

Not All Modalities Are Equal: Instruction-Aware Gating for Multimodal Videos

Bonan Ding¹ Umair Nawaz¹ Ufaq Khan¹ Abdelrahman M. Shaker¹
Muhammad Haris Khan¹ Jiale Cao² Jin Xie³ Fahad Shahbaz Khan^{1,4}

¹Mohamed bin Zayed University of Artificial Intelligence, ²Tianjin University,
³Chongqing University, ⁴Linköping University

bonan.ding@mbzuai.ac.ae, umair.nawaz@mbzuai.ac.ae

Project Page: <https://your-project-page.com>

Abstract

Pre-trained video large language models excel at visual reasoning. However, they struggle when videos arrive with auxiliary streams, such as audio, depth map, or high-frame-rate. In such a scenario, uniform fusion induces modality interference, allowing irrelevant channels to distract the model. To address this issue, we present a unified multimodal video understanding framework, named UniMVU, that performs instruction-aware fusion across video, audio, depth map, or any other modality inputs via two levels of dynamic gating: Feature-level gates emphasize salient regions within each modality, whereas modality-level gates re-weight whole streams; both are conditioned on the text instruction to adaptively balance modality importance. Our UniMVU combines cross-modal self-attention with instruction-driven inner-modality gating module and a modality-level gating module with control token; for time-aligned streams we further adopt a fast-to-slow fusion scheme that reduces redundancy. Across six benchmarks (AVQA, AVSD, Music-AVQA, ScanQA, SQA3D and MVBench), our UniMVU achieves consistent gains over static-fusion baselines achieving gains as high as 13.5% in terms of CIDEr metric. Further, our analysis shows that the gating mechanism aligns with the human-interpretable modality relevance, and ablations show the contributions of inner-modality and modality-level gating. Our UniMVU provides a simple, unified recipe for instruction-aware multimodal video understanding that scales to diverse modalities without hand-crafted fusion rules.

1. Introduction

Video large language models (Video LLMs) have recently shown promising capabilities in visual reasoning and open-ended video question answering [2, 6, 11, 19, 28]. However, adapting these models to *multimodal* video understanding,

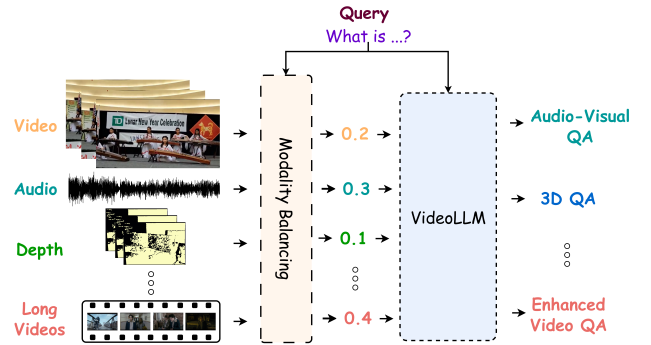


Figure 1. **Illustration of our context-dependent modality weighting.** For each sample, our UniMVU assigns context-dependent modality weights to video, audio, depth, or long-video segments, so that the Video-LLM focuses on the most relevant information, enabling one model to handle Audio-Visual QA, 3D QA, and long-video QA tasks.

where a video may be accompanied by audio, 3D depth cues, or other modality information remains challenging [6, 15, 24, 27]. Many downstream tasks require integrating such side-channel modalities beyond the original video-text training pairs [6, 15]. For instance, audio-visual question answering needs both the visual stream and the synchronized audio to comprehend content [6, 27], while 3D scene understanding can leverage depth or point cloud information in addition to RGB frames [10, 24]. In these settings, the relative importance of each modality varies with the task and the user’s instruction: a query about speech content relies primarily on audio, whereas a query about color or motion is predominantly visual [15, 27]. This variability creates a modality-importance imbalance: naively fusing all streams can inject redundancy or noise from irrelevant channels and under-utilize the informative ones [6, 15, 27].

Recent studies have extended Video-LLMs beyond RGB to include audio and 3D, but many systems still rely on static, uniform fusion of modalities (e.g., fixed cross-attention or average pooling) without considering query rel-

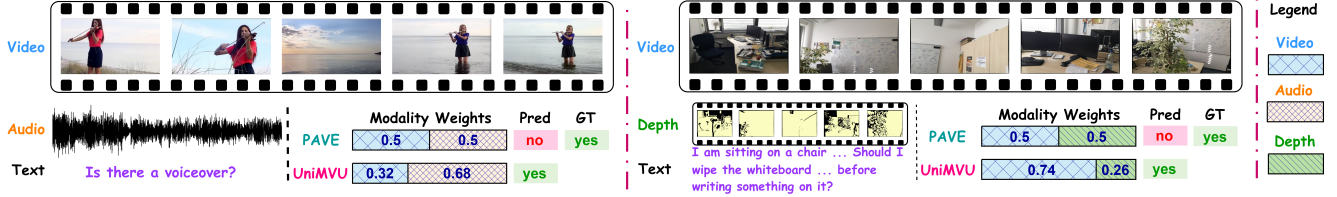


Figure 2. **Qualitative comparison between our UniMVU and uniform weighting approach.** The core idea of UniMVU is to adapt modality weights to the input query, while PAVE [15] uses a uniform weighting approach. The example on the left is audio-centric, but PAVE split modality weights equally across video and audio and predicts **no**, whereas **UniMVU** up-weights audio over video and answers **yes**, matching the ground-truth. The example on right is for 3D-VQA task, where the query depends on spatial visual cues. PAVE’s uniform weighting again leads to **no**, whereas **UniMVU** emphasizes the video stream over depth as depth modality does not contribute significantly, predicting the correct answer **yes**. Our question-conditioned modality gating allows to focus on the most informative modality instead of averaging them, yielding more reliable answers across diverse tasks. Additional results are in Fig. 6 and suppl. material.

evance [2, 6, 19, 28]. This can likely cause modality interference (see Fig. 2), where irrelevant channels dominate or distract from relevant cues, yielding marginal gains from additional modalities [4, 16, 26]. Delegation-style or query-aware fusion schemes aim to tackle these issues [7, 15, 20, 27]. However, most existing methods lack explicit instruction-driven re-weighting across modalities.

To address these problems, we propose UniMVU (see Fig. 1), a unified multimodal video understanding framework that extends LMMs with a flexible instruction-aware fusion of video, audio, 3D, or any other modality information. Our UniMVU addresses modality-importance imbalance by learning dynamic gates that modulate contributions at three levels: (i) token-level gates emphasize the most relevant regions within each modality, (ii) modality-level gates adjust the global weight of each entire modality stream and (iii) provide a unified, instruction-following LMM that supports mixed-modality inputs and joint multi-task training/inference, allowing seamless training and deployment across heterogeneous datasets and objectives.

Both sets of gates are conditioned on the textual instruction and task context, enabling the model to emphasize the modalities implicitly indicated by the query and attenuate the rest. This design reduces redundancy and suppresses noise, yielding more precise cross-modal reasoning without sacrificing generality. In summary, our work tackles the core challenge of adaptive multimodal fusion in LMMs with a novel instruction-driven dynamic gating approach.

1.1. Contributions

We present **UniMVU**, a unified large multimodal video understanding model that ingests mixed-modality inputs including video, audio, and depth/3D or any other modality together with text instruction, and enable *jointly* training across modality-mixed datasets and tasks, enabling a single model to generalize across diverse modality configurations. The core of UniMVU is an instruction-driven question-conditioned dynamic gating mechanism that operates at two

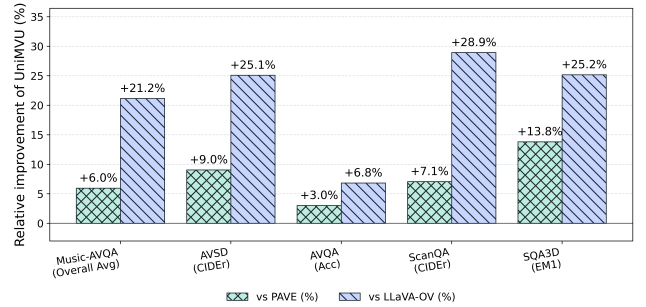


Figure 3. **Relative improvements of UniMVU over PAVE [15] and LLaVA-OV [11] across five benchmarks.** UniMVU achieves consistent gains across these benchmarks.

levels i.e., feature and modality, allowing the model to tailor fusion to each query and to overcome the limitations of fixed or single-level fusion. The learned gates explicitly re-balance modalities on a per-query basis (e.g., up-weighting audio for “what was said?” and combining RGB with depth for “what action is occurring?”), and we show that their behavior aligns with human-interpretable notions of modality relevance. We validate UniMVU on a broad suite of audio-visual and 3D-enhanced video-QA benchmarks, including AVQA, AVSD, Music-AVQA, ScanQA, SQA3D and MVBench, demonstrating consistent improvements over traditional fusion baselines and strong results on modality-specific queries (see Fig. 3).

