EQUALS: An Audio-Visual LLM with One-Stage Question-Guided Alignment and Flexible Fusion

Anonymous ACL submission

Abstract

Audio-Visual Question Answering (AVQA) has emerged as a crucial task for multimodal reasoning in human-computer interaction, requiring models to align and interpret visual and auditory signals conditioned on natural language questions. Despite recent progress, three key challenges remain: (1) difficulty in locating question-relevant segments within lengthy and redundant video streams, (2) suboptimal audio-visual alignment due to the decoupling between pretraining and task-specific supervision, and (3) insufficient flexibility in fusion strategies across diverse tasks. To address these issues, we propose **EQUALS** (on**E**-stage **Q**uestion g**U**ided **A**lignment and f**L**exible fu**S**ion), a unified end-to-end AVQA framework. **EQUALS** integrates compression, alignment, and fusion within a single stage. Specifically, we interleave optimal transport-based loss modules before and after the question-guided pooling module to achieve fine-grained semantic alignment. To enhance adaptability in fusion, we introduce **FlexFuseMoE**, a mixture-of-experts module that supports early, mid, and late fusion via flexible expert routing. Experiments on MUSIC-AVQA and its challenging variant FortisAVQA demonstrate that EQUALS achieves new state-of-the-art results with interpretability. Our findings highlight the importance of jointly modeling alignment and fusion under explicit question guidance, offering a flexible and scalable solution for audio-visual understanding.

1 Introduction

Audio-Visual Question Answering (AVQA) (Yun et al., 2021; Li et al., 2022; Yang et al., 2022; Liu et al., 2024b; Ma et al., 2024) has emerged as a central problem for natural human–computer interaction, attracting increasing attention in recent years (Han et al., 2024; Lin et al., 2024; Zhou et al., 2025; Bravo et al., 2025). The task requires integrating auditory and visual signals in a question-aware manner, which is particularly challenging in realistic acoustic environments involving speech, music, and ambient sounds. Progress in this area has been largely driven by the development of benchmarks: compared with AVQA (Yang et al., 2022) and Pano-AVQA (Yun et al., 2021), which emphasize object-level semantics and panoramic reasoning, the MUSIC-AVQA series (Li et al., 2022; Liu et al., 2024b; Ma et al., 2024, 2025) introduces fine-grained, instance-level annotations in dynamic musical performances, enabling richer question types such as comparative, temporal, and counting, and providing a comprehensive testbed for spatiotemporal reasoning.

Despite these advances, several challenges remain. First, many approaches adopt a two-stage paradigm that combines contrastive learning with masked reconstruction (Gurram et al., 2022; Wang et al., 2024; Liu et al., 2024a; Araujo et al., 2025; Wang et al., 2025b; Cheng et al., 2025); while effective for representation learning, these methods decouple alignment from downstream supervision, require complex pretraining, and hinder joint optimization. In contrast, Optimal Transport (OT) (Santambrogio, 2015) has proven effective for multi-source alignment (Zhu and Luo, 2024; Yan et al., 2024; Rho et al., 2025; Wang and Zhao, 2025; Wang et al., 2025a), with empirical studies confirming its superiority over contrastive paradigms (Chowdhury et al., 2024). Second, long video sequences contain high redundancy, with only a small fraction relevant to the query; while recent methods (Ye et al., 2024a; Li et al., 2024; Diao et al., 2025; Kim et al., 2025) introduce question-aware mechanisms, they often compress raw streams without explicitly modeling question–modality alignment. Third, fusion design remains insufficiently adaptive: early (Shi et al., 2024), mid-level (Yang et al., 2024; Awan et al., 2024; Zhou et al., 2025), and late (Senocak et al., 2023; Li et al., 2024; Ryumina et al., 2024; Bravo et al., 2025) strategies lack flexibility to dynami-

cally balance modalities, while recent Mixture-of-Experts (MoEs) (Han et al., 2024; Lin et al., 2024; Kim et al., 2025) introduce conditional computation and capacity allocation but remain restricted to vanilla architectures.

These limitations highlight the need for AVQA models that jointly optimize fine-grained cross-modal alignment under task supervision, explicitly leverage question semantics to filter redundancy, and incorporate adaptive fusion mechanisms that allocate capacity where needed. Addressing these challenges is essential to advance robust and generalizable audio-visual reasoning.

To address these limitations, we propose *EQUALS*, a unified audio–visual LLM framework that integrates compression, alignment, and fusion within a single stage. Given an audio–visual question, our method progressively refines multimodal representations through a structured align–compress–align pipeline. First, one-sided optimal transport module (AT/VT) softly aligns question tokens with fine-grained audio and visual patches, amplifying question-relevant cues while preserving modality-specific content. Next, a question-guided pooling module condenses these OT-enhanced features into compact representations that retain only evidence useful for answering the query. Finally, bidirectional optimal transport module(AV) aligns the compressed audio and video streams to ensure they refer to the same underlying event, yielding a coherent, question-consistent multimodal representation. A task-adaptive Mixture-of-Experts module then dynamically routes among complementary fusion strategies, enabling flexible and robust reasoning across diverse query types. This interleaved design—OT before pooling, OT after pooling, and MoE-based fusion—provides a principled mechanism for injecting question supervision at every stage of representation learning, before passing the resulting representation to a large language model for final reasoning.

Our contributions are threefold: (1) We present a unified training framework that *interleaves optimal-transport alignment before and after pooling*, enforcing semantic consistency both within and across modalities and producing compact, coherent multimodal features. (2) We introduce a *Question-Guided Pooling* mechanism that selectively compresses relevant audio and visual regions, ensuring focused and linguistically grounded multimodal representations. (3) We develop *FlexFuseMoEs*, a task-adaptive mixture-of-experts architecture that dynamically integrates early, mid, and late fusion strategies through diverse expert routings, which improves audio-visual reasoning performance.

2 Related Works

2.1 Audio-Visual Understanding

Recent works in audio–visual understanding largely focus on spatio-temporal alignment and semantic fusion. CAD (Nadeem et al., 2024) performs contrastive learning across spatial, temporal, and semantic dimensions, while SHMamba (Yang et al., 2024b) introduces hyperbolic-state Mamba for cross-modal exchange. VALOR (Liu et al., 2024a) strengthens multi-modal associations via grouping and captioning, and the InternVideo series (Wang et al., 2022; 2024; 2025b) improves long-video modeling through progressive training. Question-guided approaches further refine feature selection: PSTP-Net (Li et al., 2023) localizes relevant regions, MCD (Ye et al., 2024a) aligns modalities via distillation, TWM (Diao et al., 2025) applies query-guided attention to temporally salient content, and QA-TIGER (Kim et al., 2025) models temporal dynamics by selecting informative frames with Gaussian-based mechanisms.

2.2 Mixture-of-Experts in Multimodal Learning

MoE-based architectures have shown promise in multimodal fusion by dynamically routing inputs to expert modules, improving efficiency and scalability. MoPE (Jiang et al., 2024) leverages prompt-based expert selection, while FuseMoE (Han et al., 2024) introduces Laplace gating to address missing or irregular modalities. QA-TIGER (Kim et al., 2025) incorporates MoE to enhance temporal reasoning via Gaussian experts.

Recent designs further expand MoE's scope. MoME (Shen et al., 2024) uses sparsely gated visual-language experts to reduce interference, and TTME (Sun et al., 2024b) applies expert routing in tri-modal pretraining. MIMoE-FND (Liu et al., 2025) adopts a hierarchical MoE with interaction gating for fine-grained modeling in fake news detection. These works demonstrate MoE's effectiveness in handling heterogeneous modality interactions.

3 Method

We propose a unified framework for question-guided multimodal representation learning that

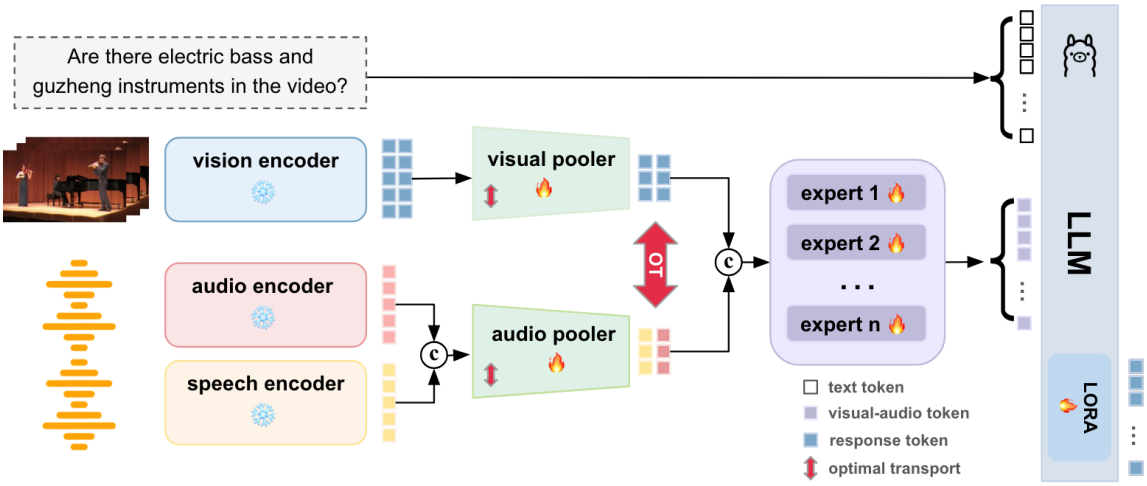

Figure 1: Overview of the proposed framework, which consists of three key components: (1) Question-guided Pooling for extracting question-aware modality features; (2) OTLoss to enforce cross-modal alignment via optimal transport before and after pooling operation; and (3) FlexFuseMoE, a task-dependent mixture-of-experts for adaptive multimodal fusion.

