QRePair: QUESTION-GUIDED RECONSTRUCTION VIA PAIRED PROTOTYPE RETRIEVAL FOR MISSING-MODALITY AUDIO-VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-Visual Question Answering (AVQA) requires reasoning over audio, visual, and linguistic cues, yet one modality may be unavailable at inference time, leaving question-relevant evidence unobserved. Semantic retrieval can provide useful priors, but may not preserve structured evidence such as source count or temporal configuration. We propose Question-Guided Reconstruction via Paired Prototype Retrieval (QRePair), which reconstructs task-relevant missing-modality representations from paired temporal prototypes. Because relevant evidence depends on the question, Question-Guided Modality Encoding first produces question-conditioned audio and visual representations. Paired Prototype Construction organizes complete training representations into temporally resolved audio-visual prototype pairs, and Paired Prototype Retrieval uses the available representation to retrieve the corresponding missing-modality prior. Question-Guided Reconstruction then adapts this prior to the current question and the feature space of a frozen answering backbone. On 5,096 audio-visual test questions from MUSIC-AVQA, QRePair achieves 71.59% and 73.49% accuracy with visual and audio inputs missing, respectively, representing gains of 2.22 and 1.43 percentage points over the reported R^2ScP results. Ablations further show that disrupting audio-visual correspondence reduces average missing-modality accuracy by 4.25 points, while removing explicit question conditioning from reconstruction reduces it by 8.96 points. QRePair also shows strong performance on AVQA and fully observed MUSIC-AVQA-R and MUSIC-AVQA 2.0 benchmarks.

1 INTRODUCTION

Audio-Visual Question Answering (AVQA) relates a question to complementary auditory and visual evidence (Li et al., 2022; Yang et al., 2022). Vision can reveal source identity and location, while audio can expose off-screen events or distinguish visually ambiguous activities (Huang et al., 2023). Answering a question can therefore require information that neither modality provides alone.

Audio or visual inputs can be unavailable at deployment time due to limited sensor coverage, transient failures, or data transmission errors (Ma et al., 2022; Wang et al., 2023). A question may require counting sources, localizing a brief event, or comparing cues across moments and modalities, each of which can depend on the missing stream. The model must then reason from the remaining cues and learned cross-modal representations. Retrieved features can provide a useful prior, although they cannot guarantee recovery of instance-specific facts that are unidentifiable from the observed input.

Existing AVQA methods strengthen question-relevant perception through spatial selection, temporal modeling, and multimodal interaction (Li et al., 2023; 2024a; Kim et al., 2025). These mechanisms improve the use of observed evidence, but cannot by themselves supply information lost when a modality is unavailable. Missing-modality approaches therefore generate absent features from the remaining cues (Park et al., 2024) or retrieve complementary evidence from training data (Zhang et al., 2026). Semantic similarity can identify samples sharing related objects, events, or scene content, but semantic agreement alone does not guarantee preservation of the structured evidence required

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

by a question, such as event multiplicity, temporal configuration, or modality-specific details. For example, two clips may depict similar musical performances while differing in how many sources are active, when they occur, or how they are ordered. This motivates retaining structured cross-modal priors and adapting them to the current question rather than relying on semantic similarity alone.

We address this setting with Question-Guided Reconstruction via Paired Prototype Retrieval (QRePair). Because the evidence relevant to a clip depends on the question, Question-Guided Modality Encoding (QG-ME) first maps the audio and visual streams into question-conditioned representations, allowing the same clip to emphasize different evidence for different questions. Paired Prototype Construction (PPC) compresses these complete training representations into a compact set of paired temporal audio-visual prototypes while preserving cross-modal correspondence. At inference, Paired Prototype Retrieval (PPR) uses the available question-conditioned representation to identify a matching prototype pair and retrieves its counterpart for the missing modality. Question-Guided Reconstruction (QGR) then adapts the retrieved prior to the current question and to the feature space expected by the frozen answering backbone. Complete-input features and answer distributions supervise reconstruction during training.

We evaluate QRePair on MUSIC-AVQA (Li et al., 2022) and AVQA (Yang et al., 2022) with visual or audio inputs missing. On the audio-visual questions of MUSIC-AVQA, QRePair obtains accuracies of 71.59% and 73.49% with visual and audio inputs missing, respectively, 2.22 and 1.43 percentage points above the reported R^2ScP values. On the official AVQA test split, the corresponding accuracies are 76.43% and 88.91%. Both datasets use their official benchmark splits, with prototypes constructed from training examples. The main contributions of this paper are as follows.

- We introduce a compact set of paired temporal prototypes that preserves audio-visual sample correspondence while reducing storage. At L2 granularity, the set retains 28.8% of the pairs with a 0.23-point reduction in average missing-modality accuracy relative to no clustering (Figure 2).
- We introduce paired observed-to-missing retrieval followed by question-guided reconstruction: the available modality identifies a prototype pair, its missing-modality counterpart provides a complementary prior, and QGR adapts that prior to the current question and the frozen answering backbone. Benchmark comparisons assess missing-modality accuracy, while feature-space analysis examines how reconstruction refines the retrieved representation (Figure 3).

2 Related Work

2.1 Missing-Modality Learning in AVQA

Missing-modality methods recover absent features or adapt reasoning to the available inputs. Within AVQA, Missing-AVQA (Park et al., 2024) predicts missing features from the available modality and the question, followed by diffusion-based enhancement. MoMKE (Xu et al., 2024) combines unimodal and joint representations through modality experts, while SimMLM (Li et al., 2025) uses dynamic expert gating and a ranking objective across modality subsets. IMOL (Zeng et al., 2025), developed for fake news video detection, combines missing-feature reconstruction with retrieval-augmented contrastive learning. These approaches support incomplete-input prediction through feature recovery, expert adaptation, or cross-sample alignment.

For AVQA, R^2ScP (Zhang et al., 2026) retrieves candidates through cross-modal similarity in a shared semantic space and applies question-guided purification. QRePair maintains explicitly paired temporal audio-visual prototypes and matches within the observed modality to retrieve the corresponding missing-modality prior. It then adapts this prior to the intermediate feature space used by the answering backbone.

Prototypes and cross-sample retrieval. Prototypical networks summarize support examples as class centroids (Snell et al., 2017), while nonparametric instance discrimination stores example embeddings in a memory bank (Wu et al., 2018). Retrieval also supports incomplete multimodal prediction: RAMER retrieves related multimodal examples to supplement missing information in emotion recognition (Fan et al., 2024). QRePair builds on these uses of stored representations by

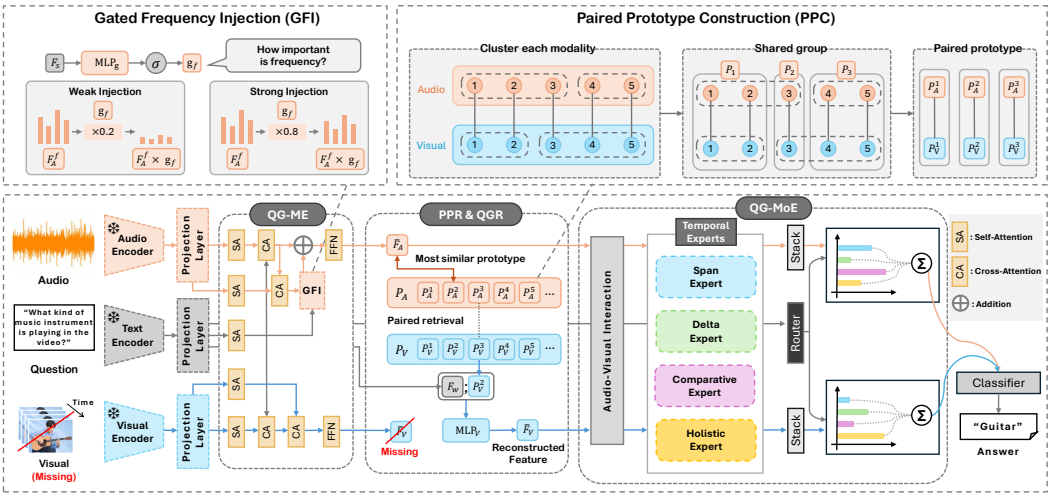

Figure 1: **Overall architecture of the proposed QRePair framework.** PPC constructs a compact set of paired temporal prototypes. PPR uses the observed modality to retrieve a complementary prototype, and QGR adapts it to the question. The supporting backbone encodes the inputs and reasons over available and reconstructed features; frequency injection supports audio encoding.

coupling paired temporal audio-visual prototypes with observed-to-missing retrieval and question-guided reconstruction for AVQA.

