere Perception, (8) Semantic Content Summarization, (9) Logical Relation Reasoning, (10) Intention & Outcome Prediction.*
Second, determine the primary modality dependency. Choose EXACTLY ONE:

- *`AV-Strong`: Requires logically combining visual and auditory cues.*
- *`Video-Strong`: Relies primarily on visual information.*
- *`Audio-Strong`: Relies primarily on auditory information.*

Note. *The modality dependency score in Task 1 measures reliance on sensory evidence, whereas the modality classification in Task 2 identifies which modality or modalities provide that evidence.*
Output Format. *Output ONLY a valid JSON object. Do not include markdown code blocks, conversational text, or explanations outside the JSON structure.*
Reference JSON Schema.

```
{
  "evaluation": {
    "question_difficulty": <int, 1-5>,
    "logic_soundness": <int, 1-5>,
    "modality_dependency_score": <int, 1-5>
  },
  "classification": {
    "category_id": <int, 1-10>,
    "category_name": "<string>",
    "modality_dependency": "<AV-Strong|Video-Strong|Audio-Strong>",
    "confidence": <float, 0.0-1.0>,
    "reasoning": "<string, 1-2 sentences>"
  }
}
```

Figure 7: Prompt used to evaluate the reasoning difficulty, logical soundness, and modality dependency of synthesized AVQA pairs, and to classify their reasoning categories and primary modalities.

Prompt 4: Segment Level Caption Synthesis (Video)

Given the provided QA pair, provide a CONCISE yet complete visual-only description of the video segment that contains relevant information to answer the question. Limit the description to no more than five sentences and avoid redundant details. Describe only directly observable visual elements: setting, people, actions, objects, and camera movement. Do not infer mood, intent, genre, cultural style, or add interpretation, and strictly avoid speculation and evaluative language.

Figure 8: Prompt used to synthesize a concise video caption focusing exclusively on observable visual elements guided by a specific QA pair.

Prompt 5: Segment Level Caption Synthesis (Audio)

Given the provided QA pair, provide a CONCISE yet complete audio-only description of the segment that contains relevant sound information to answer the question. Limit the description to no more than five sentences and avoid redundant details. State only clearly identifiable sound sources (e.g., music, instruments, environmental noises). If speech is present, accurately report the speaker and the spoken content. Do not infer mood, intent, genre, or cultural style, and strictly avoid speculation and atmospheric language.

Figure 9: Prompt used to synthesize a concise audio caption detailing identifiable sounds and speech relevant to a specific QA pair.

Prompt 6: Segment Level Caption Fusion

You are tasked with fusing the visual caption and audio caption into a single, coherent narrative based on the video content. Follow these strict rules:
1. Preserve every single sentence from both the visual caption and audio caption exactly as they appear.
2. Do NOT omit or delete any sentence in any way.
3. You may reorder the sentences (from both captions) to create a logical and temporally accurate sequence that reflects the video's events.
4. Ensure the integrated narrative flows naturally in time with the video, aligning visual actions with corresponding sounds or spoken content.
Verify before responding: Did I include every sentence from both captions?

Figure 10: Prompt used to fuse the segment-level video caption and audio caption.

Prompt 7: Segment Caption Refinement

Task: *Refine a fragmented video segment caption using the full video caption as a reference.*

Inputs:

- ***[Full Caption]:*** *[full_AV_caption]*
- ***[Segment Caption]:*** *[segment_AV_caption]*

Instructions: *Compare the [Segment Caption] with the [Full Caption] to fix shot fragmentation. Resolve incomplete actions, abrupt cuts, or missing subjects caused by rigid segmentation by accurately contextualizing actions that span across boundaries. Do not describe events outside the segment's timestamp. Keep the refined caption concise, strictly objective, and output ONLY the refined text without any formatting tags or explanations.*

Figure 11: Prompt used to refine fragmented segment captions by cross-referencing full captions to address shot fragmentation and maintain temporal constraints.