jointly addresses alignment, compression, and fusion within a single-stage architecture. Unlike prior two-stage approaches that decouple pretraining from downstream supervision, our design leverages the question as a guiding signal to adaptively extract task-relevant information from both audio and video modalities, thereby enabling efficient and question-aware multimodal fusion. An overview of the proposed architecture is illustrated in Figure 1. The framework is built upon three key components. First, *Question-Guided Pooling* generates modality representations that are explicitly conditioned on the question. Second, an interleaved *Optimal Transport Loss (OTLoss)* module enforces fine-grained cross-modal alignment, serving as a plug-in objective directly coupled with task supervision. Third, *FlexFuseMoEs*, an advanced flexible mixture-of-experts mechanism, perform dynamic and adaptive fusion of multimodal features by allocating capacity according to question-specific demands.

3.1 Task Defination

Given a multimodal input $X = \{x^{\mathrm{a}}, x^{\mathrm{v}}\}$ and a natural language question Q, our goal is to learn a question-aware fused representation that can be used for downstream tasks such as audio-visual question answering. We dynamically condition the compressive pooling process on the semantic content of Q, reducing redundant information and mitigating interference during the fusion stage.

3.2 Question-Guided Pooling

In this section, we introduce our Question-Guided Pooling design, which leverage task-specific descriptive questions as guidance to extract information from audio and video modalities using 2D and 3D attention mechanisms, respectively. Figure 2 describe the details of the module process.

3.2.1 Question Global Semantic

We begin by encoding the natural language question Q using a pre-trained BERT model (Radford et al., 2021). The [CLS] token is adopted as the global semantic representation $h_Q^{\text{bert[CLS]}} \in \mathbb{R}^d$, where d denotes the embedding dimensionality. We use BERT rather than the embeddings from the first layer of LLM, since the latter capture only local token-level features and fail to provide a coherent global representation of the entire question.

3.2.2 Question-Guided Pooling

Given audio inputs x^a and video inputs x^v, we first extract modality-specific representations using pretrained encoders:

$$\begin{aligned} h^a &= \text{Enc}_a(x^a) \in \mathbb{R}^{T_a \times B \times D_a}, \\ h^v &= \text{Enc}_v(x^v) \in \mathbb{R}^{T_v \times H \times W \times D_v}, \end{aligned} \tag{1}$$

where $T_a(T_v), F_a(F_v), B(H, W)$ and D represent the temporal, spatial, and feature dimensions of audio and video, respectively. We then flatten the frequency(spatial) and temporal dimensions into sequences, applying a $\phi(\cdot)$ operation (two-layer MLP with ReLU activation and output dimension

of LLaMA3) to obtain the features prior modality-specific pooling:

$$\begin{aligned}
&\tilde{h}^a = \phi(\text{Flatten}(h_a)),\\
&\tilde{h}^a \in \mathbb{R}^{T'_a \times D_{\text{llama}}},\ T'_a = T_a \times B,\\
&\tilde{h}^v = \phi(\text{Flatten}(h_v)),\\
&\tilde{h}^v \in \mathbb{R}^{T'_v \times D_{\text{llama}}},\ T'_v = T_v \times H \times W.
\end{aligned} \tag{2}$$

In order to effectively guide the pooling process with global question semantics, we project the question embedding ($h^{\text{BERT}}_{[\text{CLS}]}$) into the hidden spaces of both the audio and visual modalities via a learnable linear transformation , resulting in the projected embedding h_{q}. The attention weights over the modality features are computed using a softmax-based attention mechanism, as follows:

$$\begin{aligned}
&h_q = \text{Linear}(h^{\text{BERT}}_{[\text{CLS}]}),\\
&\alpha^a_t = \text{softmax}\left(\frac{h_q^\top W_a\, \tilde{h}^a_t}{\sqrt{d}}\right), \quad t = 1, \ldots, T'_a,\\
&\alpha^v_t = \text{softmax}\left(\frac{h_q^\top W_v\, \tilde{h}^v_t}{\sqrt{d}}\right), \quad t = 1, \ldots, T'_v,
\end{aligned} \tag{3}$$

where $\alpha_{a,t}$ and $\alpha_{v,t}$ represent the attention weight for the t-th feature, $h_{a,t}$ and $h_{v,t}$ are the reshaped features piror pooling operation, and W_a and W_v are the learned weight matrix for respective modality stream. This attention mechanism ensures that features most relevant to the question semantics are prioritized during the pooling operation. Finally, the compressed and informative representations h^a_{pooled} and h^v_{pooled} are obtained by taking an weighted average pooling on features $\hat{h}^a_t$ and $\hat{h}^v_t$ that are enhanced by the text-to-audio/video alignment (see Sec. 3.3.1) on $\tilde{h}^a_t$ and $\tilde{h}^v_t$ respectively:

$$\begin{aligned}
&h^a_{\text{pooled}} = \text{AttnAvgPool}\left(\left\{\alpha^a_t\, \tilde{h}^a_t\right\}_{t=1}^{T'_a}\right),\\
&h^v_{\text{pooled}} = \text{AttnAvgPool}\left(\left\{\alpha^v_t\, \tilde{h}^v_t\right\}_{t=1}^{T'_v}\right).
\end{aligned} \tag{4}$$

3.3 Optimal Transport based Alignment Cross Modalities

We design a family of Optimal Transport Loss (OTLoss) modules that enable flexible alignment across arbitrary modality pairs. Concretely, we develop three task-specific variants. *OTLoss-AT* (audio–text) and *OTLoss-VT* (video–text) transfer question semantics into the local "patch" embeddings of audio or video, thereby enhancing question-aware representation learning. *OTLoss-AV* (audio–video) focuses on aligning complementary audio–visual information after ruquestion-guided compression, preparing more informative representations for the subsequent fusion stage. To support diverse learning dynamics, these modules are interleaved at different stages of fine-tuning. Furthermore, to prevent spurious alignments caused by variable-length inputs, we modify the transport cost by imposing a large penalty on distances involving padding tokens, thereby ensuring that the transport plan is concentrated on semantically meaningful regions.

3.3.1 Text-to-Audio/Video Alignment Before Question-Guided Pooling

As shown in Figure 2,to align textual information with acoustic and visual streams, we adopt an Optimal Transport (OT) formulation that treats question tokens as queries and audio/video tokens as candidates for alignment. Specifically, We use LLaMA3(Grattafiori et al., 2024) to encode the input question as $h^l \in \mathbb{R}^{T_l \times D_{llama}}$ due to its fine-grained token-level representations and strong semantic modeling capability. To match query embedding with the dimensionality of audio and video encoders, a linear layer is applied to project the text embedding into audio and video's hidden dimensions respectively.

As mentioned in Section 3.2.2, Given the flattened audio representation $\tilde{h}^a \in \mathbb{R}^{T'_a \times D_{llama}}$ and $h_l \in \mathbb{R}^{T_l \times D_{llama}}$, we first construct the cost matrix via cosine distance:

$$C^{l2a}_{ij} = 1 - \cos(h^i_l, h^j_a), \quad C^{l2v}_{ij} = 1 - \cos(h^i_l, h^j_v), \tag{5}$$

where i indexes text tokens and j indexes audio or visual tokens.

To obtain a soft assignment, we normalize each row to form a distribution over the target modality:

$$\begin{aligned}
M^{l2a}_{i\cdot} &= \tfrac{1}{T_l}\,\text{softmax}\left(-\tfrac{1}{\tau_a} C^{l2a}_{i\cdot}\right),\\
M^{l2v}_{i\cdot} &= \tfrac{1}{T_l}\,\text{softmax}\left(-\tfrac{1}{\tau_v} C^{l2v}_{i\cdot}\right),
\end{aligned} \tag{6}$$

where τ_a, τ_v are temperature parameters controlling the sharpness of the alignment (smaller τ leads to harder assignments). Alternatively, sparse operators such as sparsemax or entmax can be used, and a straight-through estimator (STE) can approximate hard $\arg\min$ assignments.