2. Related Work

LMMs for open-ended video understanding. Large multimodal models have rapidly advanced video reasoning and open-ended video QA by leveraging language modeling with visual-temporal inputs. Flamingo [2] demonstrated few-shot transfer with cross-attention over visual tokens, while instruction-tuned video systems such as Video-ChatGPT [19] and Video-LLaMA [28] extend LMMs

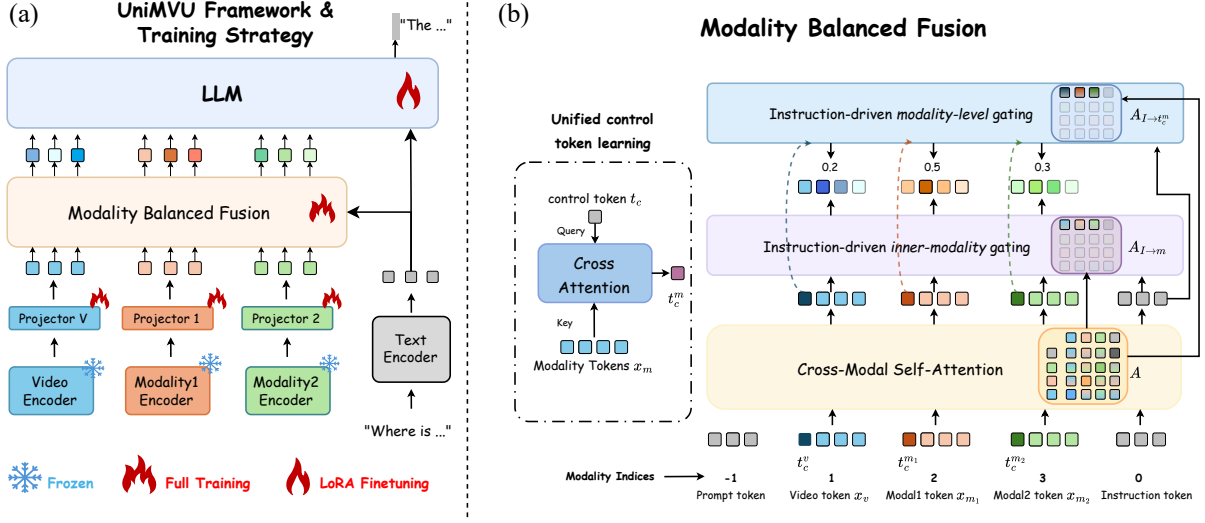


Figure 4. (a) **UniMVU Framework and Training Strategy.** UniMVU is a unified multimodal video LMM with modality-specific encoders and projectors that extract features from arbitrary inputs, which together with text questions or instructions are fed into the proposed modality-balanced fusion module to produce instruction-gated tokens. These tokens are then passed to a pre-trained LLM to generate answers. We froze all encoders, fully trained the projector with the Modality Balanced Fusion block, and simultaneously fine-tuned the LLM using LoRA. (b) **Detailed Structure of Modality Balanced Fusion of UniMVU.** The Balanced Fusion block consists of three modules i.e., Cross-Modal Self-Attention, Instruction-driven inner-modality gating, and Instruction-driven modality-level gating.

to video through temporal connectors and instruction-following recipes. VideoLLaMA-2 [6] further integrates a spatial-temporal connector and an audio branch to improve audio-video understanding in QA and captioning. Despite strong progress, these models are primarily trained on RGB-text pairs and assume a fixed fusion mechanism.

Adapting Video LLMs to additional modalities. To inject signals beyond RGB (e.g., audio, depth), recent approaches explore lightweight or delegated fusion. PAVE patches side-channel tokens into frozen Video-LLMs at the *patch* level to exploit auxiliary modalities with minimal retraining [15]. CAT learns more fine-grained, question-conditioned cross-modal interactions to extract relevant clues per modality before decoding [27]. However, many pipelines still combine streams either uniformly or through fixed connectors, without explicitly estimating each modality’s contribution to the *current* instruction. This often yields modality interference, where irrelevant channels distract the decoder [26]. Our work complements these efforts by learning instruction-conditioned gates that adaptively reweight modalities at both token and modality levels.

3. Method

Our approach unifies multimodal video understanding by coupling modality-specific encoders and projectors with an instruction-aware fusion module that precedes a pre-

trained LLM. The fusion module comprises: (1) **Self-Attention for Cross-Modal Interaction** enables token-level interactions among modalities and instruction tokens; (2) **Instruction-Driven Intra-Modality Gating** summarizes each modality into an instruction-selected token set by reweighting tokens according to instruction-guided relevance; and (3) **Instruction-Driven Modality-Level Gating** assigns a learnable, instruction-conditioned factor to each modality via a dedicated control token, thereby modulating the contribution of each modality. The fused tokens are then passed to the LLM for generating answers. Next, we present each component in detail.

3.1. Self-Attention for Cross-Modal Interaction

Unified Token Sequence Construction. We first project all input modalities into a common embedding space by their encoders and projectors and merge them into a single token sequence alongside the text instruction tokens and prompts. For each non-text modality m , given its encoder characteristic vectors, we apply a learned linear projector $P_m : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^d$ to map those features to the hidden dimension. We then flatten the features of each modality into a sequence of tokens. To distinguish the modality belonging in the token sequence, we maintain a modality index for each token: $m_i = 0$ for instruction (text) tokens, $m_i = k > 0$ for tokens from the k -th modality, and $m_i = -1$ for padding tokens and system prompts. This unified sequence represen-

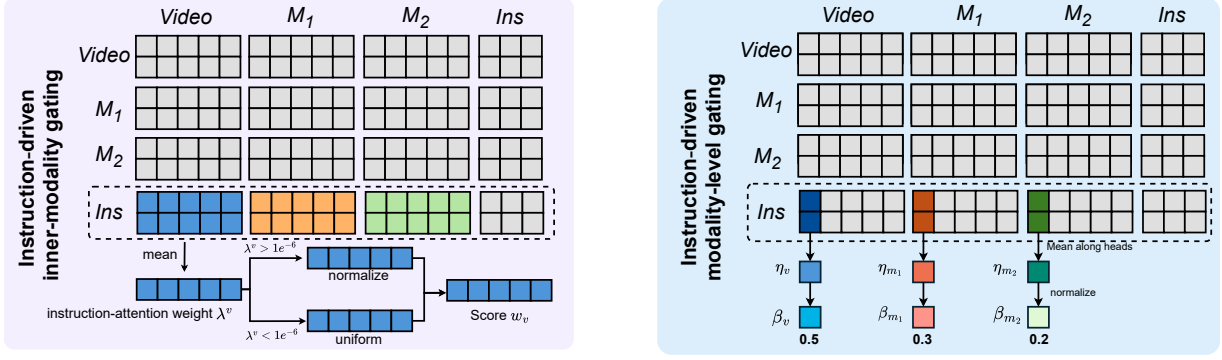


Figure 5. **Instruction-Driven Inner-Modality Gating & Modality-Level Gating.** The grids denote token-level attention maps partitioned by modality, where rows are query tokens and columns are key tokens. **Left:** inner-modality gating uses instruction-to-modality attention to compute normalized token weights within each modality. **Right:** modality-level gating uses instruction-to-control-token (first token in each modality sequence) attention to produce one scalar weight for each modality.

tation enables subsequent attention layers to reason jointly over text and all modalities, surpassing the need for task- or modality-specific modules, such as those in [15, 27].