2.2 Question-Conditioned Reasoning in AVQA

A dominant line of AVQA research focuses on question-conditioned perception. Early work such as ST-AVQA (Li et al., 2022) couples spatial grounding with temporal attention, after which PSTP-Net (Li et al., 2023), TSPM (Li et al., 2024a), and APL (Li et al., 2024b) progressively refine question-relevant cues at segment, temporal-spatial, and object levels. More recent methods strengthen question-aware temporal modeling: QA-TIGER (Kim et al., 2025) applies Gaussian experts along the temporal axis for continuous question-relevant weighting. QSTar (Li et al., 2026) studies query-guided spatial, temporal, and frequency interaction.

These methods improve reasoning by selecting, weighting, or temporally organizing question-relevant evidence from the observed modalities. However, such mechanisms primarily operate on evidence that is already present in the input; when an entire modality is unavailable, question-guided selection alone cannot supply the missing complementary representation. QRePair therefore complements question-conditioned reasoning with paired prototype retrieval and reconstruction. The retrieved representation provides a cross-modal prior for the missing stream, while QGR adapts this prior to the current question before it is processed by the reasoning backbone.

3 Method

QRePair has three core steps: **Paired Prototype Construction (PPC)**, **Paired Prototype Retrieval (PPR)**, and **Question-Guided Reconstruction (QGR)**. PPC builds compact prototype pairs from complete training representations. PPR selects a complementary prototype using the available modality, and QGR adapts it to the current question. The supporting backbone uses **Question-Guided Modality Encoding (QG-ME)** to prepare representations and **Question-Guided Mixture-of-Experts (QG-MoE)** for temporal reasoning and answer prediction. We first train this backbone on complete inputs, then construct the prototypes and learn reconstruction with the backbone fixed.

Input representation. We represent each clip with T temporal steps and a common channel dimension D. Pretrained visual and audio encoders provide frame-level features $F_V, F_A \in \mathbb{R}^{T\times D}$, together with visual patch features F_P and frequency-level audio features F_{Freq}. A text encoder

provides word-level question features $F_W \in \mathbb{R}^{N\times D}$ for N question tokens and a sentence embedding $F_S \in \mathbb{R}^D$. The question remains available in every evaluation condition; an entire audio or visual feature stream is masked to simulate missing inputs. Encoder choices and feature preparation are detailed in Appendix C.1.

3.1 Question-Guided Modality Encoding

Prototype construction and retrieval operate on encoded modality representations, so the evidence emphasized before retrieval determines which training examples are considered similar. A clip can support questions about instrument identity, temporal order, or spatial location, each requiring different cues. We therefore condition each modality on the question before constructing or querying the prototype set. QG-ME serves as the supporting representation backbone and uses standard self- and cross-attention operations (Vaswani et al., 2017).

For $m \in \{A, V\}$, self-attention (SA) models dependencies within each modality, and cross-attention (CA, with query, key, and value arguments in that order) relates it to the question; LN denotes layer normalization:

$$\begin{aligned} \hat{F}_m &= \mathrm{SA}(F_m), \qquad \hat{F}_W = \mathrm{SA}(F_W), \\ F_m^q &= \mathrm{LN}\Big(F_m + \mathrm{CA}(\hat{F}_m, \hat{F}_W, \hat{F}_W)\Big). \end{aligned} \tag{1}$$

The visual stream then attends to patch features for spatial detail. The audio stream uses the frequency information described below. Residual connections, normalization, and feed-forward layers yield $\bar{F}_V$ and $\bar{F}_A$; Appendix C.1 gives the full operations. During prototype construction these features are conditioned on the training question; during retrieval the observed stream is conditioned on the current question.

Gated Frequency Injection (GFI). For audio encoding, GFI retains question-dependent spectral detail by reintroducing frequency-level evidence that may be attenuated by frequency averaging. Cross-attention selects this evidence, and a question-conditioned scalar gate controls its contribution. Full operations are given in Appendix C.2.

3.2 Paired Prototypes and Question-Guided Reconstruction

The observed modality may be compatible with several missing-modality representations, so reconstruction requires a complementary prior as well as the current input. PPC compresses complete training representations into a compact set of paired audio-visual prototypes. PPR retrieves the missing-modality counterpart of an observed-modality match, and QGR adapts this prior to the current question.

Paired Prototype Construction (PPC). After Phase 1, the frozen QG-ME encoder extracts paired complete-input training representations $(\bar{F}_V^i, \bar{F}_A^i)$ conditioned on their training questions. PPC uses two-stage cross-modal grouping: first, FINCH (Sarfraz et al., 2019) independently clusters temporally centered and compressed audio and visual descriptors; second, their assignments are intersected at the sample level. Let $c_A(i)$ and $c_V(i)$ be the audio and visual cluster indices assigned to paired training example i, with K_A and K_V clusters, respectively:

$$c_A(i) \in \{1, \ldots, K_A\}, \qquad c_V(i) \in \{1, \ldots, K_V\}. \tag{2}$$

Audio and visual cluster labels belong to independent index spaces. For each label pair (u, v), we define the sample intersection

$$\mathcal{C}_{u,v} = \{i \mid c_A(i) = u \wedge c_V(i) = v\}. \tag{3}$$

We re-index the J non-empty intersections as $\{\mathcal{C}_j\}_{j=1}^J$, with $|\mathcal{C}_j|$ examples in group j, and average the original, uncentered temporal representations over each shared sample set:

$$P_A^j = \frac{1}{|\mathcal{C}_j|}\sum_{i\in\mathcal{C}_j} \bar{F}_A^i, \qquad P_V^j = \frac{1}{|\mathcal{C}_j|}\sum_{i\in\mathcal{C}_j} \bar{F}_V^i. \tag{4}$$

Here, a *prototype* denotes a representative temporal feature sequence for a paired group, rather than a single pooled feature vector. For $m \in \{A, V\}$, each $P_m^j \in \mathbb{R}^{T\times D}$ is obtained by averaging the

modality representations assigned to the same group while retaining their temporal indices. The shared group index j links the audio and visual prototypes, allowing a prototype retrieved in the observed modality to provide the corresponding representation in the missing modality.

We compress each prototype along the channel dimension while retaining an ordered temporal sequence for decoding. Averaging preserves temporal indices but does not explicitly align event onsets across clips. The prototype set is fixed during reconstruction training. Descriptor generation, compression, and storage accounting are specified in Appendix D.

Paired Prototype Retrieval (PPR). A reconstruction module conditioned only on the question can provide a task-relevant surrogate for the missing stream, but cannot condition that representation on the observed modality. QRePair therefore retrieves a complementary prior using the observed stream: an observed-modality match identifies a prototype pair, and its paired prototype provides a candidate representation of the missing stream.

Under the standard controlled zero-masking protocol, modality availability is identifiable from the masking pattern and is not treated as a contribution. The model receives no state label at inference. A lightweight state predictor first selects the complete, visual-missing, or audio-missing state; complete inputs bypass reconstruction, and a predicted missing state activates only the corresponding reconstruction direction. Its probability also supplies a soft gate (Appendix C.3). Let m be the modality predicted to be missing and o the other modality. Retrieval uses the encoded observed stream to select its matching prototype:

$$j^* = \arg\max_j \operatorname{sim}_o\big(\bar{F}_o, P_o^j\big)\,. \tag{5}$$

Here, sim_o is cosine similarity in the temporally centered, channel-projected space, using the stored quantized representation of each prototype. We retrieve the paired prototype $P_m^{j^*}$ at the same index. Thus, the predicted visual-missing branch matches audio keys to retrieve a visual prototype, and the predicted audio-missing branch matches visual keys to retrieve an audio prototype. During Phase 2 training, we exclude the prototype pair containing the current example by its construction-group index before similarity-based selection: example i can retrieve only pairs with $i \notin \mathcal{C}_j$.

Question-Guided Reconstruction (QGR). The retrieved prototype is a cross-modal prior, not an exact representation of the current sample. Its shared group representation need not match the evidence emphasized by the current question. QGR therefore adapts this prior to the current question and the feature space expected by the frozen reasoning backbone. QG-ME conditions the observed representation *before retrieval*, whereas QGR conditions the retrieved missing-modality prior *after retrieval*.