This one-sided OT losses are formulated as:

$$\mathcal{L}_{\text{OT}}^{l\to a} = \langle M^{l2a}, C^{l2a}\rangle + \varepsilon_a H(M^{l2a}) + \alpha_a \left\| M^{l2a}\mathbf{1} - \tfrac{1}{T_l}\mathbf{1}\right\|_2^2, \quad (7)$$

$$\mathcal{L}_{\text{OT}}^{l\to v} = \langle M^{l2v}, C^{l2v}\rangle + \varepsilon_v H(M^{l2v}) + \alpha_v \left\| M^{l2v}\mathbf{1} - \tfrac{1}{T_l}\mathbf{1}\right\|_2^2, \quad (8)$$

where $\langle A, B\rangle = \sum_{ij} A_{ij}B_{ij}$, $H(M) = -\sum_{ij} M_{ij}\log M_{ij}$ is the entropy regularizer, and the last term softly enforces the row-sum constraint. This formulation allows each text token to distribute its mass to semantically similar audio or visual tokens, producing a differentiable alignment map without extra parameters. This one-sided OT alignment offers a lightweight yet effective way to guide cross-modal interaction, yielding question-augmented audio and video representations:

$$\hat{h}_a = \tilde{h}^a + M^{l2a}h_l, \quad \hat{h}_v = \tilde{h}^v + M^{l2v}h_l. \quad (9)$$

3.4 Alignment Between Audio and Video Modalities After Question-Guided Pooling

For audio–visual alignment, we focus on compressed patch-level features obtained after question-guided pooling, with the goal of aligning semantically corresponding information across modalities to prepare for the subsequent fusion stage. Given audio and video modalities $\mathcal{M}_{audio}$ and $\mathcal{M}_{video}$, h^a_{pooled} and h^v_{pooled} are their respective patch-level representations. Leveraging the semantic structure encoded in these features, we construct two discrete distributions $\theta_a \in P(\mathcal{H}_a)$ and $\theta_v \in P(\mathcal{H}_v)$ for the audio and video streams as follows:

$$\theta_a = \sum_{k=1}^{M} u_a(k)\,\delta_{f_a(k)}, \quad \theta_v = \sum_{l=1}^{N} u_v(l)\,\delta_{f_v(l)} \quad (10)$$

where $u_a \in \Delta^M$ and $u_v \in \Delta^N$ denote the weights of corresponding probability distributions, which we set as uniform for simplicity, i.e., $u_a(k) = \frac{1}{M}$ and $u_v(l) = \frac{1}{N}$. The term δ_f denotes the Dirac delta function placed at the supporting point f in the embedding space where the distribution assigns non-zero probability (Chen et al., 2022a).

The objective is to compute the optimal plan Ω that minimizes the transport cost between the two distributions θ_a and θ_v, while preserving the underlying topological structure during cross-modal alignment. This is achieved by minimizing the Wasserstein Distance (WD), formally defined as:

$$\mathcal{L}_{\text{OT}}^{a\leftrightarrow v} = \mathcal{L}_{\text{OT}}(\theta_a, \theta_v) = \min_{\Omega\in\Psi(u_a,u_v)} \sum_{k,l} \Omega_{kl}\,\phi(h^a_k, h^v_l). \quad (11)$$

The set of feasible transport plans $\Psi(u_a, u_v)$ is defined as follows:

$$\Psi(u_a, u_v) = \Big\{\Omega \in \mathbb{R}_+^{M\times N} \,\Big|\, \Omega\mathbf{1}_N = u_a,\ \Omega^\top\mathbf{1}_M = u_v\Big\}, \quad (12)$$

and the ground cost function $\phi(f_a(k), f_v(l))$ is instantiated as the cosine distance between local patch-level embeddings from two modalities. An exact solution to the above formulation yields a sparse optimal transport plan Ω, containing at most $2\cdot\max(M,N) - 1$ non-zero entries. This sparsity property enhances both the interpretability and robustness of the resulting cross-modal alignment. Algorithmic detail on calculating WD is appended in Appendix A.

3.5 FlexFuseMoEs

We propose *FlexFuseMoEs*, a multimodal fusion module that dynamically routes features through a mixture-of-experts architecture. Building on DeepSeekMoE (Dai et al., 2024), our model employs N groups of m fine-grained experts per layer (K_s shared, $mN - K_s$ routed), which is different from vanilla MoE adopted for multimodal fusion in (Han et al., 2024). Each expert is implemented as a feed-forward network:

$$E_k(x) = W_{2,k}\cdot\text{ReLU}(W_{1,k}x), \quad \text{for } k = 1,\ldots,N-K_s. \quad (13)$$

For each input token representation f_t^l at layer l, the output combines contributions from always-active shared experts and selectively-activated routed experts:

$$h_t^l = \underbrace{\sum_{i=1}^{K_s}\text{FFN}_i(f_t^l)}_{\text{shared}} + \underbrace{\sum_{i=K_s+1}^{mN} g_{i,t}\,\text{FFN}_i(f_t^l)}_{\text{routed}} + f_t^l$$

$$g_{i,t} = \begin{cases} s_{i,t}, & \text{if } s_{i,t}\in\text{TopK}\left(\{s_{j,t}\}_{j=K_s+1}^{mN}, K_r = mK - K_s\right) \\ 0, & \text{otherwise}\end{cases}$$

$$s_{i,t} = \text{softmax}_i\big((f_t^l)^\top e_i^l\big)$$

$$e_i^l \sim \mathcal{N}(0,\sigma^2 I), \quad \sigma = 0.001. \quad (14)$$

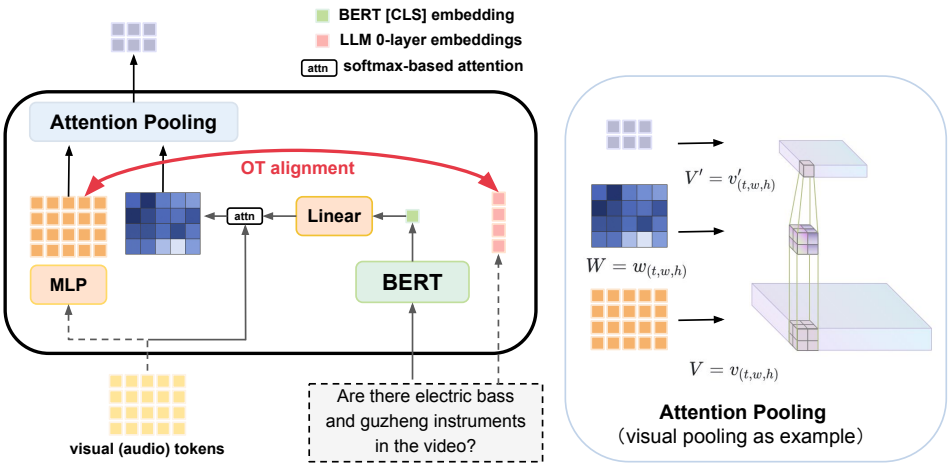

Figure 2: Question-Guided Pooling and Text-Audio/Video Alignment: Optimal transport-based alignment is first applied from text to audio or video modality using LLaMA-encoded question embeddings projected via an MLP block. Subsequently, for each modality, attention weights are computed through a softmax-based attention mechanism between the BERT [CLS] token and audio/video features to guide pooling. This ensures that semantically relevant regions are extracted based on the input question. Solid gray arrows indicate question-guided attention; dotted gray arrows represent optimal transport alignment; black arrows denote final attention pooling using the weight matrix (blue blocks) and aligned representation (orange blocks).

The gating mechanism activates only the top-K_r routed experts (default $K_r = 4$) based on their affinity scores $s_{i,t}$, while $K_s = 2$ shared experts process all tokens. This sparse activation maintains model capacity while ensuring computational efficiency. Learnable routing centroids e_i^l enable automatic specialization of experts to different input patterns. Inspired by FuseMoE (Han et al., 2024), we adopt three routing strategies to support different levels of modality specialization and parameter sharing:

- **Disjoint Routing** (separate gates, separate experts). Each modality $m \in \{\text{audio}, \text{video}\}$ uses its own gating network g_m and expert pool E_m, enabling maximal routing flexibility and modality-specific specialization.
- **Joint Routing** (shared gate, shared experts). A single gating network g dispatches all modalities to a common expert set E, encouraging fully shared multimodal representations.
- **Per-Modality Routing** (shared gate, modality-specific experts). A shared gating network g determines routing weights, while each modality retains its own expert pool E_m, balancing shared decision making with modality-aware computation.

Given modality-specific representations $f^a, f^v \in \mathbb{R}^{D_{llama}}$ obtained from question-guided pooling and alignment respectively, the fusion operation produces an output that preserves the original feature dimensionality, making it directly compatible with the downstream LLM architecture.

3.6 Total Loss

Our final training objective combines the cross-entropy loss for supervised tasks with the OT-based alignment losses:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} &+ \lambda_{\text{at}} \mathcal{L}_{\text{OT-1s}}^{t\to v}(h_l, \tilde{h}^a) \\ &+ \lambda_{\text{vt}} \mathcal{L}_{\text{OT-1s}}^{t\to v}(h_l, \tilde{h}^v) \\ &+ \lambda_{\text{av}} \mathcal{L}_{\text{OT-1s}}^{a\leftrightarrow v}(h^a_{\text{pooled}}, h^v_{\text{pooled}}),\end{aligned} \tag{15}$$

where $\lambda_{\text{at}}, \lambda_{\text{vt}}, \lambda_{\text{av}}$ are hyperparameters controlling the contribution of each OT loss component.

4 Experiments

We evaluate EQUALS on the widely used audio-visual benchmark MUSIC-AVQA and its more robust extension FortisAVQA. This dataset suite illus-

trates how audio-visual reasoning can be systematically studied under carefully designed conditions, while FortisAVQA further improves the testbed toward reducing biases and enabling more reliable evaluation. Additional details on datasets evolving and our preprocessing are provided in Appendix B.1 and B.2 respectively.