Cross-Modal Self-Attention Layer. To enable rich cross-modal interactions and prepare for the following instruction-aware weighting, we introduce a Self-attention-based cross-modal interaction layer operating on the unified sequence. This layer is built on a multi-head self-attention [22, 23]. Let $\mathbf{H} \in \mathbb{R}^{L \times d}$ denote the input unified sequence with L tokens. We first compute the projections of query, key, and value for each token as

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{H}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{H}\mathbf{W}_V, \quad (1)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learned projection parameters. We then apply rotary positional embeddings [21] to the projected queries and keys to encode their positions. Then multi-head attention is computed as

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right), \quad (2)$$

where d_k is the dimension of each head and $\mathbf{M}$ denotes an attention mask that masks out prompt and padding tokens. The resulting attention matrix $\mathbf{A} \in \mathbb{R}^{h \times L \times L}$ contains the attention weights between every pair of tokens. We then compute the attended values as

$$\mathbf{O} = \mathbf{A}\mathbf{V}. \quad (3)$$

Therefore, the output vector $\mathbf{O} \in \mathbb{R}^{L \times d}$ can fuse information from across all tokens. Afterwards, we perform Instruction-Driven Inner-Modality Gating (see § 3.2) and Instruction-Driven Modality-Level Gating (see § 3.3), which can adaptively fuse the information from different modalities. Then, we add a residual connection and layer

normalization to the attention output, followed by a feed-forward network (FFN) and another residual addition. This produces final updated token representations $\hat{\mathbf{H}}$. To preserve the integrity of the instruction tokens, we propose to keep instruction embeddings untouched. In particular, after the attention and FFN layers, we replace the output for any instruction token with its original input embedding. Formally, if $\mathbf{H}$ is the input to this cross-modal fusion module, the output $\hat{\mathbf{H}}_{\text{out}}$ for each position i can be calculated as

$$\hat{\mathbf{H}}_{\text{out},i} = \begin{cases} \mathbf{H}_i, & \text{if } m_i \in 0, -1, \\ \hat{\mathbf{H}}_i, & \text{otherwise,} \end{cases} \quad (4)$$

where m_i is the modality index of token i . In words, tokens belonging to multi-media are allowed to absorb information from the instruction through $\hat{\mathbf{H}}$, but the instruction tokens themselves remain unchanged, and do not incorporate modality information. This ensures one-way information flow: modality tokens are conditioned on the textual instruction, while the instruction representation stays intact and uncontaminated by modality noise.

3.2. Instruction-Driven Inner-Modality Gating

To better fuse the features within modality relevant to the instruction, we introduce an inner-modality attention mechanism to refine each modality’s token sequence by the proposed instruction-guided weighting strategy. Specifically, we leverage the attention map from the cross-modal self-attention to estimate the importance of each token with respect to the instruction. Let $\mathcal{I}$ be the set of instruction tokens and $\mathcal{M}_m$ be the set of tokens from modality m . Using the attention matrix $\mathbf{A}$ from Eq. 2, we define the k -th token $x_k^m \in \mathcal{M}_m$, its aggregated instruction-attention weight λ_k^m as the average attention mass over H heads coming from

instruction queries to that token, which can be written as

$$\lambda_k^m = \frac{1}{H} \sum_{h=1}^H \sum_{i \in \mathcal{I}} A_{i \rightarrow k}^{(h)}. \quad (5)$$

Then, we derive a normalized attention score for token k relative to other tokens in the same modality m as

$$w_{m,k} = \frac{\lambda_k}{\sum_{j \in \mathcal{M}_m} \lambda_j}. \quad (6)$$

In the case where the instruction has no preference, e.g. $\sum_{j \in \mathcal{M}_m} \lambda_j < 1e^{-6}$, because it paid virtually no attention to modality m , we fall back to a uniform weight $w_{m,k} = \frac{1}{|\mathcal{M}_m|}$ for all tokens in that modality. This uniform baseline ensures stability. If the instruction provides no preference, all tokens contribute equally by default.

3.3. Instruction-Driven Modality-Level Gating

Unified Learnable Control Token. We add a learnable control embedding t_c and insert it into the beginning of each modality token sequence. Before the cross-modal self-attention layer, each control token first attends to the tokens of its own modality using a single cross-attention block to get modality-specific control token t_c^m , as shown in Fig. 4(b). In this way, the unified control token collects information from each modality first.

Modality-Level Gating. After the cross-modal self-attention and inner-modality gating, we compute a scalar importance weight for each modality by cross-attention operation between the control token t_c and the instruction tokens $\mathcal{I}$. For modality m , let score η_m be derived from the cross attention of all instruction tokens as queries to t_c^m , which can be written as

$$\eta_m = \frac{1}{H} \sum_{h=1}^H \sum_{i \in \mathcal{I}} A_{i \rightarrow t_c^m}^{(h)}. \quad (7)$$

These η_m values are then normalized across modalities to produce modality-level gating coefficients β_m as

$$\beta_m = \frac{\eta_m}{\sum_1^M \eta_{m_1}}. \quad (8)$$

Finally, the modality gating coefficients β_m and the inner-modality gate score w_m from Eq. 6. The final outputs are then combined as

$$\hat{\mathbf{H}}_m = \mathbf{O}_m + \beta_m \times (w_m \cdot \mathbf{O}_m), \quad (9)$$

where $\mathbf{O}_m$ is the original self-attention output of modality m and w_m with β_m add the gated adjustments for both inner-modality and modality-level, resulting in instruction-balanced multimodal representation. Intuitively, β_m adjusts the influence of that modality, while w_m further adaptively emphasizes the important tokens of the modality m .

4. Experiments and Results

4.1. Experimental Setup

We evaluate UniMVU on three tasks requiring different modality combinations. (i) *Task A (Audio-Visual QA)*. It includes Music-AVQA [12], AVSD [1], and AVQA [25], covering music, dialog, and multiple-choice settings. (ii) *Task B (3D QA)*. It focuses on spatial reasoning in 3D scenes using ScanQA [3] and SQA3D [18]. (iii) *Task C (Video QA)*. It uses MVBench [14], a multi-video QA benchmark for fine-grained scene understanding and reasoning.

Evaluation Metrics. We use the standard metrics for each dataset: generative QA tasks (AVSD, ScanQA) use captioning metrics (CIDEr, BLEU-n, METEOR, ROUGE-L) plus exact match (EM@1); multiple-choice or classification tasks (AVQA, Music-AVQA) use accuracy; SQA3D uses EM@1 under original and refined protocols with breakdowns by question type; and MVBench uses accuracy on five subtasks (state change, fine-grained pose, object shuffle, action prediction, action sequence) and their average.

Implementation details. We compare UniMVU with PAVE (reproduced via official code) [15] and LLaVA-OneVision fine-tuned with concatenated multi-modal input or video-only input [11]. We follow the training pipeline of PAVE [15] with the same frozen modality encoders. We train the LMM, the projection layers, and our fusion blocks, where the LLM is initialized with LLaVA-OneVision [11], and fine-tuned via LoRA. We uniformly sample 32 frames per video. For audio-visual QA, audio is processed by ImageBind [9] (one token per second) and concatenated with the visual tokens. For 3D QA, we use an LLaVA-3D-based [31] encoder to extract per-frame RGB-D features aligned with video frames, summing them with video tokens to reduce memory overhead. For video QA, following PAVE, the Training data is a subset of LLaVA-Video-178K (videos > 1min, 2 QA pairs each), totaling 57k videos and 114k QA pairs. We add dense video features (2fps, 2x2 spatial resolution) as an auxiliary modality. We train UniMVU either per task or as a unified model on all tasks. More details are presented in the supplementary material.

4.2. Task A: Audio-Visual QA

Here, we assess the complementary facets of audio-visual reasoning using three public benchmarks: **Music-AVQA**, **AVSD**, and **AVQA**. Table 1 shows that UniMVU consistently outperforms PAVE [15] across all three audio-visual benchmarks and both model scales. At 0.5B, UniMVU improves Music-AVQA overall accuracy from 77.3 to 81.9 (+4.6), AVQA accuracy from 89.6 to 92.3 (+2.7), and AVSD CIDEr from 134.9 to 147.1 (+12.2). At 7B, the gains remain consistent, improving Music-AVQA overall accuracy from 81.9 to 83.7 (+1.8), AVQA accuracy from 93.4 to 94.3 (+0.9), and AVSD CIDEr from 151.6 to 165.1 (+13.5).

