# SCENEROUTE-AVQA: SCENE-CONDITIONED TRUSTWORTHY FUSION FOR AUDIO-VISUAL QUESTION ANSWERING UNDER MULTIMODAL DEGRADATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Audio-Visual Question Answering (AVQA) requires selecting question-relevant evidence from audio and visual streams. However, real-world inputs often suffer from temporal misalignment, cross-video interference, missing modalities, visual occlusion, and audio-visual noise. These degradations alter modality reliability in a sample- and question-dependent manner, making conventional fixed-weight or unified fusion methods vulnerable to unreliable evidence and over-confident predictions.We propose SceneRoute-AVQA, a scene-conditioned robust framework based on a diagnosis-routing-calibration paradigm. First, scene-conditioned degradation diagnosis uses scene cues, temporal offset estimation, and modality quality assessment to characterize degradation sources, temporal inconsistencies, and modality reliability cues. Second, question-aware reliability fusion jointly considers question semantics and sample-specific modality quality to dynamically select and fuse trustworthy audio-visual evidence. Hierarchical question-scene answer priors are further introduced to mitigate low-resource and imbalanced answer distributions in degraded scenes. Third, robust context and answer calibration exploit scene-consistent contextual information and lightweight logit residuals to stabilize decision boundaries on difficult samples.Experiments under clean conditions and five degraded AVQA settings demonstrate that SceneRoute-AVQA consistently outperforms existing methods across temporal misalignment, cross-video interference, missing modalities, occlusion, and audio-visual noise, confirming its effectiveness in degradation diagnosis, adaptive evidence fusion, and robust answer prediction.

## 1 INTRODUCTION

AVQA is not simply the concatenation of video, audio, and question features; it must determine which evidence is trustworthy for each question. For example, “How many instruments are sounding?” relies on agreement between audio and visual counting cues; “What is the person doing?” relies more on local actions; and “What is producing the sound?” requires temporally aligned audio events and visual objects. Evidence selection becomes harder under degradation: misalignment breaks synchrony, cross-video interference introduces semantically irrelevant objects, missing modalities make evidence incomplete, occlusion weakens local visual actions, and noise reduces acoustic discriminability.

As illustrated in Figure 1, clean inputs provide a consistent audio–visual–question evidence chain. When temporal misalignment, cross-video interference, missing modalities, local occlusion, or acoustic noise occurs, fixed fusion treats distorted and reliable cues equally and consequently makes incorrect predictions. SceneRoute-AVQA first diagnoses each sample’s condition together with temporal, quality, and visibility evidence; it then selects reliable modalities conditioned on question semantics and conservatively calibrates answer boundaries for difficult cases.

Recent AVQA and audio-visual large-model research has moved from joint encoding toward explicit cooperation and temporal awareness. Mirasol3B Piergiovanni et al. (2024), AVicuna Lin et al. (2024), Video-LLaMA 2 Cheng et al. (2024), LLoVi Zhang et al. (2024), Dolphin Guo et al. (2025), Crab Du et al. (2025), and Daily-Omni Zhou et al. (2025) demonstrate the value of question-



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.