4.1 Experimental Settings

We implement our framework in PyTorch and train with DeepSpeed ZeRO-1 using 4×A100 GPUs under mixed-precision. For text encoding, we follow prior multimodal LLM practice: CLIP-ViT-L/14 provides the global [CLS] representation for pooling, while token-level question features for cross-modal alignment are extracted from Llama3.

Video inputs are processed at 2Hz using InternVideo2, producing 256 spatial patches per clip. Audio is represented through a dual-branch pipeline combining BEATs and Whisper-v3 at 16 kHz, enabling both high-quality timbre modeling and robust speech perception.

The backbone LLM adopts DeepSeek-MoE with 16 routed experts and top-2 routing. We fine-tune the model with LoRA-SFT, training uses Adam with cosine decay (initial LR 3×10), batch size 2, and a uniform optimal-transport alignment weighting of = 0.3 across all modality pairs. More architectural and optimization details are provided in Appendix C.

4.2 Comparison on MUSIC-AVQA

We evaluate our method on the MUSIC-AVQA test split to validate its effectiveness. For comparison, we select and add representative baselines spanning small- and large-scale state-of-the-art methods (Ma et al., 2025; Li et al., 2024). These include small models such as AVSD (Alamri et al., 2019), LAViT (Yun et al., 2021), PSTP-Net (Li et al., 2023), LAVisH (Lin et al., 2023) and QA-TIGER (Kim et al., 2025). We further compare against recent large multimodal models, including Qwen2.5-VL (Bai et al., 2023), CAT (Ye et al., 2024b), MeerKat (Chowdhury et al., 2024), video-SALMONN (Sun et al., 2024a), VideoLLaMA2 (Cheng et al., 2024), VITA (Fu et al., 2024), GPT-4o (Hurst et al., 2024), Qwen2.5-Omni (Xu et al., 2025), MAVEN (Ma et al., 2025) and Kimi-Audio (Ding et al., 2025). Following common practice, Qwen2.5-VL, video-SALMONN, VITA, GPT-4o, Qwen2.5-Omni and Kimi-Audio are evaluated in a zero-shot setting due to their closed-source nature or lack of complete training components, while CAT, MeerKat, VideoLLaMA2 and MAVEN are fine-tuned on the benchmark with same LoRA configs for fair comparison. See additional details of baseline configurations in Appendix E.

4.3 Robustness Evaluation

As shown in Table 2, our method achieves state-of-the-art results in both head and tail splits for various question types within audio, visual and visual-audio modalities.

The disjoint fusion design improves prior bests by +0.92%(88.15% vs. 87.23% from MAVEN) on location (head) within visual modality; it also increase the best baselines sofar by +4.98%(90.60% vs. 87.97% from VideoLLaMA2) , +6.12%(82.82% vs. 76.70% from QA-TIGER), +5.98%(83.76% vs. 77.78% from VITA), and +0.83%(81.30% vs. 80.47% from QA-TIGER) on exist (head), count(head) and comparison (head and tail) question types within visual-audio modality respectively. The joint fusion variant further advances performance with +11.28%(86.87% vs. 75.59% from QA-TIGER) on location(tail) within visual modality, and +9.25%/+19.47%(80.68% vs. 71.43% from MAEN and 86.30% vs.66.83% from CAT) on location (head/tail) within audio-visual modality, as well as +3.89%(89.60% vs. 85.71% from MAVEN) on temporal (tail) within audio-visual modality. Per-modality fusion also yields large gains, including +24.24%(91.67% vs. 67.43% from MeerKat) on comparison in audio, +12.29%(85.26% vs. 72.97% from Qwen2.5-VL) on count (tail) in visual, and +17.73%(79.59% vs. 65.86% from CAT) on temporal (head) in audio-visual. These consistent improvements demonstrate the effectiveness and cross-benchmark robustness of our design for various multimodal reasoning tasks.

While models such as LAViT and LAVisH perform competitively on MUSIC-AVQA, they degrade sharply on FortisAVQA, especially in the tail set (e.g., LAViT reaches only 22.22% on EXIST-type). This suggests reliance on dataset-specific memorization rather than genuine multimodal reasoning. In contrast, our method and VideoLLaMA-2 exhibit much stronger robustness, with only 4.21% and 5.71% drops, respectively. Notably, large models in the zero-shot setting even match or surpass supervised results on MUSIC-AVQA, underscoring the strong generalization ability of

Table 1: Performance comparison on AVQA tasks of the MUSIC-AVQA dataset.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
AVSD((Alamri et al., 2019))	72.47	62.46	68.78	66.00	74.53	70.31	80.77	64.03	57.93	62.85	61.07	65.44	67.32
LAViT((Yun et al., 2021))	74.27	65.06	70.86	69.89	77.12	73.55	81.21	62.03	65.93	60.90	63.96	66.78	69.29
PSTP-Net((Li et al., 2023))	73.97	65.79	70.91	77.15	77.36	77.26	76.18	72.23	71.80	71.79	69.00	72.57	73.52
LAVisH((Lin et al., 2023))	78.18	58.74	70.98	75.65	78.75	77.21	81.41	63.54	71.98	57.76	66.38	68.30	71.13
QA-TIGER((Kim et al., 2025))	84.86	67.85	78.58	83.96	86.29	85.14	83.10	78.58	72.50	63.94	69.59	73.74	77.62
Qwen2.5-VL((Bai et al., 2023))	48.60	55.00	51.80	55.28	53.66	54.47	44.00	52.17	63.57	37.84	41.18	47.75	50.14
CAT((Ye et al., 2024b))	73.42	70.15	69.87	64.87	75.26	67.48	72.54	68.01	70.92	58.84	68.20	65.47	68.83
MeerKat((Chowdhury et al., 2024))	75.90	60.92	64.10	57.28	72.64	69.87	63.62	62.45	71.98	56.12	74.69	62.49	69.21
video-SALMONN((Sun et al., 2024a))	59.59	24.48	55.82	47.54	44.63	46.07	74.57	38.46	29.45	24.48	16.38	37.25	42.86
VideoLLaMA2((Cheng et al., 2024))	79.44	52.46	69.64	71.30	52.93	62.11	77.00	63.44	77.69	59.46	64.71	68.98	72.56
VITA((Fu et al., 2024))	59.81	45.90	54.76	50.41	34.96	42.68	54.00	49.46	46.92	27.93	41.18	43.74	45.44
GPT-4o ((Hurst et al., 2024))	65.42	36.07	50.75	72.36	62.30	67.33	56.12	54.84	59.23	37.84	42.35	50.08	54.06
Qwen2.5-Omni((Xu et al., 2025))	74.58	34.78	59.85	80.15	57.59	68.89	67.74	66.01	43.58	25.92	38.35	49.16	56.27
MAVEN((Ma et al., 2025))	79.44	54.10	72.79	80.49	**93.50**	86.99	87.00	66.67	73.85	54.95	68.24	69.94	74.60
Kimi-Audio((Ding et al., 2025))	72.31	60.73	68.02	58.80	37.35	47.82	62.91	40.82	32.90	52.20	21.00	42.96	46.03
Ours(disjoint)	86.51	80.28	84.21	83.60	91.16	88.40	90.81	82.71	86.15	80.05	86.10	85.00	85.93
Ours(joint)	**88.46**	**82.95**	**86.43**	83.39	91.09	88.28	91.11	82.00	**86.47**	**82.95**	**86.37**	**85.25**	**86.36**
Ours(per-modality)	87.63	80.87	85.13	**84.34**	90.87	**88.49**	**91.47**	**82.32**	85.99	80.94	85.92	85.14	86.17

Table 2: Performance comparison on FortisAVQA benchmark. (H: Head; T: Tail).