Size	Method	Music-AVQA				AVQA	AVSD	
		Audio Avg.	Visual Avg.	AV Avg.	Overall Avg.	ACC (%)	ROUGE.L	CIDEr
7B	CAT-FT [27]	84.9	86.1	83.2	84.3	92	—	—
	LLaVA-OV-FT (video-only) [11]	75.4	89.3	72.3	77.4	90.8	—	124.9
	PAVE* [15]	79.1	92.7	77.8	81.9	93.4	38.5	151.6
	UniMVU	81.7	93.5	79.8	83.7	94.3	39.8	165.1
	Δ over PAVE* (UniMVU)	+2.6	+0.8	+2.0	+1.8	+0.9	+1.3	+13.5
0.5B	PSTP-Net [13]	—	—	—	—	90.2	—	—
	VAST [5]	—	—	—	80.7	—	—	—
	AVAF-Net [30]	78.1	82.3	72.1	75.9	—	—	—
	AV-Master [29]	79.9	86.5	74.2	78.5	91.4	—	—
	LLaVA-OV-FT (video-only) [11]	69.6	76.3	62.8	67.6	86.4	—	117.6
	LLaVA-OV-FT* (video-audio concat) [11]	76.2	89.1	72.4	77.5	89.9	35.7	127.8
	PAVE* [15]	75.9	88.6	72.4	77.3	89.6	36.5	134.9
	UniMVU	79.5	91.8	76.7	81.9	92.3	38.2	147.1
	Δ over PAVE* (UniMVU)	+3.6	+3.2	+4.3	+4.6	+2.7	+1.7	+12.2

Table 1. **Task A: Audio-Visual QA** results on **Music-AVQA** with modality-averaged accuracy (acc.), **AVSD** with ROUGE.L, CIDEr, and **AVQA** with accuracy. “*” means our reproduced results using the publicly code from authors. Δ refers to the improvement gains while FT shows the fine-tuned variant. Our approach achieves consistent improvement in performance, compared to existing works in literature.

Notably, the improvements are balanced across audio-only, visual-only, and joint audio-visual splits on Music-AVQA, suggesting that UniMVU does not overfit to a single modality but instead improves both unimodal competence and cross-modal reasoning. On AVSD, the much larger gain in CIDEr than in ROUGE-L at both scales indicates better semantic grounding and answer relevance, which is consistent with our instruction-aware modality reweighting. In addition, improvements are most pronounced in the low-capacity regime i.e., UniMVU’s 0.5B model closes much of the gap to 7B on AVSD and AVQA, suggesting our fusion is especially valuable when model capacity is tight.

UniMVU also outperforms LLaVA-OV-FT (video-only). For instance, at 0.5B, the overall performance is improved from 67.6 to 81.9 on Music-AVQA, and the improvement is from 117.6 to 147.1 on AVSD CIDEr. This highlights that stronger audio-visual coupling is essential beyond scaling a visual backbone. It also improves over a naive video-audio concatenation baseline (“video-audio concat”), which reaches 77.5 overall score on Music-AVQA and 127.8 CIDEr on AVSD at 0.5B. UniMVU lifts these to 81.9 and 147.1, respectively. Together, these comparisons show that the gains are due to *how* information is fused and routed, not just how much information is present.

4.3. Task B: 3D QA

In this task, we evaluate grounded 3D reasoning with two complementary benchmarks, including **ScanQA** and **SQA3D**. To reflect standard evaluation practice, we show both the Original scoring and a Refined variant that normalizes minor textual mismatches. UniMVU consistently improves over the PAVE across scales and metrics as shown in Table 2. At 0.5B, UniMVU lifts ScanQA EM@1 with

an absolute gain of 2.4 and 2.8 on the original and refined metrics, respectively. It also improves the quality of the generation across the board of different metrics, having a higher absolute gain of 6.0 in terms of CIDEr. On SQA3D, UniMVU shows comparable gains on both the original and refined metrics. For instance, with 0.5B, our method has the improvement of 6.7 in terms of EM1 using the original metric, has the improvement of 6.5 using the refined metric. At 7B, improvements are smaller but steady. For instance, it has the absolute gain of 3.4 on SQA3D. Relative to prior 7B models, UniMVU surpasses LLaVA-3D and Scene-LLM on both datasets. For instance, on ScanQA EM@1 with refined metric, UniMVU has the EM1 score of 48.8, while the LLaVA-3D has the score of 45.0. On SQA3D EM@1 with refined metric, our method has the EM1 score of 60.8, while the LLaVA-3D has the score of 57.6.

The large SQA3D gains particularly on *What* (+24% rel.) and *How* (+15% rel.) at 0.5B point to better spatial relations and interaction reasoning, which are precisely the regimes where 3D cues disambiguate similar-looking frames. Similarly, on ScanQA, CIDEr outperforms BLEU, which suggests few off-topic generations and stronger content grounding. This aligns with UniMVU’s context-aware routing, which upweights the modality that is most reliable for the query. The larger gains at 0.5B than at 7B (e.g., +6.0 CIDEr, +6.7 EM@1) suggest that our fusion is especially effective under tight capacity.

4.4. Task C: VideoQA

This task evaluates question answering that depends on visual dynamics rather than static appearance. The goal is to test whether a model can extract and use temporally localized evidence across frames to answer fine-grained ques-

Size	Method	ScanQA					SQA3D			
		EM1	BLEU-4	METEOR	ROUGE.L	CIDEr	EM1	What	Is	How
7B	LLaVA-3D-7B [31]	27.0 (45.0)	14.5	20.7	50.1	91.7	55.6 (57.6)	—	—	—
	Scene-LLM-7B [8]	27.2	12	16.6	40	80	54.2	—	—	—
	LLaVA-OV-FT (video-only) [11]	27.4 (46.3)	—	13.5	47.4	95.1	55.8 (58.1)	—	—	—
	PAVE* [15]	28.9 (48.2)	16.0	19.8	48.8	102.4	57.6 (59.9)	52.3 (56.9)	69.2 (69.9)	56.3 (57.4)
	UniMVU	29.6 (48.8)	16.0	19.8	49.0	102.7	59.4 (61.6)	53.4 (57.7)	75.9 (76.8)	55.9 (56.1)
Δ over PAVE* (UniMVU)		+0.7 (+0.6)	+0.0	+0.0	+0.2	+0.3	+1.8 (+1.7)	+1.1 (+0.8)	+6.7 (+6.9)	-0.4 (-1.3)
0.5B	SceSU [10]	25.1	13.2	14.9	35.5	69.6	46.8	32.2	64.9	46.2
	DSPNet [17]	26.5	15.4	15.7	39.3	78.1	50.4	38.2	66	51.2
	LLaVA-OV-FT (video-only) [11]	20.5 (36.3)	6.5	14.3	36.9	70.5	44.1 (45.7)	—	—	—
	LLaVA-OV-FT (video-3d concat) [11]	10.2 (24.3)	4.90	7.4	20.2	34.9	— (—)	—	—	—
	PAVE* [15]	23.5 (40.4)	12.7	17.1	42.7	84.9	48.5 (50.6)	37.5 (41.6)	61.0 (62.1)	50.1 (50.3)
	UniMVU	25.9 (43.2)	13.5	18.0	44.7	90.9	55.2 (57.1)	46.5 (50.6)	67.6 (68.4)	57.4 (57.6)
	Δ over PAVE* (UniMVU)	+2.4 (+2.8)	+0.8	+0.9	+2.0	+6.0	+6.7 (+6.5)	+9.0 (+9.0)	+6.6 (+6.3)	+7.32 (+7.30)

Table 2. **Task B: 3D QA** — results on **ScanQA** validation set and **SQA3D** test set. We report top-1 Exact Match (EM@1), BLEU-4, METEOR, ROUGE.L for ScanQA, and EM@1 for SQA3D along with "What", "Is" and "How" questions. Each metric cell shows *Original* followed by *Refined* in parentheses; the refined value is lightly shaded. "*" means our reproduced results using the publicly code from authors. Δ shows the improvement of our method against PAVE.