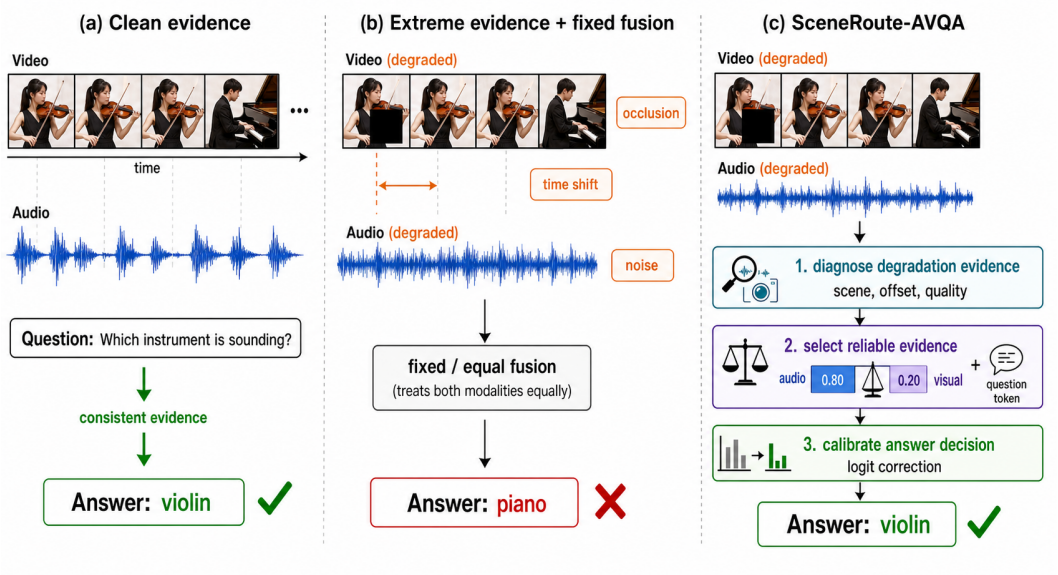



Figure 1: Audio, visual, and question evidence agree for clean samples. Under temporal misalignment, occlusion, and noise, fixed or equal-weight fusion can be dominated by erroneous evidence. SceneRoute-AVQA diagnoses degradation, selects question-relevant reliable evidence, and calibrates decision boundaries for difficult examples.

conditioned spatiotemporal audio-visual interaction. Question-Aware Gaussian Experts Zhang et al. (2025a) places question context before audio and visual encoding to select continuous, question-relevant intervals. AV-Master Zhang et al. (2025b) models each modality's contribution to a question, while Instruction-Aware Gating Ding et al. (2026) reweights modalities using textual instructions. Missing modalities, joint corruption, uncertainty modeling, and test-time robustness are likewise central issues in multimodal learning Li et al. (2024); Chen et al. (2024); Kim et al. (2025); Ramazanova et al. (2025); Lai et al. (2025). We argue that robust audio-visual reasoning must decide which modalities to trust under the given question and input quality, how to compensate for missing evidence, and how to avoid overfitting calibration in few-shot conditions. Three challenges remain:

- **Heterogeneous condition states:** each degradation affects a different evidence dimension: alignment, completeness, local visibility, or acoustic frequency and energy structure.
- **Question-dependent modality trust:** an audio or visual quality score alone cannot determine fusion; the model must account for what evidence the question requires.
- **Limited extreme-condition data:** enlarging fusion heads or fine-tuning end-to-end can overfit a small validation set; calibration should be local, conservative, and reproducible.

We therefore ask: **(1)** How can an interpretable degradation-evidence representation identify a sample's temporal, quality, and uncertainty state in mixed conditions? **(2)** How can multimodal evidence be fused stably and answer boundaries calibrated under few-shot extreme conditions? We address these questions with SceneRoute-AVQA, a scene-conditioned trustworthy reasoning framework for clean data and five degradation types. It follows the sequence "degradation diagnosis–trustworthy evidence selection–answer calibration." First, cross-video conflict suppression, acoustic reliability estimation with adaptive denoising, explicit temporal offset search, occlusion-robust visual adaptation with local visibility estimation, and modality-quality estimation jointly form condition-aware degradation evidence. Next, a question-aware trustworthy fusion module learns sample-level weights from audio, visual, and question uncertainty and uses hierarchical question–scene answer priors to reduce few-shot answer bias. Finally, Scene-Consistent Robust Context (SCRC) and a lightweight Hard-Scene Answer Delta conservatively correct difficult-scene logit boundaries. Our contributions are:

- We propose SceneRoute-AVQA, a unified scene-conditioned AVQA pipeline that connects degradation diagnosis, question-relevant evidence selection, and answer calibration.
- We introduce condition-aware degradation evidence modeling and question-aware trustworthy fusion, leveraging temporal offsets, cross-modal conflict, modality quality, local visibility, and acoustic reliability.
- We incorporate hierarchical question–scene answer priors, SCRC, and lightweight answer-residual correction to mitigate answer bias and unstable decision boundaries in few-shot extreme conditions.
- We systematically evaluate on Extreme-MUSIC-AVQA, an extension of MUSIC-AVQA, and obtain state-of-the-art performance on this benchmark.

## 2 Related Work

**Audio-Visual Question Answering and Cross-Modal Reasoning.** AVQA jointly reasons over video, audio, and language. Recent work advances temporally sensitive audio-visual understanding through time-aligned autoregressive modeling Piergiovanni et al. (2024), interleaving and context-boundary alignment Lin et al. (2024), and video-language models or benchmarks for long audio-visual understanding Cheng et al. (2024); Xie et al. (2024); Geng et al. (2025); Zhou et al. (2025); Shu et al. (2025). Other methods improve cross-modal reasoning through multi-scale adaptation, explicit cooperation, dual-path perception, or question-conditioned temporal experts Guo et al. (2025); Du et al. (2025); Zhang et al. (2025b;a). In contrast to methods focused on clean or generic video understanding, we study diagnostic reliable evidence and stable decisions when several degradations coexist.