Prompt 8: AV Interleaved Reasoning Trajectory Synthesis

Role. *You are an intelligent audio-visual analysis expert with advanced perception capabilities. Your task is to answer the user's question by rigorously analyzing the audio and video content step by step.*
Critical Context. *You will be provided with specific segments and detailed descriptive content.* ***You must treat this detailed content as your own direct sensory perception (sight and hearing).*** *You are NOT analyzing a text; you are describing the raw video and audio you are "watching".*
Strict Prohibitions.

- *Never mention words such as "caption", "text", "description", "reference", "provided info", or "input".*
- *Never say "according to the description" or "the text says".*
- *Never discuss discrepancies between visual or audio evidence and text; instead, state the perceived event as the ground truth.*

Reasoning Instructions.

1. *Construct a step-by-step reasoning chain that actively decides when source segments must be revisited.*
2. *When a segment is needed, cite it using the format* `[Segment n]`.
3. *Cite at least one segment and at most three segments in total, favoring the earliest or most decisive evidence.*
4. *Continue the reasoning without explicitly referring back to segment numbers in the narrative text.*

Final Answer Requirement. *End the response with a concise boxed answer in the format* `\boxed{your final answer here}`*, with no extra commentary afterward.*
Reference Output Pattern.

```
Reason -> [Segment n] -> Reason -> ... -> \boxed{final answer}
```

Figure 12: Prompt used to synthesize interleaved reasoning trajectories with explicit segment citations and concise grounded observations.

C More Experimental Results

To further evaluate the versatility of our approach, we extend our comparison to recent works focusing on visual latent reasoning, specifically LVR and Monet. We conduct this evaluation on the VideoMME benchmark. To ensure a fair comparison with these vision-centric models, the evaluation is performed under a pure visual setting (i.e., without audio inputs).

The results are summarized in Table 8. LATENTOMNI demonstrates highly competitive performance in the pure visual domain, outperforming both LVR and Monet across all video duration categories (Short, Medium, and Long). Notably, our model achieves the highest overall score of 60.8. This indicates that the reasoning capabilities of LATENTOMNI generalize effectively to unimodal visual tasks.

Table 8: Performance comparison with recent visual latent reasoning methods on the VideoMME benchmark (without audio inputs).

Method	Overall	Short Video	Medium Video	Long Video
LVR	36.7	39.2	36.6	34.3
Monet	51.6	52.9	56.0	46.0
Qwen2.5-Omni-7B	58.5	70.0	55.2	50.1
LatentOmni	**60.8**	**70.8**	**60.5**	**50.4**

D Additional Inference Analysis

D.1 Latent-Mode Control and Placement

Entry and exit. LatentOmni predicts when to enter latent mode while using a fixed number of states within each latent span. The trigger token `<Unified_Latent>` is generated autoregressively by the model; it is neither inserted by a rule nor supplied in the evaluation prompt. Once the trigger is emitted, the decoder generates $K = 40$ continuous states, comprising 32 visual states followed by 8 audio states. The closing token `</Unified_Latent>` is then inserted deterministically to resume text decoding. Thus, the entry positions and number of latent spans are determined by the model, whereas the span length and exit are fixed by the decoding procedure (Sec. D.2).

Trigger statistics. Table 9 summarizes latent-mode usage across the four benchmarks. The average number of spans ranges from 1.68 to 1.97 per sample, and 60.5%–75.3% of samples contain more than one span. Samples without a trigger account for 3.4% of Daily-Omni and 1.8% of WorldSense, with none observed on OmniVideoBench or LVOmniBench. These statistics show that the model varies its use of latent mode across inputs. Span closure is enforced by the decoder, so the absence of malformed or unclosed spans reflects this constraint rather than a learned ability to predict a closing token.

Table 9: Latent-mode usage at inference. Span counts are averaged over all samples in each benchmark, including those without a trigger. Multiple triggers denotes samples with at least two latent spans.