Let $d_m(P_m^{j^*})$ denote the temporal sequence decoded using the stored magnitude, temporal mean, and channel basis. With the token-pooled question vector $\bar{w} = \mathrm{Pool}(F_W)$, the modality-specific MLP ϕ_m reconstructs the feature at each time step $t \in \{1, \ldots, T\}$; $[\,;\,]$ denotes channel concatenation:

$$C_{m,t} = \phi_m([d_m(P_m^{j^*})_t; \bar{w}]). \tag{6}$$

Within the selected branch, the soft gate blends the reconstructed sequence $C_m \in \mathbb{R}^{T\times D}$ (with rows $C_{m,t}$) with the encoded missing stream to produce F_m^+, while $F_o^+ = \bar{F}_o$. QG-MoE receives these available and reconstructed features. Appendix C.3 defines the state probabilities and reconstruction gates. The controls in Table 2 examine the prototype set, audio-visual pairing, QGR, and its question conditioning.

3.3 Backbone Reasoning

Reconstructed features still need to support different temporal questions. QG-MoE is the supporting reasoning backbone that combines localization, change detection, comparison, and global context. Gated bidirectional cross-attention first relates the available and reconstructed streams, producing F_{av} (audio enriched by visual features) and F_{va} (visual enriched by audio features).

For each sequence $X \in \{F_{av}, F_{va}\}$, **Span** aggregates question-relevant temporal features, **Delta** summarizes temporal changes at several scales, **Comparative** contrasts two pooled summaries, and **Holistic** attends to the sequence as a whole. These complementary operators produce outputs e^j,

Table 1: Accuracy (%) on MUSIC-AVQA and AVQA with complete and missing modalities.

Method	Venue	Modalities		MUSIC-AVQA (Audio-Visual)						AVQA
		A	V	Exist	Localis	Count	Comp	Temp	Avg.	Avg.
Missing-AVQA	ECCV'24	•	○	77.94	58.48	67.43	65.21	58.88	65.99	46.65
		○	•	78.74	58.15	70.59	62.67	60.46	66.44	70.28
		•	•	82.94	68.70	75.13	61.30	69.87	71.27	89.96
		Average		79.87	61.77	71.05	63.06	63.07	67.90	68.96
MoMKE	MM'24	•	○	79.90	59.61	65.16	60.32	58.93	64.82	59.78
		○	•	78.74	58.15	70.59	62.67	61.18	66.44	70.12
		•	•	82.24	67.03	74.96	63.58	70.16	71.34	90.26
		Average		80.29	61.60	70.24	62.19	63.42	67.53	73.39
SimMLM	ICCV'25	•	○	79.35	63.59	65.77	54.46	68.49	65.64	59.95
		○	•	81.07	60.17	69.25	63.03	60.46	66.94	70.88
		•	•	82.78	66.92	75.81	64.02	69.24	71.57	90.32
		Average		81.07	63.56	70.28	60.50	66.06	68.05	73.72
IMOL	ACL'25	•	○	81.10	61.33	70.83	61.16	61.95	67.11	61.32
		○	•	81.72	65.87	69.80	61.98	68.25	69.21	72.38
		•	•	83.49	70.35	74.68	63.21	69.10	71.86	90.28
		Average		82.10	66.18	71.77	62.12	66.43	69.39	74.66
R^2ScP	ACL'26	•	○	81.07	65.54	71.32	61.95	67.51	69.37	63.25
		○	•	82.49	68.93	74.05	63.49	69.34	72.06	75.12
		•	•	83.59	70.57	75.95	64.08	70.75	73.19	90.64
		Average		82.38	68.35	73.77	63.17	69.61	71.54	76.35
QRePair(ours)	–	•	○	83.00	66.63	76.44	62.94	67.52	71.59	76.43
		○	•	81.68	71.52	78.58	63.67	71.17	73.49	88.91
		•	•	82.79	71.85	78.50	64.58	71.05	73.92	91.87
		Average		**82.49**	**70.00**	**77.84**	**63.73**	**69.91**	**73.00**	**85.74**

•/○: available/missing; A/V: audio/visual. Questions are always provided. Average: mean across the three observation settings. Baseline results are from Zhang et al. (2026). MUSIC-AVQA uses 5,096 Audio-Visual test questions. QRePair and the AVQA baselines use the same official test split and evaluation protocol. Baseline results are retained as published.

indexed by $\mathcal{E} = \{\text{span}, \Delta, \text{comp}, \text{hol}\}$. A question-conditioned router combines them:

$$\rho = \text{softmax}(\text{MLP}_{\text{temp}}(F_S)), \qquad f_X = \sum_{j \in \mathcal{E}} \rho_j e^j. \tag{7}$$

Here, $\rho_j \geq 0$ is expert j's question-dependent weight, with $\sum_{j \in \mathcal{E}} \rho_j = 1$. The vector f_X summarizes sequence X; $f_{av} = f_{F_{av}}$ and $f_{va} = f_{F_{va}}$ denote the two directional summaries. Both directions share expert parameters and routing weights, and all four experts are evaluated. The exact interaction and expert operations are given in Appendix C.4.

Answer prediction. An MLP uses the concatenated summaries $[f_{av}; f_{va}]$ to predict answer logits on MUSIC-AVQA. On AVQA, it produces a joint representation that scores candidate answer embeddings by cosine similarity.

3.4 Two-Phase Training

Phase 1 learns a complete-input reasoning space. After prototype construction, we freeze the backbone so that QGR learns to match fixed feature and answer-distribution targets during Phase 2.

Phase 1: backbone learning. We train the supporting QG-ME and QG-MoE backbone on complete audio-visual inputs, bypassing retrieval and reconstruction. The objective combines the answer classification loss with expert load balancing:

$$\mathcal{L}_{\text{P1}} = \mathcal{L}_{\text{QA}} + \lambda_{\text{lb}} \mathcal{L}_{\text{lb}}. \tag{8}$$

Here, $\mathcal{L}_{\mathrm{QA}} = -\sum_a y_a \log p_a$ is answer cross-entropy: y_a is the smoothed target for answer a and p_a its softmax probability. For load balancing, $\mathcal{L}_{\mathrm{lb}} = 4\sum_{j\in\mathcal{E}} \bar{\rho}_j \nu_j$, where $\bar{\rho}_j$ is expert j's minibatch mean weight and ν_j the fraction of questions for which it has the largest weight. The scalar λ_{lb} weights this auxiliary term. We then construct paired temporal prototypes from the training representations using PPC and freeze the backbone.

Phase 2: question-guided reconstruction. Phase 2 consists of a *clean stream* and a *missing stream* sharing the frozen backbone. The clean stream bypasses PPR and QGR and provides target features $\bar{F}_m^{\mathrm{cl}}$ and an answer distribution p^{cl}. The missing stream masks one modality, retrieves a complementary prototype through PPR, reconstructs its representation through QGR, and predicts p_m^{mis}. Both visual-missing and audio-missing conditions use this same process. The feature loss encourages compatibility with the clean intermediate representation, while distillation preserves the answer distribution produced from complete evidence:

$$\mathcal{L}_{\mathrm{rec}} = \sum_{m\in\{V,A\}} \frac{\|F_m^{+} - \bar{F}_m^{\mathrm{cl}}\|_F^2}{TD}, \qquad \mathcal{L}_{\mathrm{KD}} = \frac{1}{2}\sum_{m\in\{V,A\}} D_{\mathrm{KL}}\big(p^{\mathrm{cl}} \,\|\, p_m^{\mathrm{mis}}\big)\,. \tag{9}$$

Each F_m^{+} is evaluated with modality m masked. The squared Frobenius norm $\|\cdot\|_F^2$, normalized by TD, gives per-modality mean squared errors; D_{KL} is KL divergence over answers. The state loss $\mathcal{L}_{\mathrm{state}}$ averages cross-entropy over complete, visual-missing, and audio-missing labels. With scalar loss weights $\lambda_{\mathrm{rec}}, \lambda_{\mathrm{KD}}, \lambda_{\mathrm{state}}$, the Phase 2 objective is

$$\mathcal{L}_{\mathrm{P2}} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}} + \lambda_{\mathrm{state}}\mathcal{L}_{\mathrm{state}}. \tag{10}$$

Only QGR and the state predictor are updated; the prototypes and the backbone remain fixed. State gates are detached from the reconstruction and distillation losses; the state predictor is trained by its classification loss. At inference, the state predictor controls the reconstruction branch and gate; the clean stream is used only during training. Details of the optimizer and checkpoint selection are provided in Appendix E.

4 Experiments

4.1 Datasets and Experimental Setup

We evaluate QRePair on MUSIC-AVQA (Li et al., 2022) and AVQA (Yang et al., 2022) with complete and missing inputs, and on MUSIC-AVQA-R (Ma et al., 2024) and MUSIC-AVQA 2.0 (Liu et al., 2024) under fully observed distribution shifts. We use official splits and the complete-input, visual-missing, and audio-missing settings of prior comparisons (Park et al., 2024; Zhang et al., 2026). Controlled masking fixes available evidence for evaluating reconstruction and answering. Main comparisons and component ablations use the same 5,096 Audio-Visual MUSIC-AVQA test questions. AVQA uses the baselines' official test split and protocol, with checkpoint-selection validation data held out from training (Appendix B).