**Robust Multimodal Learning and Degradation Awareness.** Missing-modality methods improve robustness through relation modeling, probabilistic alignment, test-time adaptation, or feature completion Li et al. (2024); Chen et al. (2024); Ramazanova et al. (2025); Chochlakis et al. (2024). For noise and occlusion, corrupted-prediction learning, uncertainty weighting, audio-source localization, and spatial-temporal disentanglement emphasize reliable local evidence and temporal consistency Kim et al. (2025); Lai et al. (2025); Ryu et al. (2025); Bae et al. (2025); Wang et al. (2024). We adapt these observations to a condition-aware AVQA process: quality estimation and question-guided compensation handle missing modalities, explicit offset search models misalignment, and condition-specific adapters suppress cross-video conflict and unreliable visual or acoustic tokens.

**Uncertainty, Trustworthy Fusion, and Few-Shot Calibration.** Fixed fusion can amplify erroneous evidence when input quality is poor. Uncertainty-weighted learning Lai et al. (2025); Jiang et al. (2025), instruction-aware gating Ding et al. (2026), and temporal-hallucination diagnostics Li et al. (2025); Liu et al. (2025) support adapting information flow to sample conditions. Since AVQA question templates and answers are long tailed, end-to-end classifiers may have biased boundaries in few-shot extreme settings. We inject stable statistical knowledge into answer logits using hierarchical question–scene priors and a lightweight answer residual, together with question-aware trustworthy fusion.

**Positioning of SceneRoute-AVQA.** Unlike methods centered on clean inputs or one failure mode, SceneRoute-AVQA treats condition as a first-class variable and couples degradation diagnosis with question-dependent trust and conservative answer calibration.

## 3 Method

An overview of the proposed SceneRoute-AVQA framework is presented in Figure 2. Given video, audio, and question features $v$, $a$, and $q$, the model predicts scene probabilities $\pi \in \mathbb{R}^6$ and produces $p(y \mid a, v, q)$ through scene-conditioned experts and trustworthy fusion. The objective is to jointly model degradation state, question-relevant evidence, and answer calibration.

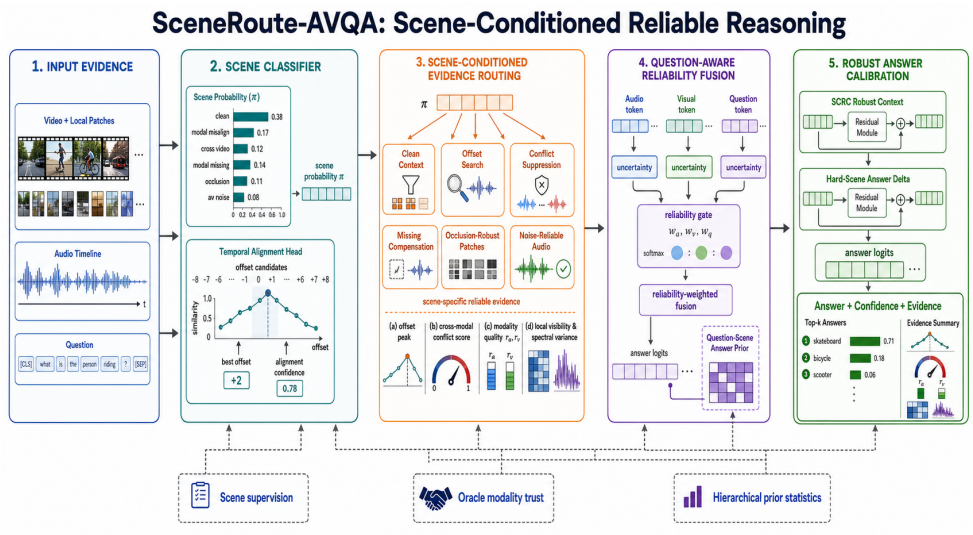



Figure 2: Overview of SceneRoute-AVQA: multi-source encoding, degradation diagnosis and routing, question-aware trustworthy fusion, and SCRC/residual calibration. Dashed branches are training-only.

### 3.1 Condition-Aware Degradation Evidence Modeling

We first construct condition-aware evidence representations. The scene classifier outputs $\pi$ for clean data and five extreme conditions, providing routing and diagnostic signals. Rather than sharing a single fusion strategy, condition adapters extract evidence specific to the degradation source: offset direction and alignment confidence for temporal misalignment; question-conditioned semantic conflict for cross-video interference; quality and compensation for missing modalities; local visibility and action evidence for occlusion; and acoustic energy, frequency dynamics, and denoised reliability for noise.