Benchmark	Avg. spans/sample	No trigger (%)	Multiple triggers (%)
Daily-Omni	1.68	3.4	60.5
WorldSense	1.87	1.8	65.7
OmniVideoBench	1.97	0.0	73.6
LVOmniBench	1.94	0.0	75.3

Placement comparison. We compare model-predicted entry with two alternative placement strategies: *Random placement*, which inserts latent spans at random reasoning steps, and *Begin-only placement*, which restricts insertion to the beginning of the reasoning sequence. Both alternative strategies are retrained rather than imposed only at inference on the LatentOmni checkpoint. All three strategies use 40 latent states per span. Table 10 reports their accuracy, alongside the base model and Explicit Text CoT for reference. LatentOmni outperforms both placement alternatives on all four benchmarks, with gains of 1.4–6.4 percentage points over Random placement and 4.9–11.8 points over Begin-only placement. These results support the effectiveness of the model-predicted placement strategy relative to the tested alternatives. The comparison fixes the per-span budget; attributing the differences solely to placement would additionally require matching the number of spans and other training and decoding conditions.

Table 10: Accuracy (%) with alternative latent-span placement strategies. Each latent span contains 40 states; this specifies a per-span budget, not a matched total budget per sample. The best result in each column is bold.

Method	Daily-Omni	WorldSense	OmniVideoBench	LVOmniBench
Qwen2.5-Omni-7B	62.9	45.4	29.3	32.0
+ Explicit Text CoT	65.6	46.6	33.2	32.1
+ Random placement	64.7	42.5	34.0	30.0
+ Begin-only placement	59.2	37.1	30.5	28.9
+ LatentOmni	**67.4**	**48.9**	**35.4**	**35.1**

D.2 Inference Cost

Evaluation setting. We evaluate the accuracy–cost trade-off of LatentOmni on Daily-Omni using the same backbone, inputs, hardware, numerical precision, batch size, and decoding configuration across methods, apart from their respective text-only or interleaved text–latent generation procedures. We report accuracy, average generated text-token and latent-state counts per sample, total autoregressive (AR) steps, end-to-end latency, and peak memory in Table 11. Latency includes audio-visual encoding, prefill, and autoregressive generation. The reported AR count is the sum of the text-token and latent-state counts. Latent states are generated autoregressively and therefore also contribute to sequential decoding cost.

Table 11: Inference cost on Daily-Omni under a shared evaluation setup. Text and latent counts, and their sum (AR steps), are averages per sample. Accuracy is in percent, end-to-end latency in seconds, and peak memory in GiB. Latency includes audio-visual encoding and prefill. Latent counts refer to continuous states rather than vocabulary tokens.

Method	Acc. ↑	Text	Latent	AR steps ↓	E2E latency(s) ↓	Peak mem.(GiB) ↓
Qwen2.5-Omni-7B	62.9	231.3	0.0	231.3	21.6	20.6
+ Explicit Text CoT	65.6	259.7	0.0	259.7	22.7	20.9
+ LatentOmni	**67.4**	178.5	67.2	245.7	22.3	20.8

Accuracy and inference cost. Relative to Explicit Text CoT, LatentOmni improves accuracy by 1.8 percentage points while reducing the average text-token count from 259.7 to 178.5. After including 67.2 latent states, the total AR count is 245.7 steps, 14.0 fewer than Explicit Text CoT (a 5.4% reduction). The latent-state count is consistent with the average of 1.68 spans on Daily-Omni and the fixed 40-state span length (Table 9). Measured end-to-end latency and peak memory are similar: 22.3 versus 22.7 seconds and 20.8 versus 20.9 GiB. Compared with the base model, LatentOmni gains 4.5 accuracy points with 14.4 additional AR steps and 0.7 seconds higher end-to-end latency. Together, these measurements indicate an improved accuracy–cost trade-off relative to Explicit Text CoT in this setting. They do not imply that an individual latent step is cheaper than a text step or that latent reasoning guarantees wall-clock acceleration.