Size	Method	MVBench					
		SC	FGP	OS	AP	AS	Avg.
0.5B	LLaVA-OV [11]	37.5	49.0	33.0	—	—	45.5
	PAVE* [15]	41.0	50.0	32.0	43.0	46.5	44.5
	UniMVU	43.0	50.5	30.0	53.5	52.0	46.7
	Δ over PAVE* (UniMVU)	+2.0	+0.5	-2.0	+10.5	+5.5	+2.2
	VideoChat2-7B	44.0	49.0	42.5	47.5	66.0	51.1
7B	VideoLLaMA2.1-7B	—	—	—	—	—	57.3
	PAVE-7B*	51.0	53.5	39.5	70.5	70.7	57.1
	LLaVA-OV-7B (Baseline)	52.0	53.0	35.5	—	—	56.7
	UniMVU	51.0	54.5	38.5	71.0	77.0	58.0
	UniMVU [†]	51.5	58.0	39.5	76.5	76.1	59.5

Table 3. **Task C: VideoQA** — **MVBench** results with 0.5B-size and 7B-size models. We report performance on state change (SC), fine-grained pose (FGP), object shuffle (OS), action prediction (AP), and action sequence (AS) subtasks. "*" denotes our reproduced results using the publicly code from authors, [†] denotes the unified training result, and Δ shows the improvements over the PAVE method.

tions about events in the sequence. We evaluate **MVBench** at both 0.5B and 7B scales, reporting subtask scores in Table 3. At 0.5B, UniMVU achieves the best average score of 46.7, outperforming both PAVE and LLaVA-OV. The largest gains appear on temporally demanding subtasks, including AP with an absolute improvement of 10.5 and AS with a gain of 5.5 over PAVE. At 7B, UniMVU also remains competitive, achieving 58.0 on average, while the unified model further improves the score to 59.5.

UniMVU’s fusion block routes attention across complementary video streams, including global clip tokens and dense low-resolution tokens, allowing the model to emphasize temporal evidence when the query depends on motion

Method	Audio	Visual	AV	Overall
LLaVA-OV-0.5B-FT-CONCAT	76.2	89.1	72.4	77.5
+ Multi-modality Self-Attention Layer	76.8	90.0	74.1	78.8
+ Inner-modality Gating	77.1	90.2	76.1	80.4
+ Modality-level Gating	79.5	91.8	76.7	81.2

Table 4. **Architectural components** ablation on Music-AVQA. Our proposed components can stably improve the performance.

Method	Audio	Visual	AV	Overall
modality-specific control token (num=1)	75.7	90.8	75.9	79.2
unified control token (num=8)	79.3	91.8	76.1	80.8
unified control token (num=1)	79.5	91.8	76.7	81.2

Table 5. **Control token design** ablation on Music-AVQA. The unified control token with number of 1 has the best results.

and event order. This selective routing helps explain the strong gains on AP/AS and the consistent improvements on other temporal subtasks.

4.5. Ablations and Analysis

Component Ablation. As shown in Table 4, we start from the baseline LLaVA-OV-0.5B-FT-CONCAT. When inserting a multi-modality self-attention layer, it yields a clear overall gain of 1.3% in terms of overall score, indicating that explicit cross-modality interaction helps joint audio-visual reasoning. When further adding inner-modality gating, it achieves the overall score of 80.4, with the largest improvement again on AV 2.0%, suggesting that refining signals within each modality improves cross-modal alignment. Finally, introducing modality-level gating pushes overall score to 81.2.

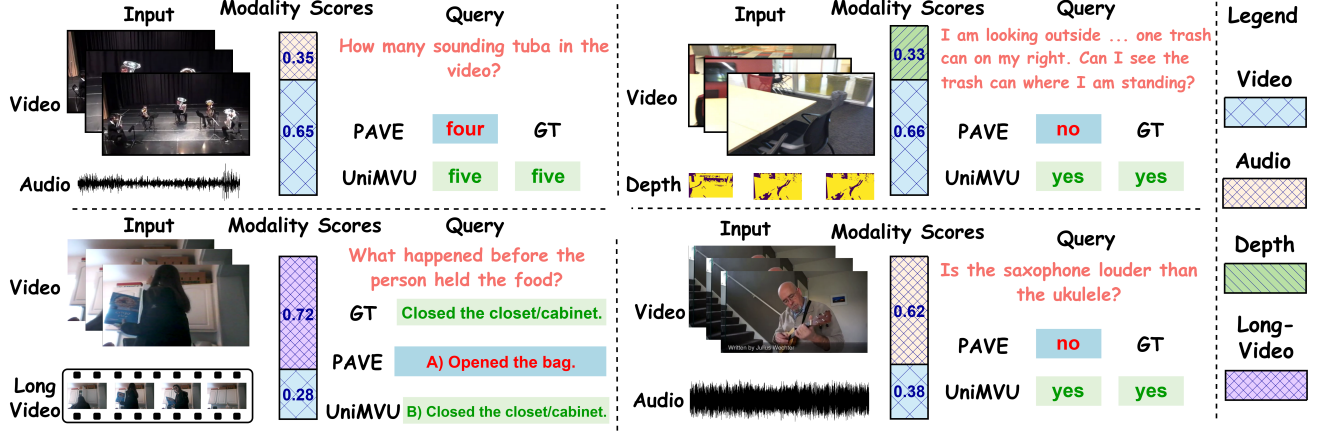


Figure 6. **Qualitative comparison.** UniMVU adapts to the evidence the question demands and up-weights audio for acoustic counting and loudness comparisons, depth/video for spatial visibility, and the long-video stream for temporal ordering, thus yielding correct answers where PAVE [15] fails due to unified modality weight assignment. The legend (right) indicates the pattern used for each modality and the scores are contextual weights produced by our gating module.

Control-Token Design. Table 5 compares different control token design strategies. (i) *Unified vs. modality-specific.* Holding the token count fixed (num=1), a unified control token clearly outperforms a modality-specific one across all splits. This suggests that a single, shared routing signal provides a cleaner interface for cross-modality fusion, avoiding competition between modality-specific tokens. (ii) *Number of control tokens.* With a unified design, using num=1 slightly but consistently beats num=8 with all metrics. Fewer control tokens can reduce sequence length and attention fragmentation, giving a modest accuracy gain at lower computational cost. Therefore, we adopt a single unified control token as default.

Gating Mechanism Design. Table 6 compares three different ways to compute the gates in our two-level fusion. We test two learned variants: (i) *MLP Score*, which applies a single MLP to the instruction-token embedding to produce modality-level gates, and (ii) *Cross-Attn Score*, which uses a lightweight cross-attention between the modality token and the instruction tokens to derive inner-modality gates, against our parameter-free gating. Despite introducing additional parameters, the learned gates do not result in improved accuracy. Our parameter-free design achieves the best overall performance. We hypothesize that learning explicit gating heads may overfit dataset priors, while parameter-free gates offer better a stronger inductive bias regularization with cheaper cross-stream calibration.

4.6. Qualitative Analysis

We present some qualitative results in Fig. 6, where we can observe that UniMVU consistently focuses on the modality that matters for the question, which also performs better than PAVE. We present the two most important modality

Method	#Params(M)	Audio	Visual	AV	Overall
UniMVU(MLP Score)	52.0	77.2	91.3	75.8	80.2
UniMVU(Cross -attn Score)	53.4	78.7	91.2	76.3	80.5
UniMVU	51.8	79.5	91.8	76.7	81.2

Table 6. **Gating mechanisms Design** ablation on Music-AVQA.

scores in the figure. Our UniMVU can assign more attention to important modality data, such as the video and depth modalities in the top-right example.

The top row shows two cases where video is more important. In the top-left example, the question “how many sounding tuba in the video?” depends more on video cues. UniMVU attends more to video and predicts the correct answer, five, whereas PAVE predicts four. The bottom row shows two cases where other modalities, such as long-video and audio, are more important. In the bottom-right example, the question “Is the saxophone louder than the ukulele?” depends more on audio cues. UniMVU attends more to audio and answers correctly: yes.

5. Conclusion

In this paper, we present UniMVU, a unified multi-modal video understanding framework that uses context-adaptive cross-modal gating to fuse diverse inputs. Our method effectively mitigates modality interference by dynamically adjusting the contribution of each modality at both token and modality levels, enabling the model to focus on the most relevant visual, audio, or depth cues. Extensive experiments across a broad range of benchmarks demonstrate that UniMVU consistently outperforms prior methods, achieving state-of-the-art results across all evaluated tasks.