**Explicit Temporal Offset Search.** For potentially asynchronous samples, we retain time sequences and compute cross-modal similarity across candidate offsets $k \in [-8, +8]$:

$$s_k = \frac{1}{T_k} \sum_t \cos\left(W_a a_t, W_v v_{t+k}\right). \tag{1}$$

The best offset $k^* = \arg\max_k s_k$ and its peak confidence become alignment evidence for subsequent modules. This branch identifies not only whether a sample is misaligned but also the direction, magnitude, and confidence of the mismatch.

**Modality Quality Estimation and Compensation.** When audio or visual evidence is missing or severely corrupted, the model estimates modality quality $r_a, r_v \in [0, 1]$ and constructs a compensatory representation from available modalities and question semantics:

$$\tilde{h} = f_{\text{comp}}(a, v, q, r_a, r_v). \tag{2}$$

The question determines what must be compensated; for instance, counting favors sound–object agreement whereas action questions favor local visual tokens.

**Cross-Video Conflict Suppression.** Cross-video interference need not imply low quality; instead, audio events and visual content can be conditionally irrelevant to one another for the question. We construct a conflict score using question-projected modality representations:

$$c_{cv} = 1 - \cos\left(g_a(a, q), g_v(v, q)\right), \tag{3}$$

where $g_a$ and $g_v$ are question-conditioned projections. A high $c_{cv}$ activates a cross-video adapter that suppresses cross-modal interaction residuals and shifts attention toward tokens more consistent with the question, rather than fusing all evidence equally.

**Occlusion-Robust Visual Adaptation and Local Visibility.** Occlusion does not necessarily invalidate an entire visual sequence; it can cover the question-relevant object, hand action, or local region. We retain ToMe patch tokens $P = \{p_i\}$ and estimate local visibility from token dispersion, question-attention entropy, and global–local consistency:

$$r_v^{\mathrm{occ}} = \sigma\left(f_{\mathrm{occ}}([\mathrm{Var}(P), H(\alpha_{q\to P}), \cos(\bar{P}, v)])\right). \tag{4}$$

Here $\alpha_{q\to P}$ is question-to-patch attention. The visual adapter repairs only low-visibility local tokens through residual updates; $r_v^{\mathrm{occ}}$ controls their contribution to fusion, preventing accidental occluder texture from becoming deterministic evidence.

**Acoustic Reliability and Adaptive Denoising.** For the `av_noise` condition, we derive acoustic reliability $r_a^{\mathrm{noi}}$ from short-time energy statistics, frequency-dynamics proxies, and local variance of VGGish features. The noise adapter combines residual denoising, local-variance estimation, and intensity-adaptive mixing:

$$\begin{aligned} a^{\mathrm{noi}} &= r_a^{\mathrm{noi}} a + (1 - r_a^{\mathrm{noi}}) f_{\mathrm{denoise}}(a), \\ r_a^{\mathrm{noi}} &= \sigma(f_{\mathrm{noi}}([\mu(a), \mathrm{Var}(a), \Delta a])). \end{aligned} \tag{5}$$

$\Delta a$ aggregates the intensity of feature changes between adjacent time steps. Lower reliability increases the denoised representation and downweights unreliable acoustic tokens, while the residual path preserves details when stable event cues remain.

### 3.2 Question-Aware Trustworthy Fusion and Few-Shot Answer Priors

Scene evidence alone is insufficient because questions depend on modalities differently. Given audio, visual, and question representations $z_a, z_v, z_q$, an uncertainty estimator outputs $u_a, u_v, u_q$, which are converted by reliability gating:

$$g_m = \sigma(-u_m), \quad m \in \{a, v, q\}. \tag{6}$$

Fusion weights are then obtained by softmax:

$$w_m = \frac{\exp(g_m)}{\sum_{j\in\{a,v,q\}} \exp(g_j)}, \qquad z = \sum_m w_m z_m. \tag{7}$$

This mechanism learns which modality to trust for each sample instead of applying fixed or average fusion. At the answer level, hierarchical question–scene priors alleviate few-shot answer bias. Let $P_{qs}(y)$ denote answer statistics conditioned on question type and scene, and let $P_q(y)$ be question-type statistics. The calibrated logit is

$$z'_y = z_y + \lambda_1 \log(P_q(y) + \epsilon) + \lambda_2 \log(P_{qs}(y) + \epsilon). \tag{8}$$

The prior is estimated only from training and validation support sets; hyperparameters are selected on validation data.