E Behavioral Grounding Diagnostic Beyond Attention

Purpose and scope. Attention visualizations describe how a model distributes attention over its inputs, but do not establish that the attended evidence determines its answer. We therefore complement the qualitative examples with an input intervention diagnostic: does removing annotated answer-supporting evidence reduce accuracy more than removing matched intervals outside that evidence? This tests the selectivity of the model's final-answer behavior, rather than the faithfulness of individual latent reasoning steps.

Interventions and metric. On 300 manually verified Daily-Omni samples, we compare the original inputs (*Clean*) with two masking conditions. *Gold-mask* masks annotated answer-supporting intervals. *Random-mask* masks intervals outside the annotated evidence, matched in duration, modality configuration, and perturbation type. Intervals outside the annotations are treated as controls; they are not guaranteed to contain no useful or redundant evidence. For accuracies expressed as percentages, we define

$$\text{Selective Drop} = \text{Acc}(\text{Random-mask}) - \text{Acc}(\text{Gold-mask}), \tag{7}$$

measured in percentage points (pp). Equivalently, this is the clean-to-gold accuracy drop minus the clean-to-random accuracy drop. A positive value means that masking annotated evidence is more disruptive than the matched control under this protocol; its magnitude is not a direct measure of reasoning faithfulness.

Table 12: Input-masking diagnostic on 300 manually verified Daily-Omni samples. Accuracy is reported in percent and Selective Drop in percentage points (pp). Gold-mask accuracy has no standalone preferred direction: interpretation requires the Clean and Random-mask conditions.

Method	Clean (%)	Gold-mask (%)	Random-mask (%)	Selective Drop (%)
Qwen2.5-Omni-7B	69.3	63.0	70.0	7.0
+ Explicit Text CoT	71.3	68.3	70.3	2.0
+ LatentOmni	74.7	60.7	73.0	**12.3**

Observed response to evidence removal. LatentOmni's accuracy decreases by 14.0 pp under Gold-mask and by 1.7 pp under Random-mask, yielding a Selective Drop of 12.3 pp. The corresponding point estimates are 7.0 pp for Qwen2.5-Omni-7B and 2.0 pp for Explicit Text CoT. Thus, on this subset, LatentOmni exhibits the largest observed difference between evidence masking and control masking, while its control accuracy remains close to its clean accuracy. This pattern is consistent with greater selectivity toward the annotated evidence under the tested interventions.

Interpretation limits. The diagnostic acts on the input, so it cannot isolate whether an answer depends on the latent reasoning states or on another path from the input to the output. Matched masking controls for the specified perturbation properties, but does not eliminate differences in content, annotation coverage, or the distribution shift induced by masking. The reported aggregate accuracies also do not establish the statistical significance of the between-model differences. We therefore interpret this experiment as subset-level behavioral evidence of selective input dependence, without inferring faithful latent reasoning, excluding shortcut learning, or establishing joint necessity of both modalities.

F Case Study

In this section, we present examples from DailyOmni benchmark to demonstrate the reasoning capabilities of LatentOmni across diverse audio-visual tasks. For clarity, we abstain from projecting the generated latent embeddings into the discrete language space, as this would yield uninterpretable tokens. Instead, we represent the latent reasoning segments using <Unified_Latent><latent_embeddings></Unified_Latent>. The selected examples encompass three representative scenarios: AV Event Alignment (Figure 13), Inference (Figure 14), and Reasoning (Figure 15).

To further analyze the model's intrinsic reasoning process within the latent space, we concurrently visualize the attention maps between the latent reasoning tokens and the original audio-visual inputs alongside the generated outputs. These visualizations intuitively illustrate how the latent states dynamically track and anchor to fine-grained multimodal evidence during generation. These qualitative analyses highlight the model's capability to preserve and reason over continuous multimodal semantics throughout the inference process.

G Limitation

While LatentOmni establishes a robust framework for unified latent reasoning across visual, auditory, and textual modalities, it inherently shares a common boundary with current state-of-the-art multimodal systems regarding modality coverage. Real-world environments are fundamentally more complex, encompassing a broader spectrum of sensory and control signals, such as 3D spatial representations, tactile physics, and motor action commands. Currently, mapping these extended physical and interactive signals into a single unified latent space remains an open challenge for the community. In future work, we aim to explore the expansion of our latent semantic bridge to accommodate a wider array of heterogeneous modalities, ultimately taking a step towards a more comprehensive and embodied omni-modal reasoning system.