Method	Audio QA				Visual QA				AVQA										Avg
	Count		Comp		Count		Local		Exist		Count		Local		Comp		Temp		
	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T	
LAViT	50.53	44.19	54.66	58.33	50.49	25.67	65.07	59.02	54.14	22.22	47.33	40.30	46.26	21.95	37.61	48.25	39.83	48.33	47.47
LAVisH	71.90	61.27	56.47	58.79	71.84	27.87	43.68	63.13	70.99	41.53	35.62	57.37	66.75	21.91	43.92	69.38	28.76	44.91	54.88
QA-TIGER	82.67	75.82	71.75	43.11	81.30	54.59	84.76	75.59	72.84	78.56	76.70	33.55	48.22	64.65	37.55	80.47	36.85	62.96	67.99
Qwen2.5-VL	88.11	57.14	83.78	13.33	77.94	72.97	46.38	68.91	57.14	76.07	36.24	53.73	56.76	60.98	11.97	76.99	25.42	63.33	57.07
CAT	55.10	57.42	57.82	63.17	59.24	56.88	63.43	68.23	67.17	73.48	58.36	50.54	62.60	66.83	61.49	42.97	65.86	60.33	69.97
MeerKat	64.21	53.95	78.67	67.43	69.90	60.81	60.10	72.47	76.09	44.02	55.84	46.67	58.67	54.93	61.71	54.74	55.42	63.59	61.35
video-SALMONN	75.26	30.23	84.00	20.00	54.85	22.97	46.89	23.77	62.41	75.21	34.36	31.71	24.00	13.43	17.09	34.21	16.10	26.67	40.84
VideoLLaMA2	89.47	67.44	64.00	40.00	84.95	67.57	55.02	74.59	87.97	44.44	61.33	52.24	70.48	51.22	59.83	62.28	49.15	68.33	66.85
VITA	**91.45**	48.28	62.67	35.00	74.03	38.89	15.74	46.19	63.16	45.30	34.34	37.29	47.22	27.27	77.78	28.95	27.27	49.50	48.66
GPT-4o	63.78	46.51	69.33	16.67	75.00	68.92	56.52	75.21	44.36	87.18	42.95	50.75	56.83	39.02	13.68	63.16	30.51	71.67	56.06
Qwen2.5-Omni	59.47	**86.05**	71.95	20.00	80.10	63.51	46.41	66.39	45.11	82.05	70.93	**73.17**	32.67	50.75	12.68	56.14	26.27	73.33	54.61
MAVEN	88.64	62.50	**85.71**	33.33	**92.31**	66.67	87.23	75.76	85.71	**89.29**	65.62	55.56	71.43	44.44	45.83	61.90	54.55	85.71	72.92
Kimi-Audio	87.89	25.58	86.67	31.67	66.50	37.84	18.66	13.11	72.18	82.91	50.56	17.07	36.67	10.45	36.75	73.68	27.12	16.67	48.14
Ours(disjoint)	84.28	79.35	68.00	89.16	86.30	67.76	**88.15**	82.34	**90.60**	82.48	**82.82**	51.20	76.99	83.70	**83.76**	**81.30**	79.24	86.14	81.47
Ours(joint)	80.16	79.35	77.33	87.50	86.06	64.47	86.41	**86.87**	90.23	88.46	78.86	50.00	**80.68**	**86.30**	79.92	78.28	77.86	**89.60**	**82.15**
Ours(per-mod.)	81.96	79.35	78.67	**91.67**	85.58	**85.26**	87.05	82.34	87.97	60.68	74.01	50.00	78.53	82.22	76.50	60.00	**79.59**	84.16	79.30

large-scale multimodal LLMs.

4.4 Analysis Studies

4.4.1 OTLoss Ablation Analysis

We conducted ablation study with our best average-performance models to verify the effectiveness of the OTLoss module. The full OTLoss setting $(\lambda_{at} + \lambda_{vt} + \lambda_{av})$ achieves the best performance, reaching 86.17% on AVQA and 82.15% on FortisAVQA. Removing any OT alignment term consistently degrades performance, while the baseline without OT losses performs worst $(83.18\%/79.90\%)$. Notably, the audio-text (λ_{at}) and video-text (λ_{vt}) alignments contribute the most significant gains, highlighting the importance of cross-modal optimal transport alignment.

4.4.2 Further Analysis

We provide three complementary analyses in the appendix for studying different pooling methods, kernel-stride configuesations and highlighting the importance of audio cues in AVQA tasks(see Appendix G, H and Appendix I).

Table 3: Performance (%) across configurations with different OT loss components. Red numbers denote relative improvements over the *None* baseline.

Configuration	AVQA	Fortis	Fortis^H	Fortis^T
$\lambda_{at} + \lambda_{vt} + \lambda_{av}$	**86.17** (+2.99)	**82.15** (+2.25)	**82.76** (+4.07)	**82.25** (+6.99)
$\lambda_{at} + \lambda_{vt}$	84.34 (+1.16)	81.05 (+1.15)	81.56 (+2.87)	80.09 (+4.83)
λ_{av}	83.80 (+0.62)	80.78 (+0.88)	80.48 (+1.79)	78.14 (+2.88)
None	83.18	79.90	78.69	75.26

5 Conclusion

We introduced EQUALS, a unified one-stage framework for audio-visual question answering that integrates question-guided alignment, feature compression, and flexible fusion. Experiments on MUSIC-AVQA and FortisAVQA show that EQUALS achieves state-of-the-art results across overall accuracy and fine-grained question types, validating its effectiveness and robustness.

6 Limitation

Although current evaluation focuses on music-performance settings, future work will extend to more diverse audio-visual data to assess the generalizability of the framework.

References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, and 1 others. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, and James R Glass. 2025. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. *arXiv preprint arXiv:2505.01237*.

Mahrukh Awan, Asmar Nadeem, Muhammad Junaid Awan, Armin Mustafa, and Syed Sameed Husain. 2024. Attend-fusion: Efficient audio-visual fusion for video classification. *arXiv preprint arXiv:2408.14441*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Luis Bravo, Ciro Rodriguez, Pedro Hidalgo, and Cesar Angulo. 2025. A systematic review on artificial intelligence-based multimodal dialogue systems capable of emotion recognition. *Multimodal Technologies and Interaction*, 9(3):28.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2022a. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022b. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Hao Cheng, Zhiwei Zhao, Yichao He, Zhenzhen Hu, Jia Li, Meng Wang, and Richang Hong. 2025. Vaemo: Efficient representation learning for visual-audio emotion with knowledge injection. *arXiv preprint arXiv:2505.02331*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

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

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, and 1 others. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding. *arXiv preprint arXiv:2502.06020*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, and 1 others. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sumanth Gurram, Andy Fang, David Chan, and John Canny. 2022. Lava: language audio vision alignment for contrastive video pre-training. *arXiv preprint arXiv:2207.08024*.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. 2024. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. 2024. Mope: Mixture of prompt experts for parameter-efficient and scalable multimodal fusion. *arXiv preprint arXiv:2403.10568*.

Hongyeob Kim, Inyoung Jung, Dayoon Suh, Youjia Zhang, Sangmin Lee, and Sungeun Hong. 2025. Question-aware gaussian experts for audio-visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13681–13690.

Guangyao Li, Henghui Du, and Di Hu. 2024. Boosting audio visual question answering via key semantic-aware cues. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5997–6005.

Guangyao Li, Wenxuan Hou, and Di Hu. 2023. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM international conference on multimedia*, pages 7808–7816.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19086–19096. IEEE.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. 2024. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*.

Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2023. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2024a. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiulong Liu, Zhikang Dong, and Peng Zhang. 2024b. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4478–4487.

Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection. *arXiv preprint arXiv:2501.12431*.

Jie Ma, Zhitao Gao, Qi Chai, Jun Liu, Pinghui Wang, Jing Tao, and Zhou Su. 2025. Fortisavqa and maven: a benchmark dataset and debiasing framework for robust multimodal reasoning. *arXiv preprint arXiv:2504.00487*.

Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. 2024. Look, listen, and answer: Overcoming biases for audio-visual question answering. *arXiv preprint arXiv:2404.12020*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Kyeongha Rho, Hyeongkeun Lee, Valentio Iverson, and Joon Son Chung. 2025. Lavcap: Llm-based audio-visual captioning using optimal transport. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, Heysem Kaya, and Alexey Karpov. 2024. Audio-visual compound expression recognition method based on late modality fusion and rule-based decision. *arXiv preprint arXiv:2403.12687*.

Filippo Santambrogio. 2015. Optimal transport for applied mathematicians.

Arda Senocak, Junsik Kim, Tae-Hyun Oh, Dingzeyu Li, and In So Kweon. 2023. Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2237–2247.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. 2024. Mome: Mixture of multimodal experts for generalist multimodal large language models. *arXiv preprint arXiv:2407.12709*.

Y. Shi, H. R. Ong, S. Yang, and Y. Fan. 2024. Deep multimodal fusion of visual and auditory features for robust material recognition. *International Journal of Computers, Communications Control*, 19(5).

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

Wenqi Sun, Jie Zhou, Yuxuan Li, Man Cheung, and James She. 2024b. Tri-modal transformers with mixture-of-modality-experts for social media prediction. *IEEE Transactions on Circuits and Systems for Video Technology*.

Cunda Wang, Weihua Wang, Xinyu Li, Qiuyu Liang, and Feilong Bao. 2025a. Otmea: Multi-modal entity alignment via optimal transport. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024. Internvideo2:

Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, and 1 others. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, and 1 others. 2025b. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Yifeng Wang and Yi Zhao. 2025. Optimal transport with mamba for multimodal inertial signal enhancement. *IEEE Transactions on Industrial Informatics*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Cairong Yan, Meng Ma, Yanting Zhang, and Yongquan Wan. 2024. Dual-path multimodal optimal transport for composed image retrieval. In *Proceedings of the Asian conference on computer vision*, pages 1741–1755.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491.

Tianyu Yang, Yiyang Nan, Lisen Dai, Zhenwen Liang, Yapeng Tian, and Xiangliang Zhang. 2024. Sasr-net: Source-aware semantic representation network for enhancing audio-visual question answering. *arXiv preprint arXiv:2411.04933*.

Qilang Ye, Zitong Yu, and Xin Liu. 2024a. Answering diverse questions via text attached with key audio-visual clues. *arXiv preprint arXiv:2403.06679*.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. 2024b. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pages 146–164. Springer.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041.

Yihan Zhou, Xiaokang Yu, Zixi Huang, Feierdun Palati, Zeyu Zhao, Zihan He, Yuan Feng, and Yuxi Luo. 2025. Multi-modal fused-attention network for depression level recognition based on enhanced audio-visual cues. *IEEE Access*.