### 3.3 Scene-Consistent Robust Context and Hard-Scene Answer Residual

Errors in extreme conditions often arise from unstable decision boundaries: the model may extract partially correct evidence, while its final prediction remains sensitive to noisy scene estimates, modality degradation, or temporal misalignment. To improve prediction stability, we introduce Scene-Consistent Robust Context (SCRC), a multi-view logit ensemble for difficult samples.

Given the same audio-visual input and question, SCRC evaluates the shared AVQA model under $K$ predefined scene-consistent context views $\{c^{(k)}\}_{k=1}^{K}$. Each view provides a distinct but plausible condition setting for scene-aware inference. The shared model produces an answer logit for each view:

$$z^{(k)} = f_\theta(a, v, q; c^{(k)}), \qquad k = 1, \ldots, K, \tag{9}$$

where all views reuse the same backbone, fusion module, and answer head parameters $\theta$. The SCRC base logit is obtained by aggregating the view-specific predictions:

$$z_{\mathrm{base}} = \sum_{k=1}^{K} \alpha_k z^{(k)}, \qquad \sum_{k=1}^{K} \alpha_k = 1. \tag{10}$$

The ensemble weights $\{\alpha_k\}_{k=1}^{K}$ are fixed using the validation set (or set uniformly when no view-specific preference is observed).

A single scene-conditioned inference view can be unreliable when scene prediction or modality-quality estimation is imperfect. By aggregating complementary context-conditioned logits, SCRC reduces prediction variance and makes the final decision boundary less sensitive to an erroneous individual context view. Importantly, SCRC introduces no additional backbone capacity: it reuses the same parameters across all views and only adds a lightweight multi-view inference and logit aggregation procedure.

We further introduce Hard-Scene Answer Delta. Given the SCRC base logit $z_{\text{base}}$, the model learns a lightweight scene-conditioned residual:

$$z = z_{\text{base}} + \alpha \sum_{s} \pi_s \, \gamma_s(h, q, r) \, \Delta_s(h, q, r), \tag{11}$$

where $\Delta_s$ is a lightweight answer-residual expert, $\gamma_s$ is the gate strength, $r$ denotes reliability and degradation evidence, and $\alpha$ is initialized to a small value. This local correction addresses residual decision bias in hard scenes while avoiding the overfitting risk of re-learning the full model.

# 4 Experiments

## 4.1 Data and Settings

**Extreme-MUSIC-AVQA.** We construct Extreme-MUSIC-AVQA from MUSIC-AVQA to evaluate AVQA under clean and extreme inputs. The dataset has 10,675 samples from six conditions: 2,935 clean samples and 1,548 samples each of modal_misalign, cross_video, modal_missing, occlusion, and av_noise. To broaden content coverage for the clean condition, we add and annotate approximately 1,500 YouTube audio-visual QA examples. The five remaining conditions apply temporal shifts, cross-video replacement, modality removal, visual occlusion, and audio-visual noise, respectively, to original MUSIC-AVQA samples. Thus, scenes have similar QA semantics but distinct evidence degradations.

We report final results on 2,055 test samples spanning all six conditions. Inputs are frame-level CLIP/ViT features, ToMe local-patch features, VGGish audio features, and CLIP textual-question features. Model selection, context strength, and prior weights are determined on validation data; the test set is used only for final reporting.

**Data Integrity and Evaluation Protocol.** All degradation transformations are applied after the clean source sample and its QA semantics are identified, so the question and answer remain tied to the same underlying event while the available evidence changes. We keep source-video identities separated across train, validation, and test whenever the source metadata permit this check; statistics used by the hierarchical priors are estimated without test annotations. Cross-video replacement draws an incompatible visual stream only within the designated split, and temporal offsets, occlusion masks, noise levels, and missing-modality flags are generated without reusing test features during tuning. This protocol separates two questions that are often conflated in robust AVQA: whether the model recognizes the intended answer semantics and whether it selects reliable evidence after the input has been degraded.