Figure 13: LatentOmni example: AV Event Alignment. LatentOmni accurately anchors task-relevant audio-visual frames within the latent space. As demonstrated by the attention visualization, deeper colors indicate higher attention weights precisely localized on the key multimodal clues.

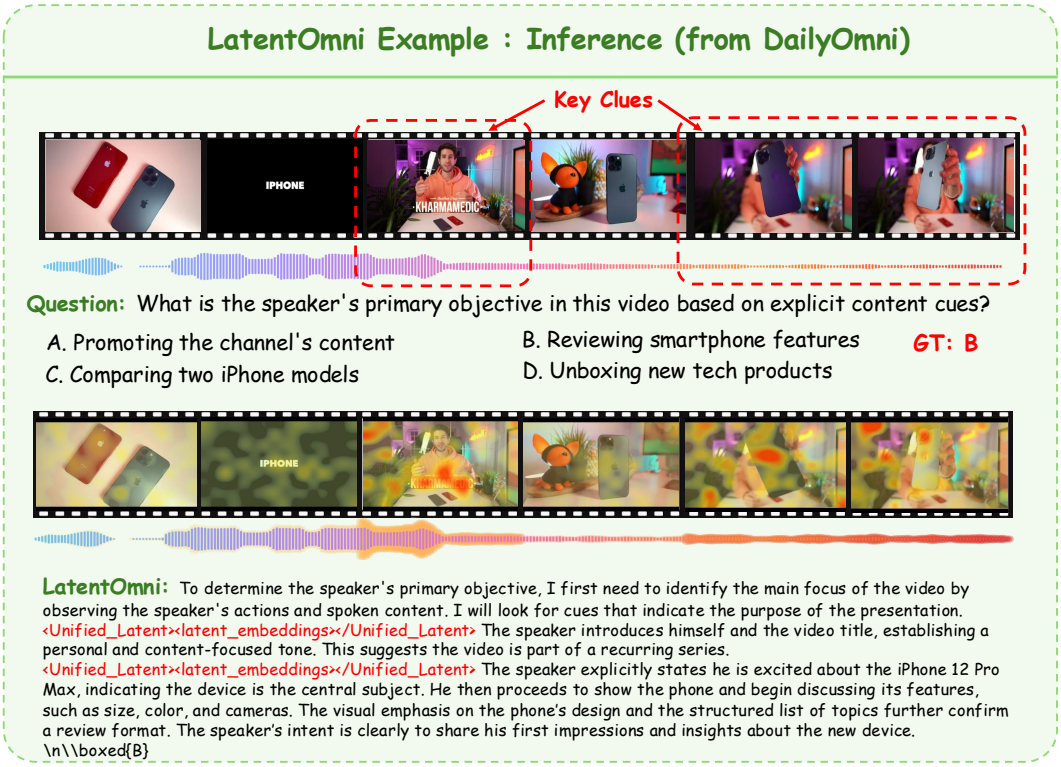
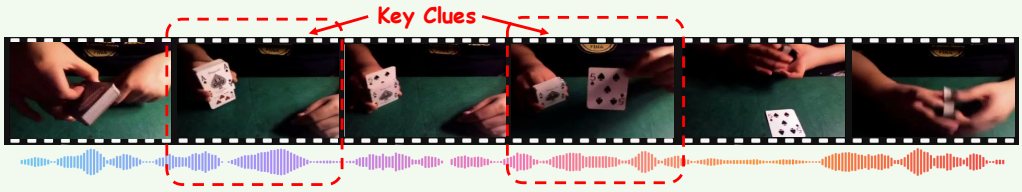


Figure 14: LatentOmni example: Inference. LatentOmni accurately anchors task-relevant audio-visual frames within the latent space. As demonstrated by the attention visualization, deeper colors indicate higher attention weights precisely localized on the key multimodal clues.