Sidan Zhu and Dixin Luo. 2024. Enhancing multimodal contrastive learning via optimal transport-based consistent modality alignment. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 157–171. Springer.

A OTLoss Algorithms Math Details

In Algorithm1, we detail the computation of the Wasserstein distance used in our OTLoss. Given two modalities, the procedure initializes a scaled unity matrix and transport plan, constructs a cosine-based cost matrix, and applies an exponential decay to obtain a scaled similarity matrix. The Sinkhorn iteration is then performed by alternately updating the row and column scaling factors to refine the transport plan. Finally, the Wasserstein distance is computed via a Frobenius dot product between the cost matrix and the learned transport plan. This iterative process yields a stable and differentiable alignment map that can be seamlessly integrated into our framework for cross-modal optimization.

Algorithm 1: EQUALS: Wasserstein Distance Computation in OTLoss

Input: Modality 1: $\{\mathcal{M}1_i\}_{i=1}^{k}$; Modality 2: $\{\mathcal{M}2_j\}_{j=1}^{k}$; Total Optimal Transport steps: Ω; Initial scaled unity matrix: $\sigma = \frac{1}{k}\mathbf{1}\mathbf{1}^\top$; Initial Transport Plan: $\Omega^{(1)} = \mathbf{1}\mathbf{1}^\top$; Cosine similarity matrix: $C_{ij} = c(I_i, A_j)$; similarity matrix decay factor: β; Scaled similarity matrix: $\tau_{ij} = e^{-\frac{C_{ij}}{\beta}}$.

Output: Learned Optimal Transport Plan: Ω; Wasserstein Distance: $D_{\text{Wasserstein}}$.

for $t \in \{1, 2, 3, \cdots, \Omega\}$ **do**
  $\mathbf{Q} \leftarrow \tau \odot \Omega^{(t)}$; `// ⊙ is Hadamard product`
  for $l \in \{1, 2, 3, \cdots, L\}$ **do**
    $\delta \leftarrow \frac{1}{k\mathbf{Q}\mathbf{1}}$, $\sigma \leftarrow \frac{1}{k\mathbf{Q}^\top\delta}$;
  $\Omega^{(t+1)} \leftarrow \text{diag}(\delta)\mathbf{Q}\text{diag}(\sigma)$;
$D_{\text{Wasserstein}} \leftarrow \langle \mathbf{C}^\top, \Omega \rangle$; `// ⟨·,·⟩ is the Frobenius dot-product`
return $\Omega, D_{\text{Wasserstein}}$;

B AVQA Datasets

B.1 Datasets Evolution Description

To systematically evaluate the multimodal reasoning capacity of AVQA systems, a series of progressively refined datasets (You et al., 2025) in the context of musical instrument performance have been developed. The MUSIC-AVQA dataset (Li et al., 2022) serves as an early benchmark, introducing a large-scale collection of 9,288 videos (150+ hours) featuring performances across 22 instruments, with 45,867 question-answer (QA) pairs. These pairs are split into training (32,087 QA pairs), validation (4,595), and testing (9,185) sets, covering 33 question templates spanning 9 types (e.g., existential, location, counting, comparative, temporal). This dataset pioneered the evaluation of basic spatio-temporal and audio-visual understanding but revealed biases in certain question categories.

To address these limitations, MUSIC-AVQA v2.0 (Liu et al., 2024) rebalanced the dataset by adding 1,230 new videos (totaling 10,518) and generating 54,000 QA pairs. The splits were adjusted to 36,700 (training), 5,250 (validation), and 10,819 (testing) pairs, while ensuring no dominant answers exceeded 60% for binary questions or 50% for multi-class questions. This mitigation targeted biases in 15 templates, particularly improving representation in existential, counting, and comparative questions.

Further advancing robustness evaluation, MUSIC-AVQA-R (Ma et al., 2024) expanded the original test set by rephrasing each question 25 times via human-machine collaboration, scaling from 9,129 to 211,572 questions. It introduced a vocabulary of 465 words (5× larger than MUSIC-AVQA) and categorized questions into head (common) and tail (rare) samples to assess out-of-distribution performance. The dataset preserved training/validation biases but enabled granular evaluation through three metrics: head accuracy, tail accuracy, and overall accuracy. Most recently, FortisAVQA (Ma et al., 2025) replaced fixed threshold-based splitting with conformal-inspired optimization, dynamically adapting head-tail splits to ensure coverage and compactness across diverse distributions. Together, these datasets provide a rigorous framework for evaluating AVQA systems, from foundational benchmarks (MUSIC-AVQA) to bias-aware (v2.0) and robustness-focused (MUSIC-AVQA-R, FortisAVQA) iterations. Statistics are summarized in **??**, while Figure 4 illustrates real and synthetic samples.

B.2 Data Preprocessing

We applied a two-step preprocessing pipeline to both MUSIC-AVQA and Fortis-AVQA datasets: (1) audio-video separation, and (2) JSON reformatting.

1. Audio-Video Separation. For each video, we used `MoviePy` to extract the audio track and

Table 4: Question Statistics Comparison of MUSIC AVQA Datasets

Dataset	#QA	#Type	Key Features
MUSIC-AVQA	45.8K	33	9 question types
MUSIC-AVQA2	54K	33	Balanced templates
MUSIC-AVQA-R	211K	33	25× rephrased; 465 vocab.

(a) Real performance.

(b) Synthetic video.

Figure 3: Representative samples from the MUSIC-AVQA dataset (Li et al., 2022). (a) Question: Are there violin and cello sound? Answer: no. (b) Question: What is the instrument on the right of the piano? Answer: cello.

save it as a `.wav` file. The video (without audio) was saved as a new `.mp4` file. The process was parallelized with `multiprocessing.Pool`. Corrupted files were detected via exceptions and logged for exclusion.

Listing 1: Audio-Video Separation (Simplified)

```
for video in video_list:
    try:
        clip = VideoFileClip(video)

        clip.audio.write_audiofile(out_audio)

        clip.without_audio().write_videofile(
            out_video
        )

    except Exception as e:
        log_corrupted(video)
```

2. JSON Reformatting. Original JSON entries were converted to a unified EQUALS-compatible format. For each entry, placeholders in `question_content` (e.g., `<LR>`) were replaced using values in `templ_values`.

Listing 2: JSON Preprocessing (Simplified)

```
for entry in data:
    question =
     fill_template(entry["question_content"],

     entry["templ_values"])
    new_entry = {
        "video_path":
     get_video_path(entry["video_id"]),
        "audio_path":
     get_audio_path(entry["video_id"]),
        "image_path": "",
        "text": "",
        "answer": entry["answer"],
        "question": question,
        "task": entry["type"]
    }
```

All transformed samples were saved into new JSON files with a standardized directory layout.

C Training Details

C.1 Text, Video and Audio Encoding

Text inputs are encoded using the CLIP-ViT-Large-Patch14 transformer, with the [CLS] token used as the global semantic representation to calculate pooling attention score, meanwhile also embedded with Llama3 to get the more fine-grained token-level representation for alignment operation. Video frames are sampled uniformly at 2 Hz, with each 10-frame clip processed into 256 spatial patches using a tubelet size of 1, resulting in a 1408-dimensional embedding space. For audio, we segment clips into 10-frame windows, each with 25 frequency bands and downsampled to 16k Hz. A dual-branch architecture using BEATs and Whisper v3 encodes the audio frames, and sinusoidal positional encoding is applied, resulting in 2048-dimensional embeddings.

C.2 Pooling and Alignment

We use a global pooling layer with a softmax activation, a pooling size of 512, and pooling temperature of 0.1 to control smoothness. An OT loss is applied to guide cross-modality alignment, with a loss weight coefficient of 0.3.

C.3 Backbone and MoE Structure

The core model is based on DeepSeek-MoE (deepseek-moe-16b-chat variant), employing a top-2 routing strategy. It includes 16 routed experts and 2 shared experts, with 4 experts selected per token. The hidden size and output size are 4096, the intermediate size is 8192, and the MoE-specific intermediate size is 1024. Expert scoring uses a

softmax function with normalized top-k probability. An auxiliary loss is applied with a coefficient of 0.001, and sequence-level auxiliary supervision is enabled. The activation function is SiLU.

C.4 Optimization and Training

Training is performed on 4×NVIDIA A100 GPUs using mixed-precision (fp16). The initial learning rate is set to 3×10^{-5}, with a 200-step warm-up starting from 1×10^{-6}, followed by cosine decay to a minimum of 3×10^{-6}. We use the Adam optimizer with a gradient accumulation step of 4. The training and evaluation batch size is 2. We fix the random seed to 42 across all libraries to ensure reproducibility.

The training leveraged DeepSpeed with gradient accumulation (4 steps, micro-batch size 2/GPU) and gradient clipping (max norm=3.0). FP16 mixed-precision training employed dynamic loss scaling (1-256 range, 1000-step window). Memory optimization used ZeRO stage 1 for optimizer state partitioning. The AdamW optimizer (lr=2×10, =(0.9,0.999), =1×10) incorporated decoupled weight decay (0.01).