## 4.2 Implementation Details

Following the common AVQA experimental protocol, answer prediction is formulated as open-ended classification over a fixed answer vocabulary. Each video is represented by cached frame-level CLIP/ViT features and ToMe local-patch features; audio is represented by 128-D VGGish features, with a documented deterministic spectral fallback used only when the TensorFlow VGGish extractor is unavailable. Questions are encoded with sentence-level CLIP textual features. The answer classifier outputs logits over the candidate vocabulary and predicts $\hat{y} = \arg\max_y p(y \mid a, v, q)$. Training uses answer cross-entropy as the main loss; when scene labels are available, the degradation-diagnosis branch is supervised with an auxiliary scene-classification loss. All baseline and SceneRoute-AVQA variants use the same train/validation/test split, cached features, answer vocabulary, and seed-42 protocol to avoid feature-extractor or data-split confounds.

The reported SceneRoute-AVQA model uses a 512-D hidden representation, six scene classes, eight attention heads, two Transformer layers, dropout 0.3, adapter hidden dimension 128, and 42 answer classes. Training is divided into Stage-A/Stage-B schedules of 30 and 20 epochs, respectively. We use AdamW with learning rate $3\times10^{-4}$, weight decay $1\times10^{-4}$, batch size 8 for the reported cached-feature snapshot, and early stopping on validation accuracy with patience 10. The locked result fixes random seed 42 for Python, NumPy, PyTorch, and CUDA. The experimental environment is an NVIDIA GeForce RTX 5070 GPU with 8151 MiB GPU memory, 64GB system memory, Windows 10/11 build 10.0.26200, Python 3.10.20, PyTorch 2.11.0+cu128.

### 4.3 Evaluation Metrics

We use answer accuracy as the primary metric because AVQA requires selecting the correct answer from the candidate answer vocabulary:

$$\mathrm{Acc} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \mathbb{1}[\hat{y}_i = y_i]. \tag{12}$$

In addition to overall accuracy, we report per-condition accuracy for clean, modal_misalign, cross_video, modal_missing, occlusion, and av_noise. We also report macro-scene average accuracy, computed as the mean of the six per-condition accuracies, because the benchmark intentionally contains more clean samples than each individual degraded condition. For robustness analysis, this macro score prevents improvements on the larger clean subset from hiding failures on missing-modality or occlusion cases.

**Reporting and Analysis Protocol.** We report overall accuracy together with the six condition-wise values, macro-scene average, feature-aligned baseline comparisons, and module-level ablations. Hyperparameters are selected by validation overall accuracy, with macro-scene accuracy used only as a tie breaker. The final reported test result is a locked seed-42 configuration; therefore, it should be interpreted as a reproducible snapshot rather than a claim of a multi-seed statistical test. We additionally record negative results to identify calibration or adaptation choices that fail to transfer from validation data to the held-out mixed-condition test set.

### 4.4 Baselines and Fair Comparison Protocol

We compare SceneRoute-AVQA with representative AVQA baselines and robust-fusion variants under a feature-aligned protocol. PSTP, TSPM, QA-TIGER, AV-Master, and our ablated variants are evaluated on the same Extreme-MUSIC-AVQA split whenever compatible. To make the comparison fair, we keep the input features, answer vocabulary, scene labels, batch construction, and seed fixed across reimplementations. For methods whose public implementation assumes a different encoder or unavailable feature format, we either provide a feature-compatible reimplementation or explicitly mark the result as not directly comparable. Oracle scene routing is reported only as a diagnostic upper bound and is not used as the deployable result.

**Research Questions.** Our experiments address four practical questions: (RQ1) does condition-aware diagnosis improve overall AVQA accuracy under mixed clean and degraded inputs; (RQ2) does the proposed mechanism improve macro-scene performance rather than only the larger clean subset; (RQ3) which degradation types benefit from routing, trustworthy fusion, and calibration; and (RQ4) do the gains remain after input features, answer vocabulary, data split, and random seed are controlled? Tables 1 and 2 answer RQ1–RQ3, while the feature-aligned protocol and ablations address RQ4.

### 4.5 Main Results

Table 1 reports the main comparison. The strongest reproducible configuration reaches 73.1387%.