LatentOmni Example : Reasoning (from DailyOmni)



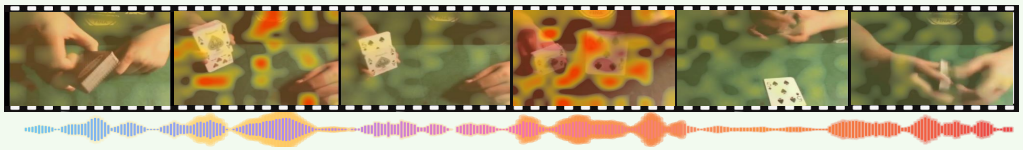
Question: Why does the person tap the edge of the card deck on the green table while the speaker is talking about showing 'their card'? **GT: C**

A. To create a distraction before performing a move.

B. To demonstrate the color of the table.

C. To visually highlight the process of show the card.

D. To emphasize the sound of the cards.



LatentOmni: To determine the purpose of tapping the card deck on the table, I need to identify the moment when the action occurs and correlate it with the speaker's instructions about revealing \"their card.\" I will first locate the segment where the speaker mentions \"their card\" and observe the corresponding physical action. <Unified_Latent><latent_embeddings></Unified_Latent> The speaker explicitly links the action of tapping the deck to squeezing a specific card, which is then revealed. This indicates the tap is not for sound or distraction, but to manipulate the deck. I now check the immediate context of the tap to confirm its role in the trick. <Unified_Latent><latent_embeddings></Unified_Latent> The rising card is presented as the revealed card, confirming the tap is part of the reveal mechanism. The speaker's instruction to \"show them that this is actually their card\" directly follows the tap and rising card, proving the tap is a key step in the visual trick. The action is not about sound or color, but about demonstrating the card's movement and identity. \n\\boxed{C}

Figure 15: LatentOmni example: Reasoning. LatentOmni accurately anchors task-relevant audio-visual frames within the latent space. As demonstrated by the attention visualization, deeper colors indicate higher attention weights precisely localized on the key multimodal clues.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we present our claims and contributions. We propose LatentOmni, a cross-modal reasoning framework that reconstructs inputs within a latent space; OSPE, a novel position embedding method and LatentOmni-Instruct-35K, a dataset for latent reconstruction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendices G

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: Our paper focuses on empirical methodology and architectural design; we do not present formal theorems, lemmas, or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive details in Sec. 4.1 and Appendices B, which enables reproduction of the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have publicly released the training and inference code, data construction pipeline, and LatentOmni-Instruct-35K dataset. The project repository provides the implementation and instructions for reproducing our experiments: `https://github.com/yfanDai/LatentOmni`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Comprehensive details are provided in Sec. 4.1 and Appendices B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: In our benchmark evaluation, we employ standardized configurations and deterministic evaluation protocols. Since the primary metric is the absolute accuracy on a fixed set of questions, reporting statistical dispersion measures such as standard deviation, variance, or error bars is not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments in this study were conducted on a server equipped with eight NVIDIA H800 GPUs and 2,000 GB of system memory.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our framework and dataset advance cross-modal reasoning by enabling more efficient latent-space reconstruction supervision and providing high-quality AV interleaved CoT data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our framework and dataset do not have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Where applicable, we credit and cite the owners and authors of prior works, along with their associated codebases or datasets, and specify the license under which these assets were made publicly available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce LatentOmni-Instruct-35K, a dataset for latent reconstruction, and document its synthesis pipeline comprising three stages: AVQA Data Synthesis & Filtering, Segment-Level Caption Synthesis, and AV-Interleaved Reasoning Trajectory Synthesis; more detailed documentation, including license and usage limitations, will be provided upon official release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not involve human subjects or study participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: As described in Section 3.3 and Appendix B, we employ existing large language models to synthesize, evaluate, and refine audio-visual question-answer pairs, segment descriptions, and reasoning trajectories for constructing LatentOmni-Instruct-35K. We document the roles of these models and their corresponding instructions in the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.