After learning the backbone with complete inputs, we train QGR and the state predictor for 15 epochs with the backbone and prototypes fixed. Following the reporting convention of prior missing-modality AVQA studies (Park et al., 2024; Zhang et al., 2026), QRePair results use a fixed random seed (seed 42), except for the multi-seed results in Appendix A.2. PPC uses FINCH clustering (Sarfraz et al., 2019) to construct paired temporal prototypes, with low-rank channel compression and simulated INT4 quantization. We select the second FINCH hierarchy level (L2) using the validation-set accuracy–storage trade-off and fix it before test evaluation. Figure 2 subsequently reports test-set sensitivity across alternative hierarchy levels.

4.2 Results and Analysis

Missing-modality settings. Table 1 evaluates QRePair under complete and missing-modality settings. On MUSIC-AVQA, QRePair obtains accuracies of 71.59 and 73.49 with visual and audio inputs missing, respectively, 2.22 and 1.43 percentage points above the reported R^2ScP values. Its average across the three settings reaches 73.00, exceeding R^2ScP by 1.46 points. The gains vary by question type: counting improves by 5.12 and 4.53 points in the two missing settings, compared with

0.01 and 1.83 points for temporal questions. The gains are largest on counting questions (+5.12 and +4.53 points) and smallest on temporal questions (+0.01 and +1.83 points), suggesting that the benefit varies across question types.

On the official AVQA test split, QRePair obtains 76.43 with visual input missing and 88.91 with audio input missing. The reported R^2ScP values are 63.25 and 75.12, giving differences of 13.18 and 13.79 percentage points. Relative to complete inputs, QRePair loses 15.44 and 2.96 points, compared with 27.39 and 15.52 for R^2ScP. These within-model drops describe missing-input behavior alongside complete-input accuracy.

Table 2: Component ablations and accuracy (%) on the 5,096 Audio-Visual MUSIC-AVQA test questions with missing inputs.

Ablation group	Setting	Prototype prior	Correct A–V pair	QGR	Question in QGR	V miss	A miss	Avg_2
No retrieved prior	Backbone only	✗	–	✗	–	59.24	62.30	60.77
	Direct reconstruction	✗	–	MLP	✓	64.25	71.35	67.80
Paired retrieval	Shuffled A–V pairs	✓	✗	✓	✓	64.80	71.78	68.29
Question-guided reconstruction	w/o QGR	✓	✓	✗	–	67.41	57.28	62.34
	w/o question conditioning	✓	✓	✓	✗	69.19	57.97	63.58
Full model	**QRePair**	✓	✓	✓	✓	**71.59**	**73.49**	**72.54**

Backbone only: no retrieval or reconstruction; –: not applicable. Direct reconstruction: an MLP reconstructs the missing features from the question alone, without a prototype prior. Question in QGR: explicit question input to QGR or this MLP; the backbone is always question-conditioned. Avg_2: mean of V miss and A miss, before rounding.

With both modalities available, QRePair obtains 78.07% on the full 9,129-question MUSIC-AVQA test set (Table F1).

Full-modality settings. QRePair obtains 78.07% on MUSIC-AVQA with both modalities available, compared with 77.62% for QA-TIGER and 78.98% for QSTar (Li et al., 2026). Appendix F reports these competitive full-modality results and the MUSIC-AVQA-R and MUSIC-AVQA 2.0 evaluations. Across three random seeds, visual- and audio-missing accuracies on the MUSIC-AVQA Audio-Visual subset are 71.01 ± 0.61 and 73.11 ± 0.50, respectively (mean $\pm$ sample SD; Appendix A.2), indicating limited variation in these runs.

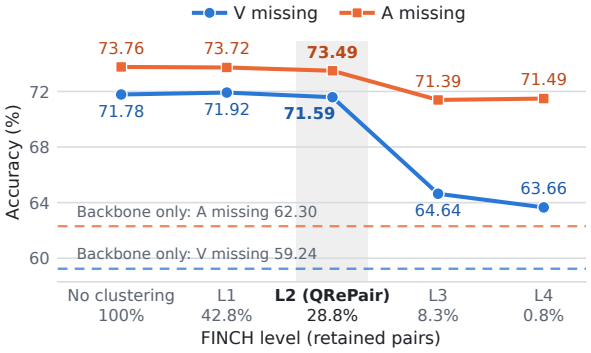

Figure 2: Test-set FINCH-level sensitivity on 5,096 Audio-Visual MUSIC-AVQA questions. Percentages: retained pairs; shaded L2: validation-selected setting; dashed lines: backbone only. The codec and Phase 2 protocol are fixed across levels.

4.3 Ablation and Component Analysis

Ablation study. Table 2 examines paired priors and reconstruction on the 5,096 Audio-Visual test questions. The backbone alone reaches $\text{Avg}_2 = 60.77$. Direct reconstruction from the question raises this to 67.80 (+7.03 points), while full QRePair reaches 72.54 (+4.74 further). This suggests that a question-conditioned surrogate for the missing stream already helps, while the retrieved prior provides an additional benefit when combined with QGR. Shuffling audio-visual correspondence lowers accuracy to 68.29 (−4.25), consistent with a role for correct cross-modal pairing beyond access to a prototype set.

Using retrieved prototypes without QGR yields 62.34 (-10.20 points), suggesting that retrieval alone supplies an insufficiently adapted representation. QGR without explicit question input reaches 63.58, only 1.24 points above direct use of the prototype, whereas question-conditioned QGR reaches 72.54 ($+8.96$). This pattern suggests that question-specific adaptation contributes substantially beyond an unconditioned nonlinear mapping. The encoder and reasoning backbone remain question-conditioned in all controls, so this comparison concerns the explicit question input to QGR. Appendix A.1 reports the full-test-set ablations; Appendix A.3 analyzes expert routing. The unadapted prototype raises visual-missing accuracy over the backbone (59.24 to 67.41), but lowers audio-missing accuracy (62.30 to 57.28). QGR improves both conditions, suggesting that retrieved priors benefit from task-specific adaptation.

Prototype granularity and storage efficiency. L1–L4 denote successive levels of the FINCH hierarchy (Sarfraz et al., 2019): higher levels merge clusters into coarser groups and retain fewer prototype pairs. Figure 2 analyzes granularity with the codec fixed and reports retained pairs relative to no clustering (100%). The validation-selected L2 setting retains 28.8% of pairs with $\mathrm{Avg}_2 = 72.54$, versus 72.77 without clustering; L4 retains only 0.8% but incurs a larger accuracy loss. Combining L2 with rank-4 channel compression and INT4 quantization gives an estimated $188\times$ reduction in prototype storage relative to unclustered full-channel FP16. This bit-count estimate assumes packed INT4 coefficients; Appendix D details storage accounting for simulated quantization.

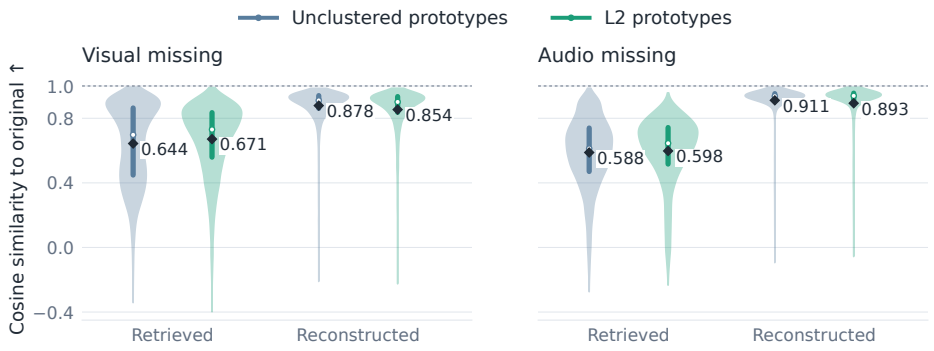

Figure 3: Cosine similarity to the original missing-modality representation before and after reconstruction. Unclustered prototypes and L2 prototypes are compared for visual-missing (left) and audio-missing (right) inputs. Violins show similarity distributions; annotated diamonds mark mean values.