For parameter-efficient adaptation, we adopt a LoRA fine-tuning scheme implemented via the `peft` library. Unless otherwise specified, the LoRA rank is set to $r = 8$ and the scaling factor to $\alpha = 32$, which provides a moderate increase in effective capacity without incurring prohibitive memory overhead. A dropout rate of 0.1 is applied on the LoRA branches to mitigate overfitting. LoRA adapters are injected only into the attention projection layers, i.e., `q_proj`, `k_proj`, `v_proj`, and `o_proj`, while the remaining backbone parameters are kept frozen during supervised fine-tuning. Inference-only LoRA merging (i.e., `lora_inference_mode = False`) is not used; instead, all reported results are obtained by explicitly applying the learned LoRA updates at run time.

D Framework Components

Intervideo (Wang et al., 2022) integrates generative and discriminative learning by employing Masked Autoencoders (MAE) for action understanding and video-language contrastive learning for semantic alignment. Cross-model attention (CMA) modules further enhance representations during supervised training. This unified approach achieves superior action recognition, video-language alignment, and open-world task performance with efficient training.

Whisper (Radford et al., 2023) employs a Transformer-based encoder-decoder trained on vast web-scale audio-transcript pairs using weak supervision. Its multitask training framework enables a single model to handle speech recognition, translation, and speaker identification by representing tasks as token sequences. This design enhances generalization, robustness, and adaptability without fine-tuning on specific datasets.

Beats (Chen et al., 2022b) iteratively optimizes acoustic tokenizers and audio SSL models by converting audio signals into semantically rich discrete labels. Starting with a random-projection tokenizer clustering audio features, subsequent iterations refine the tokenizer via knowledge distillation from the SSL model. Using a Vision Transformer backbone and a Masked Audio Modeling task, BEATS enhances label prediction, achieving state-of-the-art audio classification with superior robustness and semantic alignment.

DeepSeekMoE (Dai et al., 2024) enhances expert specialization and reduces redundancy using fine-grained expert segmentation and shared expert isolation. This approach splits experts into smaller units with flexible activation and assigns some experts to capture common knowledge. It achieves significant performance gains, nearly matching dense models at 2B parameters and equaling DeepSeek 7B at 16B parameters with only 40% computation, while supporting strong scalability and adaptability.

E Baseline Models

E.1 Traditional Deep Learning Models

AVSD (Alamri et al., 2019) employs a late-fusion architecture: visual features are extracted using a pretrained I3D model, audio features are obtained via a pretrained AENet, while both the dialog history and the current question are encoded using LSTM networks. These four modalities are then fused at the semantic level to form a unified representation, which is used to rank a set of candidate answers and select the most appropriate one. This architecture enables effective integration of multimodal context for generating accurate and contextually relevant responses in dynamic scenes.

PSTP-Net (Li et al., 2023) progressively identifies spatio-temporal regions relevant to the question

through a series of modules. The Temporal Segment Selection Module (TSSM) locates key video segments, the Spatial Region Selection Module (SRSM) pinpoints important visual regions within these segments, and the Audio-guided Visual Attention Module (AVAM) performs fine-grained perception to associate audio with visual content. This step-by-step approach reduces redundancy and enhances both efficiency and accuracy.

LAVisH (Lin et al., 2023) adapts frozen Vision Transformers for audio-visual tasks by injecting trainable latent tokens that compress modality-specific information and enable efficient cross-modal fusion via attention bottlenecks. Its bidirectional design transfers information between audio and visual modalities, improving joint representation. It achieves strong performance on audio-visual tasks with far fewer tunable parameters.

LAViT (Yun et al., 2021) addresses spherical spatial and audio-visual reasoning in 360° videos by integrating visual, audio, and language modalities through separate unimodal encoders and a multimodal encoder. It uses quaternion-based spatial representations for spherical geometry and includes an audio skewness prediction task. Visual features come from Faster R-CNN, audio features from a VGG-like CNN on stereo channels, and language features from BERT. Answers and grounding are predicted via a decoder.

QA-TIGER (Kim et al., 2025) is an audio-visual QA framework that explicitly injects question information into the temporal modeling process. Instead of relying on uniform sampling or discrete Top-K selection, QA-TIGER adopts Gaussian-based continuous temporal weighting to adaptively emphasize both consecutive and non-consecutive frames relevant to the question. A Mixture-of-Experts module further provides multiple Gaussian specialists, allowing the model to activate question-appropriate temporal experts. This design enables fine-grained temporal reasoning and question-aware frame selection, leading to strong performance across AVQA benchmarks.

E.2 Multimodal Large Language Models

QWen2.5-VL (Bai et al., 2023) integrates a powerful visual encoder with the Qwen-7B language model and employs a three-stage training pipeline: pre-training on web-scale image-text pairs, multi-task pre-training with fine-grained data, and supervised fine-tuning for dialogue enhancement. It excels in multilingual conversations, multi-image interactions, and fine-grained visual understanding, outperforming existing vision-language models.

CAT (Ye et al., 2024b) is a multimodal LLM tailored for audio-visual reasoning, designed to reduce ambiguity and strengthen grounding in dynamic scenes. It enhances standard MLLMs through three components: a clue-aggregator that extracts question-relevant cues from audio and video, mixed-modality training including the AVinstruct dataset for richer cross-modal alignment, and an ambiguity-aware preference optimization strategy encouraging precise and unambiguous answers.

MeerKat (Chowdhury et al., 2024) is an audio-visual large language model designed for fine-grained cross-modal reasoning. It incorporates an optimal-transport–based alignment module to better match audio and visual embedding spaces, together with a cross-attention mechanism that enforces local audio-visual consistency. To support diverse temporal and spatial tasks, Meerkat is trained on AVFIT, a curated collection of 3M instruction-tuning samples covering multiple audio-visual settings. The resulting model demonstrates strong generalization across challenging downstream tasks—including audio-referred grounding, cross-modal localization, and fact-checking—and achieves state-of-the-art results on the Meerkat-Bench benchmark.

video-SALMONN (Sun et al., 2024a) introduces a unified end-to-end audio-visual LLM designed to handle not only visual frames and general audio events but also fine-grained speech understanding. To balance speech-level temporal precision with efficient multimodal modeling, the method proposes a multi-resolution causal Q-Former that bridges pretrained audio-visual encoders with the LLM backbone. In addition, diversity-oriented objectives and an unpaired audio-visual mixing scheme are adopted to prevent modality imbalance during training.

VideoLLaMA2 (Cheng et al., 2024) uses a Spatial-Temporal Convolution connector and an Audio Branch to process visual and audio data separately before fusion in a fine-tuned large language model. The STC reduces tokens while preserving spatiotemporal order, and the Audio Branch

employs BEATs for audio features and joint training. This design achieves state-of-the-art video and audio understanding.

VITA (Fu et al., 2024) builds on Mixtral 8×7B and extends its vocabulary through bilingual instruction tuning to enhance understanding of Chinese and English. It integrates visual and audio modalities into the language model using multimodal alignment and instruction tuning, introducing state tokens to differentiate input query types and enabling interaction without wake words. VITA employs a dual-model deployment architecture: one for response generation and another for continuous environmental monitoring, supporting audio-interruption-based interaction.

GPT-4o (Hurst et al., 2024) employs a shared Transformer backbone to jointly model text, audio, and visual inputs in a common semantic space. Through multimodal instruction tuning, it achieves robust cross-modal alignment and task understanding, enabling coherent text generation, low-latency spoken dialogue, and visual reasoning like image interpretation and multimodal question answering. Its native multimodal training enhances generalization and inference efficiency across diverse real-world applications.

Qwen2.5-Omni (Xu et al., 2025) adopts block-wise processing in its audio and visual encoders and introduces TMRoPE, a time-aligned positional embedding that interleaves audio and video tokens to maintain temporal consistency. To enable simultaneous text and speech generation, Qwen2.5-Omni employs a Thinker–Talker architecture, where the Thinker (LLM) handles text reasoning and the Talker autoregressively generates audio tokens based on shared hidden states. A sliding-window DiT further supports low-latency streaming speech decoding. Qwen2.5-Omni matches or surpasses similarly sized multimodal models such as Qwen2.5-VL and Qwen2-Audio, and achieves state-of-the-art results on several multimodal and speech-instruction benchmarks.

MAVEN (Ma et al., 2025) employs a Multi-faceted Cycle Collaborative Debiasing (MCCD) strategy to mitigate bias and improve generalization across in- and out-of-distribution scenarios. It extracts unimodal representations with modality-specific encoders, fuses them via a fine-tuned generative model, and captures unimodal biases through separate bias learners. The MCCD strategy amplifies distribution differences between unimodal and multimodal logits while maintaining bias consistency via cycle guidance, reducing reliance on harmful biases.

Kimi-Audio (Ding et al., 2025) adopts a lightweight 12.5 Hz audio tokenizer, an LLM-centric encoder–decoder architecture that consumes continuous audio features and generates discrete audio tokens, and a streaming detokenizer implemented via flow-matching. Training is performed on over 13M hours of diverse audio (speech, sound events, music), followed by curated instruction tuning to support tasks such as audio understanding, audio QA, transcription, and speech-based dialogue. Initialized from a pretrained LLM, Kimi-Audio is continually trained on joint audio–text tasks and then fine-tuned for instruction following. Across a broad set of audio benchmarks—covering ASR, audio classification, audio QA, and speech interaction—Kimi-Audio achieves competitive or state-of-the-art performance, making it a strong baseline for audio-centric multimodal reasoning.