Acknowledgments

The computations were enabled by resources provided by NAISS at Alvis, partially funded by the Swedish Research Council through grant agreement no. 2022-06725; by LUMI, hosted by CSC (Finland) and the LUMI consortium; and by Berzelius, provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre (NSC). This work was also supported in part by the National Key Research and Development Program of China (Grant No. 2022ZD0160400) and the National Natural Science Foundation of China (Grant Nos. 62206031 and 62271346).

References

- [1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019. 5, 2
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2
- [3] Daichi Azuma, Taiki Miyayoshi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 5, 2
- [4] Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025. 2
- [5] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023. 6
- [6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 2, 3
- [7] Zesen Cheng, Sicong Leng, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2
- [8] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2195–2206. IEEE, 2025. 7
- [9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 5
- [10] Yunqi Jiang, Jianwei Zhang, Chaoyang Lin, Yi Yu, and Zhenguo Yang. Scene-guided attention network for spatial understanding in 3d scenes. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 616–624, 2025. 1, 7
- [11] Bo Li, Yuanhan Zhang, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 5, 6, 7
- [12] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118, 2022. 5, 2
- [13] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM international conference on multimedia*, pages 7808–7816, 2023. 6
- [14] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 5, 2
- [15] Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. Pave: Patching and adapting video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3306–3317, 2025. 1, 2, 3, 4, 5, 6, 7, 8
- [16] Naver AI (LLaVA-AV-SSM). Does audio matter for modern video-llms and their benchmarks? *arXiv preprint arXiv:2509.17901*, 2025. 2
- [17] Jingzhou Luo, Yang Liu, Weixing Chen, Zhen Li, Yaowei Wang, Guanbin Li, and Liang Lin. Dspnet: Dual-vision scene perception for robust 3d question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14169–14178, 2025. 7
- [18] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 5, 2
- [19] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 1, 2
- [20] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*, 2023. 2
- [21] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4
- [22] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024. 4

- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [4](#)
- [24] FENGYUN WANG, Sicheng Yu, Jiawei Wu, Jinhui Tang, Hanwang Zhang, and Qianru Sun. 3d question answering via only 2d vision-language models. In *Forty-second International Conference on Machine Learning*, 2025. [1](#)
- [25] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022. [5](#), [2](#)
- [26] Qinghao Ye, Haiyang Xu, et al. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024. [2](#), [3](#)
- [27] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios, 2024. [1](#), [2](#), [3](#), [4](#), [6](#)
- [28] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. [1](#), [2](#)
- [29] Jiayu Zhang, Qilang Ye, Shuo Ye, Xun Lin, Zihan Song, and Zitong Yu. Av-master: Dual-path comprehensive perception makes better audio-visual question answering. *arXiv preprint arXiv:2510.18346*, 2025. [6](#)
- [30] Xujian Zhao, Yixin Wang, and Peiquan Jin. Audio-visual adaptive fusion network for question answering based on contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10483–10491, 2025. [6](#)
- [31] Chenming Zhu, Tai Wang, et al. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. [5](#), [7](#), [2](#)

Not All Modalities Are Equal: Instruction-Aware Gating for Multimodal Videos

Supplementary Material

In this suppl. material, we are going to show (1) additional results of unified training; (2) more detailed experiment implementation; (3) more qualification results.

6. Unified Multi-Task Training

Protocol. We trained a single *unified* UniMVU model on 0.5B model size on all datasets jointly for 2 epochs using α -sampling with $\alpha = 0.5$. Encoders are frozen; projection layers, fusion module and LoRA adapters on the LLM are trainable. At inference time, the same model is evaluated across Tasks A–C without task-specific fine-tuning. The unified training model is indicated as UniMVU[†], the separate training model is indicated as UniMVU*, and Δ means the improvement of UniMVU[†] over PAVE*.

6.1. Task A: Audio–Visual QA

We evaluate on three benchmarks with their standard metrics: *Music-AVQA* (Audio Avg., Visual Avg., AV Avg., and Overall), *AVQA* (Accuracy), and *AVSD* (ROUGE.L and CIDEr). Results are reported in Tab. 7.

UniMVU[†] retains strong audio–visual competence across all three datasets. Relative to baseline PAVE*, UniMVU[†] improves *Music-AVQA* by **+1.3** (Audio Avg.), **+1.6** (Visual Avg.), **+2.4** (AV Avg.), and **+2.0** (Overall); *AVQA* accuracy by **+1.5**; and *AVSD* by **+1.3** on ROUGE.L and **+11.0** on CIDEr, indicating better semantic grounding and content coverage in generation with multi-task ability.

Notably, the cross-modal category shows the most pronounced improvement on *Music-AVQA* (**+2.4** on AV Avg.), while single-modality gains are smaller but consistent (**+1.3** on audio, **+1.6** on visual). This pattern suggests that the model’s gating effectively exploits complementary cues rather than collapsing to a single modality. The absolute audio and visual accuracies remain well separated, yet both improve over PAVE*, evidencing balanced reliance on sound and vision.

Compared to the task-specific trained UniMVU*, unified training is ~ 0.4 – 2.6 points lower across metrics, averaging a modest ~ 1.6 -point drop while still exceeding PAVE* on every metric. This trade-off indicates that the unified objective preserves cross-task generalization and cross-modal alignment without per-task specialization, while keeping a single model that remains competitive across QA and dialog generation.

6.2. Task B: 3D QA

We evaluate on ScanQA (EM@1, BLEU-4, METEOR, ROUGE.L, CIDEr) and on SQA3D (overall EM@1 and a

breakdown by question type *What/Is/How*); results are summarized in Tab. 8.

Relative to baseline PAVE*, the unified training UniMVU[†] delivers consistent gains on *both* EM@1 and generation metrics on ScanQA: **+1.2** EM@1 (and **+1.6** on refined score), **+1.8** BLEU-4, **+0.8** METEOR, **+1.5** ROUGE.L, and **+4.8** CIDEr. On SQA3D, UniMVU[†] improves overall EM@1 by **+2.3** (**+2.0**), with the largest absolute gains on the question types *What* (**+3.4/+2.8**) and *How* (**+2.4/+2.4**); *Is* questions typically closer to attribute checks also improve (**+2.7/+2.6**). These trends indicate that with 3D information, the proposed gating mechanism reliably elevate geometric cues when the instruction implies spatial reasoning, without sacrificing linguistic quality.

Compared with separate training UniMVU*, UniMVU[†] is modestly lower on strict EM@1 and slightly behind on some generation metrics, yet remains competitive overall and achieves the strongest BLEU-4 on ScanQA. Importantly, it still outperforms the PAVE* baseline across all metrics and SQA3D types, evidencing robust multi-task generalization from unified training even without task-specific fine-tuning.

6.3. Task C: VideoQA

We use MVBench as the primary benchmark and report accuracy on state change (SC), fine-grained pose (FGP), object shuffle (OS), action prediction (AP), and action sequence (AS), together with the overall average (Avg.); results are summarized in Tab. 9. Beyond MVBench, the unified model also generalizes to VideoMME and MLVU, where it achieves average accuracies of 45.5 and 47.1, respectively, outperforming our PAVE reproduction (45.0 and 39.65).

UniMVU[†] improves MVBench Avg. by **+4.1** over PAVE, with pronounced gains on long-horizon temporal tasks AP (**+21.0**) and AS (**+16.5**). This pattern indicates that cross-task supervision does not dilute temporal sensitivity; rather, the gating and cross-modal attention preferentially route evidence to frames most relevant to the query, enhancing temporal comprehension without sacrificing short range, frame and object centric cues.

Compared with the separately trained UniMVU*, the unified model remains competitive overall, showing clear advantages on long-range temporal subtasks (AP/AS) while staying close on short-range perception (SC/FGP/OS). In other words, unified multi-task training strengthens temporal comprehension without losing local visual sensitivity, consistent with our instruction-driven gating design that routes evidence to frames most relevant to the query.