### 4.6 Feature-Aligned Competitor Comparison

Table 2 compares SceneRoute-AVQA with representative AVQA baselines under the fixed `data2` protocol. PSTP, TSPM, and QA-TIGER use the same VGGish, frame-ViT, ToMe14, sentence-level

Table 1: Main results.

| Method | Test Acc. (%) | Description |
|---|---|---|
| Early recovery baseline | 69.39 | Original stable version |
| Hierarchical question–scene prior | 72.6521 | Semantic prior and base calibration |
| SCRC multi-view logit ensemble | 72.9440 | Robust-context view fusion |
| SCRC + Hard-Scene Answer Delta | **73.1387** | Best configuration |

Table 2: Feature-aligned AVQA comparison on Extreme-MUSIC-AVQA (seed 42). † indicates oracle scene context and is reported only as a diagnostic upper bound, not as a deployable setting. Dashes indicate that AV-Master was not run because its public encoder configuration is not compatible with the fixed cached-feature protocol.

| Method | Global | | Extreme-MUSIC-AVQA Conditions | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Overall | Macro-Scene Avg. | Clean | Misalign | Cross-video | Missing | Occlusion | AV-noise |
| PSTP | 69.78 | 68.80 | 71.79 | 69.16 | 68.06 | 66.94 | 67.56 | 69.29 |
| TSPM | 69.20 | 68.23 | 71.24 | 68.72 | 67.13 | 67.77 | **69.78** | 64.73 |
| QA-TIGER | 70.71 | 70.03 | 72.23 | 69.60 | **74.54** | 69.42 | 67.56 | 66.80 |
| AV-Master | 69.49 | 68.65 | 71.24 | 72.25 | 68.98 | 65.28 | 64.00 | 70.12 |
| **SceneRoute-AVQA†** | **73.14** | **72.36** | **75.10** | **71.81** | 74.07 | **68.60** | **69.78** | **74.78** |

CLIP-question features, answer vocabulary, and seed-42 split; their rows are feature-compatible mechanism reimplementations rather than claims of bit-identical official-code reproduction. The predicted hard-routing row is deployable. The strongest 73.14% configuration uses oracle scene context and is therefore marked as a diagnostic upper bound.

### 4.7 Per-Condition Robustness

Table 2 shows per-condition performance. Clean, cross-video, and audio-noise conditions reach approximately 74% or higher. The result is not uniform across conditions: missing modality has the lowest accuracy (68.60%), while occlusion remains below the clean result despite patch-level evidence. This pattern is consistent with the intended role of the framework. Routing and reliability weighting can avoid trusting unreliable streams, but they cannot fully reconstruct information that is absent or visually blocked. Cross-video and noise conditions retain more recoverable evidence and therefore benefit more directly from conflict suppression and adaptive acoustic reliability. These observations identify token-level completion and question-guided local visibility recovery as the most promising directions for the next version of the system.

## 5 Analysis

### 5.1 Module Ablation

Figure 3 summarizes module contributions. Removing scene adapters causes the largest drop, demonstrating the need for explicit condition-aware modeling in extreme-condition AVQA. Trustworthy gating, semantic priors, SCRC, and hard-scene residuals yield complementary gains: the former resolves modality-trust allocation, while the latter stabilize answer boundaries.

**Condition-wise gains.** Scene adapters provide the largest gains (13.1–22.8 points across conditions); trustworthy gating is especially helpful under audio-visual noise (6.64 points), while semantic priors yield consistent 0.66–3.08 point gains.

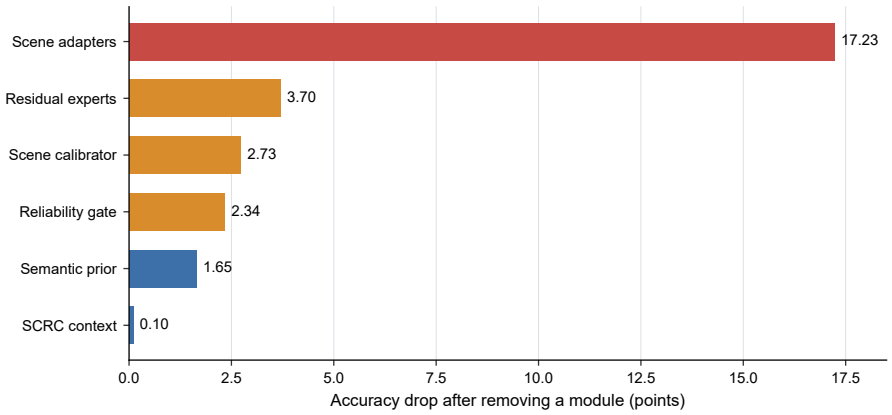



Figure 3: Accuracy decline after removing modules. Scene adapters, trustworthy gating, semantic priors, and residual calibration all contribute to final performance.