F Attentive Visualization Analysis

To verify the effectiveness of our proposed Question-Guided Pooling mechanism, we visualize the attention scores computed between the question semantics and video frames over time. As shown in Figure 4, we take a representative example where the question is "*Is the instrument on the left louder than the instrument on the right?*". Our model generates a temporal attention map based on the semantic alignment between the question and the visual frames. Notably, the visual attention consistently emphasizes the spatial regions associated with "instrument", "left", and "right" across frames, aligning well with the core semantics of the question. The 4 attention weights, shown below each frame, further highlight the temporal consistency and discriminative capacity of learned attention. These results demonstrate the interpretability and efficacy of our Question-Guided Pooling strategy in capturing modality-question interactions for audiovisual reasoning.

G Attentive Pooling vs. Average Pooling

To evaluate the effectiveness of our Question-Guided Pooling (QGP), we compare it with average pooling (AP). As shown in Table 5 and Ta-

Figure 4: Attention visualization of Question-Guided Pooling. Red and blue boxes denote the highest and lowest attention scores at each frame respectively. Effective and interpretable guidance from question semantics is demonstrated by the meaningful attention distribution. Above: MUSIC-AVQA sample. Below: FortisAVQA sample.

ble 6, QGP consistently outperforms AP on both MUSIC-AVQA and FortisAVQA, with overall accuracy gains of +1.03% and +1.92%, respectively. The improvements are particularly pronounced in audio-visual reasoning, e.g., +3.73% on *Existence* and +2.43% on *Counting* in MUSIC-AVQA, and up to +10.02% on compositional AVQA in FortisAVQA. These results highlight that leveraging question semantics for pooling preserves task-relevant cross-modal cues, whereas uniform averaging dilutes them. As shown in Table 1, our model with joint fusion design achieves a new state-of-the-art accuracy of 86.36 overall, outperforming the current sota method by +8.74% (86.36 vs. 77.62% from MAVEN). Our model also witnesses such performance advantages in nearly every modality's questions, outperforming the second best model on audio tasks by +7.45% (86.43% vs. 78.58% from QA-TIGER), visual tasks by 3.35% (88.49% vs. 85.14% from QA-TIGER), and audio-visual tasks by +11.51% (85.25% vs. 73/74% from QA-TIGER). Interestingly, the difference compared with the second best model in visual modality's subtype questions is the smallest, namely +0.38% improvement (count: 84.34% vs. 83.96% from QA-TIGER) and –2.34% decrease in local tasks (91.16% vs. 93.50% from MAVEN) within the visual modality. The largest improvement is evidenced in comparison questions within audio modality (+12.80% compared with CAT, 82.95% vs. 70.15%), comparison tasks (+16.78% compared with QA-TIGER, 82.95% vs. 69.59%), and temporal tasks (+11.68% compared with MeerKat, 86.37% vs. 74.69%) within audio-visual modality.

The significant outperformance in audio and audio-visual modalities indicates that our method facilitates the audio information to be robustly fused with visual information for multimodal question answering. And the best performance varies across different questions under different fusion strategies, which verifies the flexibility of our Flex-FuseMoE design, which can adaptively realize the best fusion strategy under different task conditions.

We further visualize the attention scores between question semantics and video frames over time, which provides additional evidence supporting the validity of our design, which can be found in Appendix F.

H Kernel and Stride Configurations

We investigate the impact of different kernel and stride configurations under the *joint routing* fusion strategy. As shown in Figure 5, the best performance is achieved with an audio and visual kernel-stride setting of **[6,8]** and **[6,6,8]**, reaching an accuracy of 86.36% on MUSIC-AVQA, 82.15% on FortisAVQA and 82.25% on FortisAVQA-Tail test set respectively, except for FortisAVQA-Head set, the best accuracy 82.76% is achieved when kernel is set as **[8,10]** and stride as **[8,8,10]**. Interestingly, the results suggest that smaller kernel sizes do not always yield better performance. Instead, a comparatively larger but task-appropriate kernel and stride configuration strikes a better balance between spatial-temporal compression efficiency and model effectiveness, highlighting the importance of task-adaptive design rather than simply adopting minimal receptive fields.

I Qualitative Examples

We provide qualitative case studies comparing our EQUALS framework with Qwen2.5-VL. Each figure shows a representative sample from MUSIC-AVQA or FortisAVQA, with the corresponding question, ground truth, and predictions. We provide qualitative case studies comparing EQUALS with Qwen2.5-VL. While Qwen2.5-VL performs reasonably well on purely visual queries, it fails to capture audio-dependent reasoning. In these cases,

Table 5: Average Pooling vs.Question Guided Pooling (Joint Fusion) on AVQA tasks of the MUSIC-AVQA dataset.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
AP(joint)	85.97	81.44	85.31	83.31	89.88	85.10	87.38	79.57	85.49	81.85	84.61	85.53	85.33
QGP(joint)	88.46	82.95	86.43	83.39	91.09	88.28	91.11	82.00	86.47	82.95	86.37	85.25	86.36
Δ (QGP–AP)	+2.49	+1.51	+1.12	+0.08	+1.21	+3.18	+3.73	+2.43	+0.98	+1.10	+1.76	-0.28	+1.03

Table 6: Average Pooling vs. Question Guided Pooling (Joint Fusion) on FortisAVQA . (H: Head; T: Tail).

Method	Audio QA				Visual QA				AVQA										Avg
	Count		Comp		Count		Local		Exist		Count		Local		Comp		Temp		
	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T	
AP(joint)	78.43	78.35	76.00	85.83	83.65	63.16	85.15	85.68	89.10	83.33	77.82	48.19	76.69	85.19	73.93	68.26	76.89	88.61	80.23
QGP(joint)	80.16	79.35	77.33	87.50	86.06	64.47	86.41	86.87	90.23	88.46	78.86	50.00	80.68	86.30	79.92	78.28	77.86	89.60	82.15
Δ(QGP-AP)	+1.73	+1.00	+1.33	+1.67	+2.41	+1.31	+1.26	+1.19	+1.13	+5.13	+1.04	+1.81	+3.99	+1.11	+5.99	+10.02	+0.97	+0.99	+1.92

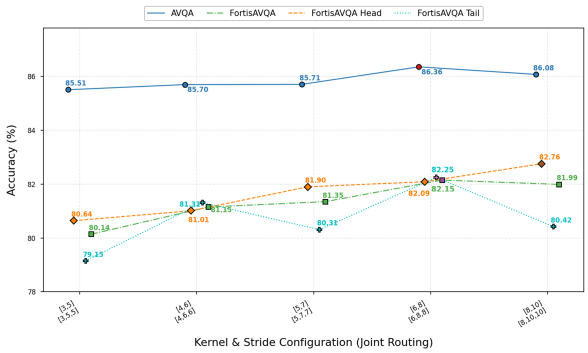

Figure 5: Accuracy comparison of different kernel and stride configurations under the per-modality routing fusion strategy. The best performance is achieved primarily with the [6,8], [6,6,8] configuration, except [8,10], [8,8,10] for FortisAVQA-Head test set, suggesting that in most case moderately-sized kernels and stides can better balance compression and representation capacity than smaller ones in most cases.

Figure 6: Sample eva00001188.mp4.

Sample ID: eva00001188.mp4

Question: Where is the lowest instrument?

Ground Truth: left

EQUALS: left

Qwen2.5-VL: Predicts drum set at the **bottom** of the stage.

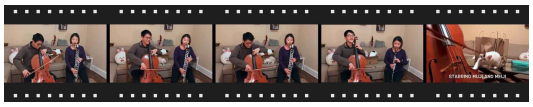

Figure 7: Sample 00004924.mp4.

Sample ID: 00004924.mp4

Question: How many instruments did not sound from beginning to end?

Ground Truth: zero

EQUALS: zero

Qwen2.5-VL: In the video, there are three instruments: violin, cello, and piano. The piano does not play continuously, and thus is **predicted as the instrument that did not sound from beginning to end**.

its outputs are either **incorrect** or based on **speculative guesses from visual cues only**, without leveraging the audio modality. In contrast, EQUALS consistently produces accurate predictions by explicitly incorporating audio features, demonstrating the necessity of cross-modal alignment for robust AVQA.

Figure 8: Sample 00004166.mp4.

Sample ID: 00004166.mp4

Question: Is this sound from the instrument in the video?

Ground Truth: yes

EQUALS: yes

Qwen2.5-VL: The video shows three individuals in a practice room. One plays piano, another a wind instrument, and a third sings. **The model does not confirm the audio source, stating it is not possible without sound**.

Figure 9: Sample 00007348.mp4.

Sample ID: 00007348.mp4

Question: What is the left instrument of the first sounding instrument?

Ground Truth: trumpet

EQUALS: trumpet

Qwen2.5-VL: The left instrument in the image is a trombone, which is being played by the musician on the left. The first sounding instrument, which is the one that starts playing, **appears to be the trombone**.