From retrieved priors to reconstructed features. Figure 3 shows that reconstruction increases mean cosine similarity from 0.644 to 0.878 for visual features and from 0.588 to 0.911 for audio features with unclustered prototypes. With L2 prototypes, the increases are 0.671 to 0.854 and 0.598 to 0.893, leaving reconstruction-similarity gaps of 0.024 and 0.018. These small gaps suggest that compact prototypes retain useful reconstruction priors. The feature-space improvement is consistent with Table 2, where direct use of retrieved prototypes yields lower QA accuracy than QGR. Together, the analyses suggest that question-guided adaptation may help align retrieved priors with the complete-input feature space and support answering; cosine similarity alone does not isolate the effect of question conditioning.

5 Conclusion

We presented QRePair, a missing-modality AVQA framework that combines question-guided modality encoding, paired temporal prototype construction and retrieval, and question-guided reconstruction. On MUSIC-AVQA, QRePair improves visual- and audio-missing accuracy over the reported R^2ScP results by 2.22 and 1.43 percentage points, respectively. Ablations further show that disrupting audio-visual pairing reduces Avg_2 by 4.25 points, while removing explicit question conditioning from QGR reduces it by 8.96 points, supporting the roles of paired priors and question-specific adaptation. We hope this work can provide new insights into the design of robust multimodal reasoning systems under missing modalities.

AI USE STATEMENT

We used generative AI tools to improve the grammar, clarity, and overall flow of the manuscript. These tools assisted with refining sentence structure and wording, as well as improving transitions between sentences and paragraphs.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Qi Fan, Hongyu Yuan, Haolin Zuo, Rui Liu, and Guanglai Gao. Leveraging retrieval augment approach for multimodal emotion recognition under missing modalities. *arXiv preprint arXiv:2410.02804*, 2024. URL https://arxiv.org/abs/2410.02804.

Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22910–22921, 2023.

Hongyeob Kim, Inyoung Jung, Dayoon Suh, Youjia Zhang, Sangmin Lee, and Sungeun Hong. Question-aware Gaussian experts for audio-visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13681–13690, 2025. doi: 10.1109/CVPR52734.2025.01277.

Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M. Bakker, and Michael S. Lew. COCA: COllaborative CAusal regularization for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12995–13003, 2023. doi: 10.1609/aaai.v37i11.26527.

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*, pp. 19108–19118, 2022. doi: 10.1109/CVPR52688.2022.01852. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_Learning_To_Answer_Questions_in_Dynamic_Audio-Visual_Scenarios_CVPR_2022_paper.html.

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*, pp. 7808–7816, 2023. doi: 10.1145/3581783.3612293.

Guangyao Li, Henghui Du, and Di Hu. Boosting audio visual question answering via key semantic-aware cues. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 5997–6005, New York, NY, USA, 2024a. Association for Computing Machinery. doi: 10.1145/3664647.3680803.

Kun Li, Michael Ying Yang, and Sami Sebastian Brandt. Query-guided spatial–temporal–frequency interaction for music audio–visual question answering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8CnU2kchiw.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24068–24077, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_SimMLM_A_Simple_Framework_for_Multi-modal_Learning_with_Missing_Modality_ICCV_2025_paper.html.

Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3306–3314, 2024b. doi: 10.1609/aaai.v38i4.28116.

Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2299–2309, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00228.

Xiulong Liu, Zhikang Dong, and Peng Zhang. 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*, pp. 4478–4487, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Liu_Tackling_Data_Bias_in_MUSIC-AVQA_Crafting_a_Balanced_Dataset_for_WACV_2024_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. Look, listen, and answer: Overcoming biases for audio-visual question answering. In *Advances in Neural Information Processing Systems*, volume 37, pp. 9507–9531, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/123a18dfd821c8b440f42a00a27648d6-Abstract-Conference.html.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Kyu Ri Park, Hong Joo Lee, and Jung Uk Kim. Learning trimodal relation for audio-visual question answering with missing modality. In *Computer Vision – ECCV 2024*, pp. 42–59. Springer, 2024. doi: 10.1007/978-3-031-72633-0_3.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Sarfraz_Efficient_Parameter-Free_Clustering_Using_First_Neighbor_Relations_CVPR_2019_paper.html.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://arxiv.org/abs/1805.01978.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 438–446, 2024. doi: 10.1145/3664647.3681683.

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*, pp. 3480–3491, 2022. doi: 10.1145/3503161.3548291. URL https://mn.cs.tsinghua.edu.cn/avqa/.

Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. IMOL: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30921–30933, 2025. doi: 10.18653/v1/2025.acl-long.1494. URL https://aclanthology.org/2025.acl-long.1494/.

Jiayu Zhang, Shuo Ye, Qilang Ye, Zihan Song, Jiajian Huang, and Zitong Yu. Retrieving to recover: Towards incomplete audio-visual question answering via semantic-consistent purification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20985–20995, 2026. doi: 10.18653/v1/2026.acl-long.961. URL https://aclanthology.org/2026.acl-long.961/.

A Additional Experiments and Analyses

A.1 MUSIC-AVQA Ablations on the Full Test Set

Table A1 evaluates the same six configurations as Table 2 on all 9,129 MUSIC-AVQA test questions. Full QRePair achieves the highest Avg_2 of 75.91, consistent with the Audio-Visual subset results.

Table A1: Component ablations and accuracy (%) on all 9,129 MUSIC-AVQA test questions with missing inputs.

Ablation group	Setting	Prototype prior	Correct A–V pair	QGR	Question in QGR	V miss	A miss	Avg_2
No retrieved prior	Backbone only	✗	–	✗	–	60.12	64.55	62.33
	Direct reconstruction	✗	–	MLP	✓	66.08	75.24	70.66
Paired retrieval	Shuffled A–V pairs	✓	✗	✓	✓	65.91	75.28	70.59
Question-guided reconstruction	w/o QGR	✓	✓	✗	–	70.00	58.94	64.47
	w/o question conditioning	✓	✓	✓	✗	71.92	60.27	66.10
Full model	**QRePair**	✓	✓	✓	✓	**74.59**	**77.24**	**75.91**

Ablation groups, settings, and column definitions follow Table 2. Avg_2 is computed before rounding.

A.2 Results Across Random Seeds

Table A2 reports QRePair results for seeds 42, 43, and 44 on the 5,096 Audio-Visual MUSIC-AVQA test questions used in Table 1.

Table A2: Accuracy (%) across random seeds on the 5,096 Audio-Visual MUSIC-AVQA test questions.

Seed	Complete	V miss	A miss	Avg
42	73.92	71.59	73.49	73.00
43	73.53	71.08	73.29	72.63
44	73.37	70.37	72.55	72.10
Mean $\pm$ SD	73.61 ± 0.28	71.01 ± 0.61	73.11 ± 0.50	72.58 ± 0.45

Complete denotes both modalities available. $\mathrm{Avg} = (\mathrm{Complete} + \mathrm{V\ miss} + \mathrm{A\ miss})/3$. The final row reports the mean and sample standard deviation across the three seeds, computed before rounding.

A.3 Question-Guided Expert Routing

Figure A1 examines the temporal experts within QG-MoE. Span receives the largest weight for visual localization and audio-visual temporal questions, whereas Comparative is favored for audio counting and comparison. These patterns describe associations between question types and routing weights. They do not establish causal expert specialization, which would require interventions such as expert removal or routing permutation.

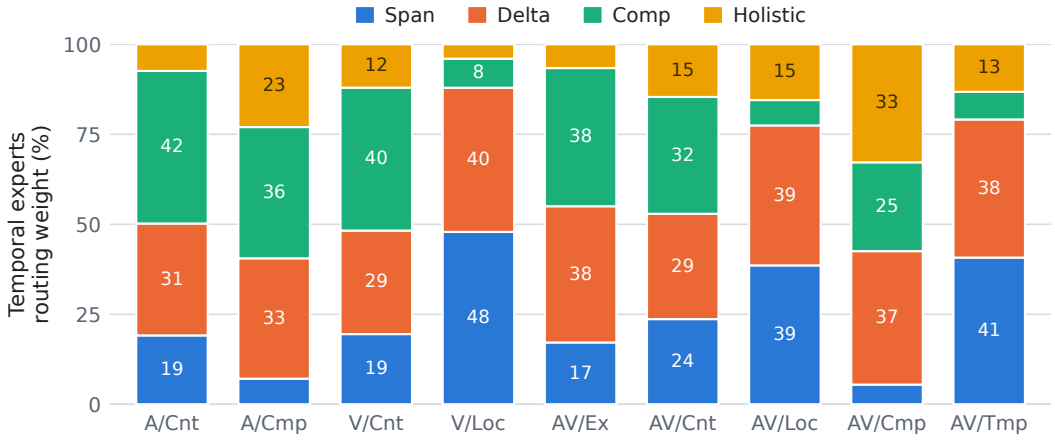

Figure A1: Mean routing weights of the temporal experts in QG-MoE by question type on the 9,129 MUSIC-AVQA test questions.