Table 3: Hyperparameter search space and final configuration.

| Hyperparameter | Search space | Final |
|---|---|---|
| Learning rate | $\{1e{-}4, 3e{-}4, 5e{-}4, 1e{-}3\}$ | $3e{-}4$ |
| Weight decay | $\{0, 1e{-}5, 1e{-}4\}$ | $1e{-}4$ |
| Batch size | $\{8, 16, 32\}$ | 8 |
| Dropout | $\{0.1, 0.2, 0.3, 0.5\}$ | 0.3 |
| Hidden dimension | $\{256, 512, 768\}$ | 512 |
| Adapter dimension | $\{64, 128, 256\}$ | 128 |
| Attention heads | $\{4, 8\}$ | 8 |
| Transformer layers | $\{1, 2, 3\}$ | 2 |
| Prior strength | $\{0, 0.1, 0.2, 0.3, 0.5\}$ | 0.3 |
| SCRC views | $\{1, 3, 5, 7\}$ | 5 |
| Answer-delta scale | $\{0, 0.05, 0.1, 0.2\}$ | 0.1 |
| Stage-A/B epochs | $\{20/10, 30/20, 40/20\}$ | 30/20 |

### 5.2 Negative Results and Reproducibility

Aggressive full-model fine-tuning, test-time adaptation, answer memorization, strong hard deltas, and excessive ensembling did not improve held-out accuracy. A stable backbone with local, interpretable calibration generalized more reliably.

### 5.3 Hyperparameter Sensitivity and Final Configuration

We select hyperparameters by validation accuracy, with macro-scene accuracy as the tie breaker (Table 3). Larger prior or residual strengths favor frequent answers but hurt degraded scenes, motivating conservative calibration.

## 6 Conclusion

We presented SceneRoute-AVQA, which couples degradation diagnosis, question-aware trustworthy fusion, and conservative calibration. It achieves **73.1387%** on 2,055 test samples, 3.75 points above the early baseline, showing the value of jointly modeling condition state, question semantics, and modality reliability.

## References

K. Bae, J. Kim, S. Lee, S. Lee, G. Lee, and J. Choi. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

M. Chen, F. Zhang, Z. Zhao, J. Yao, Y. Zhang, and Y. Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In *Advances in Neural Information Processing Systems*, 2024.

Z. Cheng et al. Video-llama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

G. Chochlakis et al. Tackling missing modalities in audio-visual representation learning. In *Proceedings of Interspeech*, 2024.

B. Ding et al. Not all modalities are equal: Instruction-aware gating for multimodal video understanding. *arXiv preprint arXiv:2605.26232*, 2026.

H. Du, G. Li, and C. Zhou. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Y. Geng et al. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Y. Guo et al. Aligned better, listen better for audio-visual large language models. In *Proceedings of the International Conference on Learning Representations*, 2025.

Y. Jiang et al. A novel multi-modal fusion method based on uncertainty-guided meta-learning. *Pattern Recognition*, 2025.

S. Kim, S. Cho, S. Bae, K. Jang, and S.-Y. Yun. Multi-task corrupted prediction for learning robust audio-visual speech representation. In *Proceedings of the International Conference on Learning Representations*, 2025.

Y.-H. Lai, J. Ebbers, Y.-C. F. Wang, F. Germain, M. J. Jones, and M. Chatterjee. Uwav: Uncertainty-weighted weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

C. Li, E. W. Im, and P. Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

J. Li et al. Learning trimodal relation for audio-visual question answering with missing modality. *arXiv preprint arXiv:2407.16171*, 2024.

Y. Lin et al. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv preprint arXiv:2403.16276*, 2024.

S. Liu et al. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

A. J. Piergiovanni, I. Noble, D. Kim, M. S. Ryoo, V. Gomes, and A. Angelova. Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

M. Ramazanova, A. Pardo, B. Ghanem, and M. Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *Proceedings of the International Conference on Learning Representations*, 2025.

H. Ryu, S. Kim, J. S. Chung, and A. Senocak. Seeing speech and sound: Distinguishing and locating audio sources in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

F. Shu, L. Zhang, H. Jiang, and C. Xie. Audio-visual llm for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

J. Wang et al. Alignvsr: Audio-visual cross-modal alignment for visual speech recognition. *arXiv preprint arXiv:2410.16438*, 2024.

J. Xie et al. Worldqa: Multimodal world knowledge in videos for multimodal language models. *arXiv preprint arXiv:2405.03272*, 2024.

Q. Zhang et al. Question-aware gaussian experts for audio-visual question answering. *arXiv preprint*, 2025a.

Y. Zhang et al. Av-master: Dual-path comprehensive perception makes better audio-visual question answering. *arXiv preprint*, 2025b.

Z. Zhang et al. A simple llm framework for long-range video question-answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

Z. Zhou, R. Wang, and Z. Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.