Size	Method	Music-AVQA				AVQA	AVSD	
		Audio Avg.	Visual Avg.	AV Avg.	Overall Avg.	ACC (%)	ROUGE.L	CIDEr
7B	PAVE* [15]	79.1	92.7	77.8	81.9	93.4	38.5	151.6
	UniMVU*	81.7	93.5	79.8	83.7	94.3	39.8	165.1
	UniMVU [†]	78.9	92.8	77.2	81.6	92.2	39.5	162.7
	Δ over PAVE* (UniMVU [†])	-0.2	+0.1	-0.6	-0.3	-1.2	+1.0	+11.1
0.5B	PAVE* [15]	75.9	88.6	72.4	77.3	89.6	36.5	134.9
	UniMVU*	79.5	91.8	76.7	81.9	92.3	38.2	147.1
	UniMVU [†]	77.2	90.2	74.8	79.3	91.1	37.8	145.9
	Δ over PAVE* (UniMVU [†])	+1.3	+1.6	+2.4	+2.0	+1.5	+1.3	+11.0

Table 7. Task A (Audio–Visual QA) results on Music-AVQA, AVQA, and AVSD under 7B and 0.5B model scales. [†] denotes the unified multi-task training setting.

Size	Model	ScanQA					SQA3D			
		EM@1	BLEU-4	METEOR	ROUGE.L	CIDEr	EM@1	What	Is	How
7B	PAVE* [15]	28.9 (48.2)	16.0	19.8	48.8	102.4	57.6 (59.9)	52.3 (56.9)	69.2 (69.9)	56.3 (57.4)
	UniMVU*	29.6 (48.8)	16.0	19.8	49.0	102.7	59.4 (61.6)	53.4 (57.7)	75.9 (76.8)	55.9 (56.1)
	UniMVU [†]	29.2 (48.8)	17.83	20.1	49.01	104.2	58.1 (60.4)	52.3 (56.6)	70.7 (71.8)	60.4 (61.1)
	Δ over PAVE* (UniMVU [†])	+0.3 (+0.6)	+1.83	+0.3	+0.21	+1.8	+0.54 (+0.5)	+0.0 (-0.3)	+1.5 (+1.9)	+4.1 (+3.7)
0.5B	PAVE* [15]	23.5 (40.4)	12.7	17.1	42.7	84.9	48.5 (50.6)	37.5 (41.6)	61.0 (62.1)	50.1 (50.3)
	UniMVU*	25.9 (43.2)	13.5	18.0	44.7	90.9	55.2 (57.1)	46.5 (50.6)	67.6 (68.4)	57.4 (57.6)
	UniMVU [†]	24.7 (42.0)	14.5	17.9	44.2	89.7	50.8 (52.6)	40.9 (44.4)	63.7 (64.7)	52.5 (52.7)
	Δ over PAVE* (UniMVU [†])	+1.2 (+1.6)	+1.8	+0.8	+1.5	+4.8	+2.3 (+2.0)	+3.4 (+2.8)	+2.7 (+2.6)	+2.4 (+2.4)

Table 8. Task B (3D QA) results on ScanQA and SQA3D under 7B and 0.5B model scales. [†] denotes the unified multi-task training setting.

Model	SC	FGP	OS	AP	AS	Avg.
PAVE* [15]	41.0	50.0	32.0	43.0	46.5	44.5
UniMVU*	43.0	50.5	30.0	53.5	52.0	46.7
UniMVU [†]	37.5	49.0	30.5	64.0	63.0	48.6
Δ over PAVE* (UniMVU [†])	-3.5	-1.0	-1.5	+21.0	+16.5	+4.1

Table 9. Task C (VideoQA) — unified multi-task training with 0.5B model size results on MVBench.

7. Implementation and Experiment Details

7.1. Tasks and Datasets

Task A: Audio–Visual QA. We train and evaluate on AVSD [1] (open-ended) and on AVQA [25] and Music-AVQA [12] (multiple-choice). AVSD contains 79k question–answer (QA) pairs across 7985 videos (10 questions per video); for evaluation, we use the AVSD@DSTC7 test split with 1000 questions. AVQA provides 40k QA pairs paired with 40k videos (MC); we evaluate on the official evaluation split. Music-AVQA comprises 32k QA pairs over 9277 videos; we report results on its official test split.

Task B: 3D QA. ScanQA [3] includes $\sim$ 41k QA pairs over $\sim$ 800 RGB-D indoor scans. We evaluate on its validation set (4675 questions) covering both positional reasoning and object recognition. SQA3D [18] contains 650 scenes, 6.8k unique situations, 20.4k situation descriptions, and 33.4k

questions; we report performance on its test set (3,519 questions) spanning five question types.

Task C: VideoQA. Following PAVE [15], we train UniMVU on a subset of LLaVA-Video-178K (videos $>$ 1 min; two QA pairs per video), yielding 57k videos and 114k QA pairs. For evaluation, we use MVBench [14], a multi-video QA benchmark, and report per-subtask scores and the overall average.

7.2. Training Set

Data preparation. For fair comparison, we used the pre-processed features of each modality from PAVE [15], which means all modality-specific encoders are kept *frozen*. For the video frames, we follow PAVE by sampling 32 frames uniformly from each video, then encoding these raw RGB frames with A pretrained vision encoder in LLaVA-One-Vision [11]. For audio, features were extracted by ImageBind’s audio encoder to convert the soundtrack into audio tokens. Specifically, we split the audio track into short segments (use a roughly 1-second window per token) and obtain an embedding for each segment. For 3D data, features are extracted by the LLaVA-3D [31]. In practice, the 3D scene is represented as an RGB-D sequence which have the same frame extractor ratio as video. To limit the token count, 3D tokens are added with video features after the fusion module rather than concatenation. This yields a compact set of tokens that carry both appearance and geom-