B Datasets and Evaluation Protocol

We use the official splits of MUSIC-AVQA (Li et al., 2022) and AVQA (Yang et al., 2022). Prototype construction and compression-basis estimation use training examples only. Backbone checkpoint selection uses validation data; for AVQA, this validation set is held out from the training split, independently of the official test set.

MUSIC-AVQA uses the updated annotations with 31,927 training, 4,568 validation, and 9,129 test questions. Tables 1, 2, and A2 and Figure 2 use the 5,096 Audio-Visual test questions; Tables A1 and F1 use all 9,129 test questions. The official AVQA test annotation is named `val_qa.json`. QRePair and the baselines use the same official test split and evaluation protocol.

For the standard visual-missing and audio-missing settings, we zero the input features of the unavailable modality while retaining the question. The model receives no missing-state label at inference; the state predictor selects the reconstruction direction, and its probability supplies the soft gate. Complete-input evaluation uses the same state predictor. All table values for other methods are retained from the cited publications.

C Detailed Backbone Architecture

C.1 Detailed Question-Conditioned Modality Encoding

This section specifies the input features and standard attention, residual, normalization, and feed-forward operations summarized in Section 3.1.

Input representation. Using a pretrained CLIP (Radford et al., 2021) encoder and Token Merging (ToMe) (Bolya et al., 2023), we obtain visual patch features $F_P \in \mathbb{R}^{T\times M'\times D}$ and frame-level features $F_V \in \mathbb{R}^{T\times D}$ by averaging the merged patches. A pretrained BEATs (Chen et al., 2023) encoder provides frequency-level audio features $F_{\text{Freq}} \in \mathbb{R}^{T\times K\times D}$, whose frequency average yields $F_A \in \mathbb{R}^{T\times D}$. The audio and visual features are aligned to T temporal steps. The CLIP text encoder produces word-level question features $F_W \in \mathbb{R}^{N\times D}$ and a sentence embedding $F_S \in \mathbb{R}^{D}$. Here, M' is the number of merged visual patches per time step, K the number of audio frequency tokens, and N the number of question tokens. All features are projected to a common dimension D; FFN denotes a position-wise feed-forward network.

Question-Guided Modality Attention. This hierarchical question-guided attention block progressively injects multi-granularity cues into the question-aligned audio-visual features.

We first apply SA to each modality feature to capture intra-modal dependencies:

$$\hat{F}_V = \mathrm{SA}(F_V), \quad \hat{F}_A = \mathrm{SA}(F_A), \quad \hat{F}_W = \mathrm{SA}(F_W). \tag{C.1}$$

The SA-refined $\hat{F}_V$ and $\hat{F}_A$ then serve as queries while $\hat{F}_W$ serves as key and value in a CA, aligning audio-visual features with the word-level semantics of the question:

$$F_V^q = \mathrm{LN}\Big(F_V + \mathrm{CA}(\hat{F}_V, \hat{F}_W, \hat{F}_W)\Big), \quad F_A^q = \mathrm{LN}\Big(F_A + \mathrm{CA}(\hat{F}_A, \hat{F}_W, \hat{F}_W)\Big). \tag{C.2}$$

Spatial and frequency attention further enrich these question-aligned features with fine-grained modality-specific context. We first apply SA with residual connections to the patch feature F_P and the frequency feature F_{Freq}:

$$\hat{F}_P = F_P + \mathrm{SA}(F_P), \qquad \hat{F}_{\text{Freq}} = F_{\text{Freq}} + \mathrm{SA}(F_{\text{Freq}}). \tag{C.3}$$

For the visual stream, $\hat{F}_P$ serves as key and value with F_V^q as query in a CA, supplementing the visual representation with patch-level spatial cues:

$$\tilde{F}_V = \mathrm{LN}\Big(F_V^q + \mathrm{CA}(F_V^q, \hat{F}_P, \hat{F}_P)\Big), \qquad \bar{F}_V = \tilde{F}_V + \mathrm{FFN}(\tilde{F}_V). \tag{C.4}$$

C.2 Detailed Gated Frequency Injection

Frequency averaging can attenuate spectral cues relevant to instruments or sound sources. GFI reintroduces frequency-level evidence under question conditioning: cross-attention selects relevant spectral information, while a scalar gate controls its overall contribution. The refined frequency features $\hat{F}_{\text{Freq}}$ are computed with the self-attention and residual operation above.

For the audio stream, F_A^q attends to $\hat{F}_{\text{Freq}}$ to capture fine-grained spectral information:

$$F_A^f = \mathrm{CA}(F_A^q, \hat{F}_{\text{Freq}}, \hat{F}_{\text{Freq}}). \tag{C.5}$$

The resulting F_A^f is integrated into the audio stream through Gated Frequency Injection.

Gated Frequency Injection. With σ denoting the sigmoid function, a scalar gate g_f conditioned on the sentence embedding F_S controls the contribution of frequency information to the audio representation:

$$g_f = \sigma(\mathrm{MLP}(F_S)), \qquad \tilde{F}_A = \mathrm{LN}\Big(F_A^q + g_f \cdot F_A^f\Big), \qquad \bar{F}_A = \tilde{F}_A + \mathrm{FFN}(\tilde{F}_A). \tag{C.6}$$

The resulting $\bar{F}_V$ and $\bar{F}_A$ retain modality-specific evidence for subsequent prototype retrieval and reconstruction.

C.3 State Prediction and Reconstruction Gating

Under controlled feature masking, the state predictor provides reconstruction routing and soft gating. It produces probabilities for complete, visual-missing, and audio-missing inputs from pooled encoded features:

$$\mathbf{p} = \mathrm{softmax}\big(\mathrm{MLP}_{\mathrm{state}}([\mathrm{Pool}(\bar{F}_V); \mathrm{Pool}(\bar{F}_A)])\big)\,. \tag{C.7}$$

Here, Pool aggregates each modality sequence over time. Let $s \in \{\text{complete}, V\text{-missing}, A\text{-missing}\}$ be the input state and $\hat{s} = \arg\max_s p(s)$ its prediction. Complete predictions pass both encoded streams directly to the reasoning module. Otherwise, only the predicted reconstruction direction is executed, with m denoting the missing modality and o the other modality.

Let $p(s = \text{V-missing})$ and $p(s = \text{A-missing})$ denote the corresponding entries of $\mathbf{p}$. With a scheduled scalar $\zeta \in [0, 1]$, we define a soft reconstruction gate for each missing state and apply the gate for the selected direction m:

$$\begin{aligned} g_V &= \zeta\, p(s = \text{V-missing}), \qquad g_A = \zeta\, p(s = \text{A-missing}), \\ F_m^+ &= (1 - g_m)\bar{F}_m + g_m C_m, \quad F_o^+ = \bar{F}_o. \end{aligned} \tag{C.8}$$

The resulting F_V^+ and F_A^+ are passed to QG-MoE. In Phase 1, PPR and QGR are bypassed and $F_m^+ = \bar{F}_m$. State-supervision and gate-schedule details are in Appendix E. When sensor availability is explicitly supplied by a system, that information could instead provide the state for reconstruction routing and gating.

C.4 Detailed QG-MoE Architecture

The reasoning backbone processes the two streams after the selected reconstruction branch. This section specifies the operations summarized in Section 3.3.

Audio-visual interaction. Gated bidirectional cross-attention relates the modality representations. The sentence embedding F_S, broadcast across the T time steps for concatenation, conditions the directional gates:

$$g_{v\to a} = \sigma\big(\mathrm{MLP}_{v\to a}([F_A^+; F_V^+; F_S])\big)\,, \qquad g_{a\to v} = \sigma\big(\mathrm{MLP}_{a\to v}([F_V^+; F_A^+; F_S])\big)\,. \tag{C.9}$$

The visual-to-audio branch updates the audio stream through gated residual fusion:

$$F_{av} = F_A^+ + g_{v\to a} \odot \mathrm{CA}(\mathrm{LN}(F_A^+), F_V^+, F_V^+), \qquad F_{av} \leftarrow F_{av} + \mathrm{FFN}_A(F_{av}). \tag{C.10}$$

Here, $\odot$ denotes element-wise multiplication. The reverse representation F_{va} is computed analogously with its own gate and parameters.