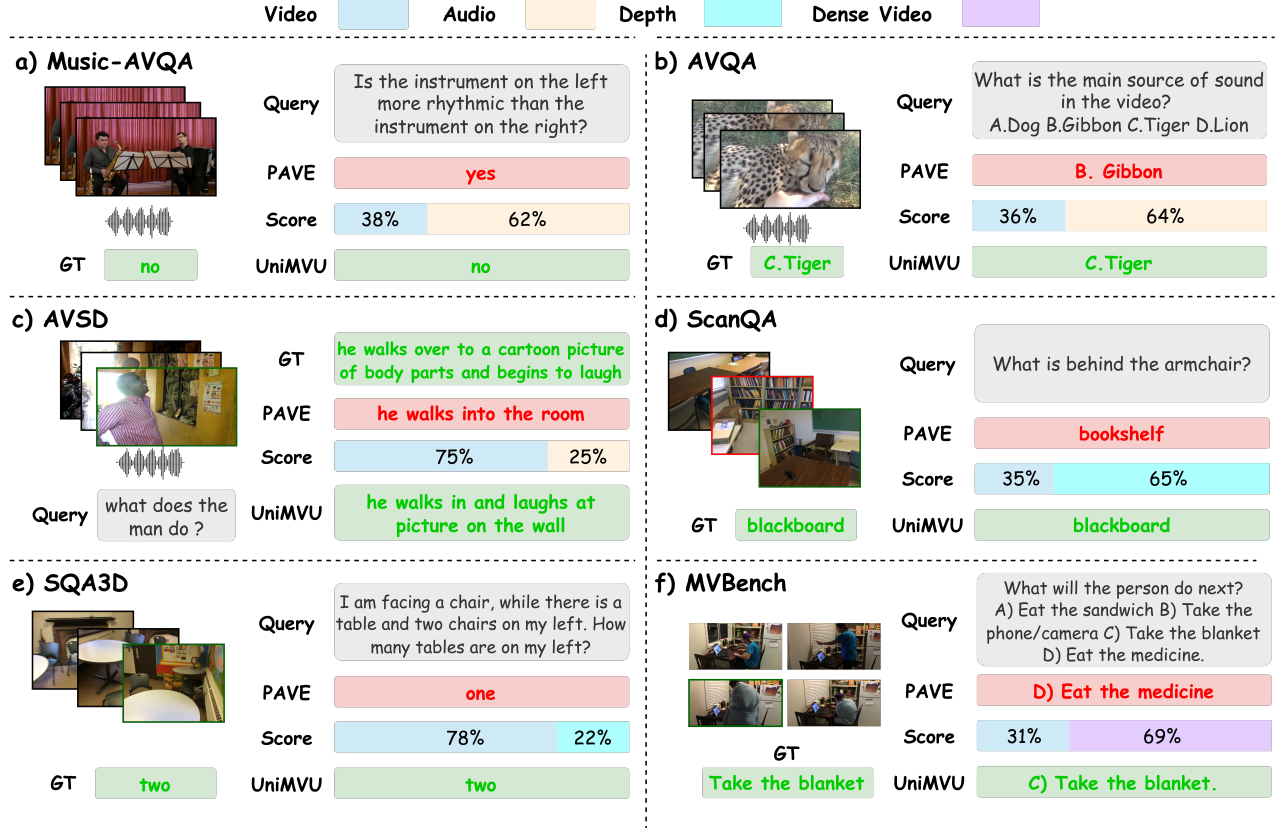


Figure 7. Additional qualitative results on three tasks across six benchmarks. Here, we show that query-conditioned modality balancing fixes PAVE’s errors. (a) **Music-AVQA**: rhythm judgment benefits from **audio** emphasis rather than its spatial understanding, and that is how UniMVU answers “no” where PAVE says “yes.” (b) **AVQA**: identifying the **main source of sound** prioritizes audio, letting UniMVU pick “C. Tiger” instead of PAVE’s “B. Gibbon.” (c) **AVSD**: describing the man’s action relies on **video** motion cues, and that is why UniMVU puts more weightage on video modality and grasps the crucial (laughs at the picture) detail. (d) **ScanQA**: UniMVU leans on the **depth** modality to better understand the broader scene context and locates “blackboard” behind the armchair. Whereas PAVE gets distracted by the “bookshelf” as shown with a red border in the second frame. (e) **SQA3D**: the green-boxed final frames reveal *two* tables and UniMVU’s scoring helps count correctly where PAVE predicts “one.” (f) **MVBench**: anticipating the next action favors the **dense-video** stream; UniMVU chooses “C) Take the blanket,” while PAVE picks an implausible alternative.

etry information of the scene.

We explore two training regimes:

(1) **Task-specific training**, where separate UniMVU instances are finetuned on each individual dataset. We use 1 cross-modal self-attention layer with gating modules; the hidden dimension aligns with LLM and has 14 attention heads. For LoRA layers in the LLM, we use LoRA $r = 64$ and LoRA $\alpha = 128$ for 0.5B models and LoRA $r = 128$ and LoRA $\alpha = 256$ for 7B models. The learning rate are set as $2e^{-5} \sim 4e^{-5}$, warm up rate set as 0.03.

(2) **Unified multi-task training**, a single UniMVU is jointly trained over all datasets and tasks. Each batch is sampled from a mixed, shuffled pool using α -sampling where the probability for dataset i is proportional to N_i^α ; we use $\alpha = 0.5$ to upweight smaller sets. The unified model is trained for 2 epochs over the combined data with the same learning rate ($1e^{-5}$). Unless noted otherwise, we use

AdamW, a cosine schedule with linear warmup, and early stopping on per-task validation. In practice, unified training converges in roughly the same number of iterations as a single-task run, but replaces multiple separate runs with one model and one training pass, yielding lower end-to-end wall-clock and a single deployable checkpoint. All experiments use $2 \times \text{H200}$ GPUs. At inference time, the unified UniMVU can be conditioned on any combination of modalities and instructions, serving as a single model for all evaluated tasks.

Training time reference. For separate training, the per-dataset times (0.5B / 7B) are: Music-AVQA, 2 epochs (12h / 13h); AVQA, 1 epoch (13h / 14h); AVSD, 2 epochs (7h / 8h); SQA3D, 2 epochs (13h / 14h); ScanQA, 2 epochs (12h / 14h); and LLaVA-178K subset, 1 epoch (23h / 28h), totaling **80h** for 0.5B and **91h** for 7B across all separate runs. On the same hardware, unified multi-task training for the

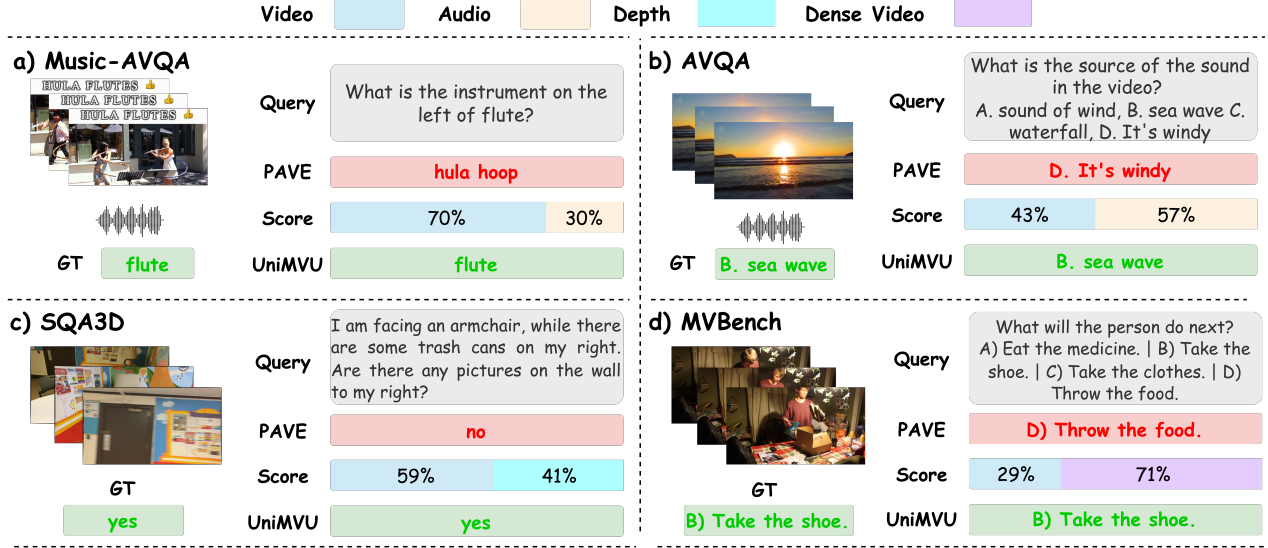


Figure 8. More additional qualitative results on three tasks across six benchmarks.

0.5B model completes in **81h**, offering comparable end-to-end training time while consolidating training into a single run and producing one checkpoint for all tasks.

8. Additional Qualitative Results

We provide illustrative examples from each task category to demonstrate how UniMVU’s instruction-aware gating flexibly prioritizes the correct modality in response to the question as shown in the Fig. 7. This figure shows how UniMVU’s learned modality scores translate into concrete fixes over PAVE by reallocating attention toward the evidence each question truly requires. And more examples are shown in Fig. 8.

In (a) **Music-AVQA**, rhythmicity is an acoustic property, and that is why UniMVU elevates the audio score and answers no, showing sensitivity to temporal patterns in the waveform that are invisible to appearance cues, whereas PAVE’s yes reflects a visual bias due to uniform modality weighting. In (b) **AVQA**, the task is to identify the *main sound source*. Here, the UniMVU again assigns more weight to audio and selects “C. tiger,” while PAVE’s “B. gibbon” reveals overreliance on what looks plausible, not what sounds correct. In (c) **AVSD**, the example requires a grounded description of an action and that is why by up-weighting the video score, UniMVU captures the micro-event (“laughs at the picture on the wall”), which PAVE misses with a generic answer (“walks into the room”). This is also consistent with our CIDEr gains that reward semantic relevance.

Furthermore, in (d) **ScanQA**, we intentionally mark the second frame in red to highlight a visually dominant *bookshelf*. Here, PAVE anchors to this distractor and answers

“bookshelf.” UniMVU on the other hand increases the depth modality score, and effectively widens the 3D context to reason about what is behind the armchair, and recovers the correct answer. In (e) **SQA3D**, the last frames are outlined in green to reveal there are indeed two tables. Even with a higher video share than depth, UniMVU leverages fused visual cues to separate adjacent planar surfaces and counts each table correctly, whereas PAVE undercounts it by one. Finally, in (f) **MVBench**, anticipating the next action requires longer-horizon evidence. UniMVU increases the dense-video weight and selects “C) Take the blanket,” while PAVE’s choice suggests overfitting to a local frame. Across all panels, the absolute percentages are less important than their relative ranking. Whether audio outranks video for acoustic queries, or dense video outranks video for long-horizon reasoning, UniMVU consistently avoids salient but irrelevant distractors and grounds its prediction in the right modality, explaining the systematic corrections we observe over PAVE.