Temporal experts. For either directional sequence $X \in \{F_{av}, F_{va}\}$, we add a learnable temporal positional embedding $P \in \mathbb{R}^{T\times D}$, giving $\hat{X} = X + P$. Span, Delta, and Comparative use separate question-conditioned scorers with feature-wise modulation (Perez et al., 2018):

$$\psi(U, F_S)_t = W_{\mathrm{out}}\,\mathrm{GELU}(\mathrm{LN}(\gamma \odot W_{\mathrm{in}} U_t + \eta))\,, \qquad [\gamma, \eta] = W_{\mathrm{cond}} F_S. \tag{C.11}$$

For an input sequence U, U_t is its feature at time t and $\psi(U, F_S)_t$ is a scalar score. The learned projections $W_{\mathrm{in}}, W_{\mathrm{out}}, W_{\mathrm{cond}}$ map features to compatible dimensions; γ and η are question-conditioned channel-wise scale and offset vectors. GELU denotes the Gaussian error linear unit.

Span pools question-relevant regions, while **Delta** captures multi-scale temporal changes:

$$\begin{aligned} a^{\mathrm{span}} &= \mathrm{softmax}(\psi_{\mathrm{span}}(\hat{X}, F_S)), \quad e^{\mathrm{span}} = \sum_t a_t^{\mathrm{span}} \hat{X}_t, \\ a^{\Delta} &= \mathrm{softmax}(\psi_{\Delta}(D_{\Delta}, F_S)), \qquad e^{\Delta} = \sum_t a_t^{\Delta} D_{\Delta,t}. \end{aligned} \tag{C.12}$$

The softmax normalizes temporal weights a_t^{span} and a_t^{Δ} over $t = 1, \ldots, T$. Here, $D_{\Delta} = \phi_{\Delta}([\Delta_1 \hat{X}; \Delta_2 \hat{X}; \Delta_4 \hat{X}])$, where $\Delta_k \hat{X}$ denotes the k-step temporal difference with zero padding at the end.

Comparative forms two complementary summaries using $\alpha_t = \sigma(\psi_{\text{comp}}(\hat{X}, F_S)_t)$:

$$s_1 = \frac{\sum_t \alpha_t \hat{X}_t}{\sum_t \alpha_t + \epsilon}, \qquad s_2 = \frac{\sum_t (1-\alpha_t)\hat{X}_t}{\sum_t (1-\alpha_t) + \epsilon}, \tag{C.13}$$
$$e^{\text{comp}} = \phi_{\text{comp}}([s_1; s_2; s_1 - s_2; s_1 \odot s_2]),$$

where $\epsilon = 10^{-6}$. The **Holistic** expert uses the question as a query to aggregate the entire sequence:

$$e^{\text{hol}} = \text{CA}(\phi_h(F_S), \hat{X}, \hat{X}). \tag{C.14}$$

A question-guided router mixes all four outputs, indexed by $\mathcal{E} = \{\text{span}, \Delta, \text{comp}, \text{hol}\}$:

$$\rho = \text{softmax}(\text{MLP}_{\text{temp}}(F_S)), \qquad f_X = \sum_{j \in \mathcal{E}} \rho_j e^j. \tag{C.15}$$

The projections ϕ_Δ, ϕ_{comp}, and ϕ_h use linear–GELU, linear–GELU–linear, and linear mappings, respectively. Both sequences share expert parameters and routing weights, yielding f_{av} and f_{va}. The router depends only on the question, while each expert processes the directional feature sequence; all four experts are evaluated in this dense mixture.

Answer prediction. We concatenate f_{av} and f_{va} and apply a single MLP for answer prediction. On MUSIC-AVQA, the MLP predicts logits over the answer vocabulary. On AVQA, it produces a joint representation that scores the candidate answer embeddings by cosine similarity.

A load-balancing term $\mathcal{L}_{\text{lb}}$ encourages diverse expert use during backbone training.

Expert load balancing. For a minibatch of B questions and $E = 4$ experts, the load-balancing loss is

$$\mathcal{L}_{\text{lb}} = E \sum_{j=1}^{E} \bar{\rho}_j \nu_j, \qquad \bar{\rho}_j = \frac{1}{B} \sum_{b=1}^{B} \rho_{b,j}, \qquad \nu_j = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left[j = \arg\max_k \rho_{b,k} \right]. \tag{C.16}$$

Here, $\rho_{b,j}$ is the routing weight of expert j for question b, and $\mathbf{1}[\cdot]$ is the indicator function. The top-choice frequency ν_j is used only for this auxiliary loss; inference combines all four experts with soft routing weights.

D Prototype Construction and Storage Details

D.1 Clustering Descriptors and Paired Groups

Features use $T = 60$ temporal steps and $D = 512$ channels. Before clustering, we subtract each training representation's temporal mean, normalize its flattened centered component, project channels to rank 16, and apply simulated INT4 quantization followed by ℓ_2 normalization. FINCH level-2 clustering (Sarfraz et al., 2019) is applied separately to these audio and visual descriptors. Prototype sets at all hierarchy levels are constructed from training representations only. L2 was selected using validation-set accuracy and storage as a practical operating point and fixed before test evaluation. The test-set comparisons across L1–L4 in Figure 2 are reported only as a sensitivity analysis and were not used for model selection. The non-empty sample intersections in Eq. (3) define the paired groups. Within each group, both temporal prototypes are averaged over the same sample indices using the original, uncentered representations in Eq. (4). The resulting prototypes are then centered and compressed with a rank-4 channel basis and simulated INT4 quantization for retrieval and decoding, retaining the temporal mean and magnitude separately. Thus, rank 16 is used for clustering descriptors and rank 4 for the final prototype representation. Both projections use leading eigenvectors of modality-specific channel second-moment matrices computed from centered, normalized training representations.

Training-time retrieval candidates. For each Phase 2 training example i, the stored group assignment identifies the prototype pair constructed from a group containing i. We remove that pair's index from the candidate set before selecting the highest-similarity match in either reconstruction direction. The prototype values remain fixed; exclusion changes only the eligible retrieval candidates.

D.2 Prototype Representation and Storage Accounting

Each prototype is decomposed into its temporal mean and centered component. We encode the centered component with a modality-specific low-rank channel basis and quantization, retaining its magnitude and temporal mean for decoding. Compression acts on the channel dimension so that each paired prototype retains an ordered temporal sequence for reconstruction. Averaging sequences preserves temporal indices but does not explicitly align event onsets across clips. The prototypes remain fixed during Phase 2. The descriptor and rank settings above specify the two compression stages.

For Figure 2, each clustering setting follows the same Phase 2 training protocol with the representation codec fixed. Storage ratios are calculated for the encoded prototype set of both modalities, including temporal means, magnitudes, and quantization scales. The FP16 reference uses the full channel dimension with the same mean and magnitude terms. The INT4 estimate assumes packed coefficients and excludes shared bases and runtime buffers; evaluation uses simulated quantization. Thus, the reported ratio describes the calculated storage required by the encoded prototype representation, rather than an empirical peak-memory measurement.

E Training Details

We use AdamW (Loshchilov & Hutter, 2019) and select the backbone checkpoint with the highest validation accuracy within the first 15 epochs. We then train QGR and the state predictor for 15 epochs with the backbone and prototypes fixed, and evaluate the final Phase 2 checkpoint. Following the reporting convention of prior missing-modality AVQA studies, QRePair results use a fixed random seed (seed 42), except for the multi-seed results in Appendix A.2. Table E1 lists the training settings. Both phases use FP16 mixed precision without learning-rate warmup. Phase 2 uses a distillation temperature of one and no answer-label classification loss. The fully observed comparisons in Table F1 and Appendix F.2 evaluate the complete QRePair model with both modalities available.

Table E1: Training settings. MUSIC denotes both MUSIC-AVQA and MUSIC-AVQA 2.0.

Setting	Phase 1	Phase 2
AdamW (β_1, β_2)	$(0.9, 0.999)$	$(0.9, 0.999)$
Initial learning rate	2×10^{-4}	10^{-3}
Weight decay	0.02	0
Learning-rate schedule	Step decay	Cosine decay
Schedule parameters	Every 10 epochs, factor 0.1	Minimum 10^{-6}
Batch size (MUSIC / AVQA)	$128/128$	$128/64$
QA label smoothing	0.1	–
Loss weights	$\lambda_{\mathrm{lb}} = 0.02$	$\lambda_{\mathrm{rec}} = \lambda_{\mathrm{KD}} = \lambda_{\mathrm{state}} = 1$

State supervision and reconstruction-gate schedule. The state-prediction loss averages cross-entropy over complete, visual-missing, and audio-missing inputs. During Phase 2, the reconstruction-gate coefficient is $\zeta(r) = \min(1, r/250)$, where r counts completed optimizer steps; $\zeta = 1$ at inference.

F Full-Modality Results

F.1 MUSIC-AVQA

Table F1 compares full-modality performance with task-specific baselines and zero-shot omni models. QRePair obtains 78.07% overall accuracy, compared with 77.62% for QA-TIGER and 78.98% for QSTar. These results complement the primary missing-modality comparison in Table 1.

Table F1: Full-modality accuracy (%) on MUSIC-AVQA.

Method	Audio QA			Visual QA			Audio-Visual QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
Zero-shot omni models													
GPT-4o	65.68	37.04	55.12	72.77	62.20	67.42	55.87	54.94	59.57	38.24	42.58	50.35	55.72
Qwen2.5-Omni (7B)	62.93	39.56	54.31	67.92	59.02	63.42	51.92	55.10	60.98	36.42	40.27	49.12	53.83
Ming-Lite-Omni	62.44	40.74	54.44	67.50	58.29	62.84	52.73	53.91	62.17	37.24	41.85	49.63	53.98
Task-specific models													
ST-AVQA	78.18	67.05	74.06	71.56	76.38	74.00	81.81	70.80	64.51	66.01	63.23	69.54	71.52
COCA	79.35	67.68	75.42	75.10	75.43	75.23	83.50	66.63	69.72	64.12	65.57	69.96	72.33
PSTP-Net	73.97	65.59	70.91	77.15	77.36	77.26	76.18	72.23	71.80	**71.79**	69.00	72.57	73.52
LAVISH	82.09	65.56	75.97	78.98	81.43	80.22	81.71	75.51	66.13	63.77	67.96	71.26	74.46
APL	82.40	70.71	78.09	76.52	82.74	79.69	82.99	73.29	66.68	64.76	65.95	70.96	74.53
TSPM	84.07	64.65	76.91	82.29	84.90	83.61	82.19	76.21	71.85	65.76	71.17	73.51	76.79
QA-TIGER	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
QSTar	85.64	**72.05**	**80.63**	83.46	84.90	84.17	**83.81**	**79.76**	**72.72**	70.03	**72.38**	**75.98**	**78.98**
QRePair	**86.53**	69.87	80.38	**84.38**	86.12	**85.26**	82.79	78.50	71.85	64.58	71.05	73.92	78.07

Avg: QA-count-weighted accuracy. Zero-shot omni results: Li et al. (2026), Table 1. QSTar: Li et al. (2026), Table 1. Other task-specific baseline values: Kim et al. (2025). **Bold**/underline: highest/second-highest reported accuracies among task-specific models. Additional references: ST-AVQA (Li et al., 2022), COCA (Lao et al., 2023), and LAVISH (Lin et al., 2023).

F.2 Fully Observed Distribution Shifts

We evaluate QRePair with both audio and visual inputs available.

Head and tail evaluation. We further evaluate the MUSIC-AVQA-trained model on MUSIC-AVQA-R without additional training. As shown in Table F2, QRePair obtains an overall accuracy of 70.89, 2.90 points above the reported QA-TIGER value. On tail audio-visual counting and localization questions, accuracy increases from 33.55 to 37.02 and from 64.65 to 76.60, respectively. These results indicate improved aggregate performance under the benchmark's question and answer distribution shifts.

Cross-distribution evaluation. Table F3 evaluates QRePair on MUSIC-AVQA 2.0 under Bias and Balance. QRePair obtains the highest reported overall accuracy among the compared methods in all four train–test combinations, with differences of 1.12–2.03 points relative to QA-TIGER. When trained on Balance, it obtains 77.92 on Bias and 77.55 on Balance, exceeding QA-TIGER by 1.35 and 1.12 points, respectively.

Table F2: Full-modality accuracy (%) on MUSIC-AVQA-R.

Method	Audio QA				Visual QA				Audio-Visual QA										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	
FCNLSTM	66.23	36.48	64.78	51.24	61.75	5.31	54.86	51.06	64.76	78.52	62.69	7.23	46.66	57.30	43.13	71.67	37.02	30.78	54.12
BiLSTM	73.68	46.32	21.51	**77.58**	64.30	0.00	53.92	42.01	**87.51**	21.14	62.85	2.18	35.16	43.75	27.61	74.38	17.58	31.32	48.84
HCAttn	61.67	41.63	59.09	47.14	56.52	9.20	67.01	53.16	66.57	61.13	59.53	12.48	37.05	42.48	48.81	60.12	33.82	39.26	51.90
MCAN	75.02	60.16	58.89	50.09	64.58	26.69	66.48	62.25	51.29	67.29	64.76	25.28	46.11	61.61	50.57	52.40	34.64	58.05	57.27
PSAC	53.01	56.68	57.41	48.12	49.55	26.43	72.96	60.69	50.56	55.54	56.70	19.58	41.98	52.30	38.13	58.92	26.68	46.24	50.45
HME	62.60	53.95	54.97	58.29	50.95	16.46	73.25	58.60	65.74	66.49	63.18	17.18	33.79	46.03	53.20	69.57	33.95	41.57	53.66
HCRN	55.53	53.31	47.17	32.44	41.87	23.55	39.40	51.27	41.81	65.45	54.58	19.57	36.62	42.72	33.33	36.87	40.47	44.13	43.92
AVSD	54.00	47.84	60.61	47.79	60.34	10.07	74.78	61.43	66.28	61.98	46.21	8.06	33.00	40.35	51.98	66.00	40.14	41.52	52.33
Pano-AVQA	50.57	43.45	50.78	44.93	47.28	15.50	67.19	65.51	52.37	22.04	52.21	21.52	44.35	61.69	45.61	40.49	35.00	49.33	47.40
ST-AVQA	56.40	41.48	62.28	57.59	59.86	12.94	63.31	54.00	73.35	77.26	48.31	8.41	35.35	40.49	53.30	62.44	40.25	38.15	52.80
LAVISH	61.73	43.99	65.06	60.38	65.53	11.13	70.21	64.73	77.83	79.46	49.88	14.87	41.76	41.20	**59.26**	65.10	41.84	46.26	57.63
TSPM†	81.65	71.80	67.66	49.56	78.29	47.56	80.58	73.18	69.15	**82.79**	77.09	**38.64**	42.24	57.37	52.07	68.86	39.23	49.36	66.30
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
QRePair	**85.98**	72.31	62.56	62.45	**84.55**	**57.33**	**88.61**	**79.26**	68.21	77.82	**82.18**	37.02	**54.56**	**76.60**	48.40	70.65	**47.34**	58.43	**70.89**

H/T: head/tail. Avg: QA-count-weighted accuracy. Baseline values: Kim et al. (2025). **Bold**/underline: highest/second-highest reported accuracies. †: reproduced with official code in the source paper. QRePair is trained on MUSIC-AVQA v1.

Table F3: Full-modality accuracy (%) on MUSIC-AVQA 2.0.

(a) Test on Bias split

Train	Method	A-QA	V-QA	AV-QA	Avg
Bias	ST-AVQA	76.86	77.70	69.59	73.07
	LAVISH	76.73	80.96	70.80	74.59
	QA-TIGER	**79.13**	84.83	72.37	76.93
	QRePair	78.99	**87.14**	**74.43**	**78.66**
Balance	ST-AVQA	76.18	77.20	67.96	71.92
	LAVISH	75.56	80.83	69.27	73.51
	LAST	77.10	82.99	70.86	75.24
	LAST-Att	77.29	83.47	71.05	75.45
	QA-TIGER	77.07	85.93	71.20	76.57
	QRePair	**79.73**	**87.34**	**72.76**	**77.92**

(b) Test on Balance split

Train	Method	A-QA	V-QA	AV-QA	Avg
Bias	ST-AVQA	73.34	76.82	64.51	69.40
	LAVISH	73.14	79.70	65.01	70.39
	QA-TIGER	**77.57**	84.84	67.43	73.91
	QRePair	77.51	**86.69**	**70.26**	**75.94**
Balance	ST-AVQA	75.50	77.67	66.32	71.02
	LAVISH	76.15	81.32	68.28	73.18
	LAST	78.08	83.29	69.72	74.85
	LAST-Att	78.56	84.07	70.30	75.44
	QA-TIGER	79.90	86.95	70.22	76.43
	QRePair	**80.80**	**87.43**	**71.73**	**77.55**

Avg: QA-count-weighted accuracy. Baseline values: Kim et al. (2025). **Bold**/underline: highest/second-highest reported accuracies within each test/training block. LAST and LAST-Att follow Liu et al. (2024).