AVRT: AUDIO-VISUAL REASONING TRANSFER THROUGH SINGLE-MODALITY TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in reasoning models have shown remarkable progress in text-based domains, but transferring those capabilities to multimodal settings, e.g., to allow reasoning over audio-visual data, still remains a challenge, i.a., because of the limited availability of high-quality reasoning data in targeted multimodal combinations. To address this problem, we introduce AVRT, a novel framework that generates high-quality audio-visual reasoning traces from single-modality teacher models. We generate independent vision- and audio-reasoning traces via models specialized to reason over their respective modalities and merge the resulting traces with an LLM merger model. The resulting multimodal traces are used in a supervised fine-tuning (SFT) cold start to adapt the target model to audio-visual reasoning traces first, before training it in a second reinforcement learning stage on larger-scale data. Evaluated on seven benchmarks, our 3B and 7B models achieve state-of-the-art results among comparable-size models on OmniBench for audio-visual and MMAR for audio-only reasoning, with the 3B model also leading on DailyOmni, showing that cross-modal training transfers to single-modality tasks and establishing a new training pipeline for multimodal reasoning models. [1]

1 INTRODUCTION

Humans perceive the world by combining information from multiple modalities through diverse sensory inputs. With the wide availability of multimodal data, such as videos, multimodal understanding in general and audio-visual understanding in particular has drawn more and more interest from the research community. Recent advancements in this area, also in combination with large language models, have shown remarkable performance in audio-visual understanding (Cheng et al., 2024; Xu et al., 2025a; Liu et al., 2025b; Comanici et al., 2025; OpenAI, 2024).

In parallel, the emergence of reasoning-capable language models has led to new capabilities with respect to the analysis and understanding of a given scenario, exemplified by OpenAI's o-series (Jaech et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025). These advances have been significantly driven by reinforcement learning techniques (Shao et al., 2024). These reasoning capabilities have been successfully extended to inputs such as vision-text models (Huang et al., 2025; Dong et al., 2025) and audio-text models (Xie et al., 2025; Wen et al., 2025; Ghosh et al., 2025), demonstrating chain-of-thought capabilities within the respective modalities. However, audio-visual reasoning has not yet reached the same level of advancement as its single-modality counterparts, i.a., due to the challenge of integrating information and reasoning cues across different modalities at scale, as well as due to the practical lack of audio-visual reasoning data in general. Existing approaches try to address this problem, e.g., by generating reference reasoning chains from large foundation teacher models that were trained with all target modalities (Du et al., 2025) or by trying to approach the problem by extending reinforcement learning formulations, e.g., by improved credits assignment or by context summarization (Yang et al., 2025; Kulkarni & Fazli, 2026).

This paper proposes a new pipeline for Audio-Visual Reasoning Transfer (AVRT) based on single-modality teachers. To this end, AVRT extracts detailed chains of thought independently from specialized visual and audio teachers, then merges them with a text-only aggregator into single cross-modal reasoning traces. An overview is shown in Fig. 1. Namely, we combine traces from specialized single-modality reasoning models via a text-only LLM merger model to generate a coherent

[1] All code, data, and checkpoints will be made available.

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.

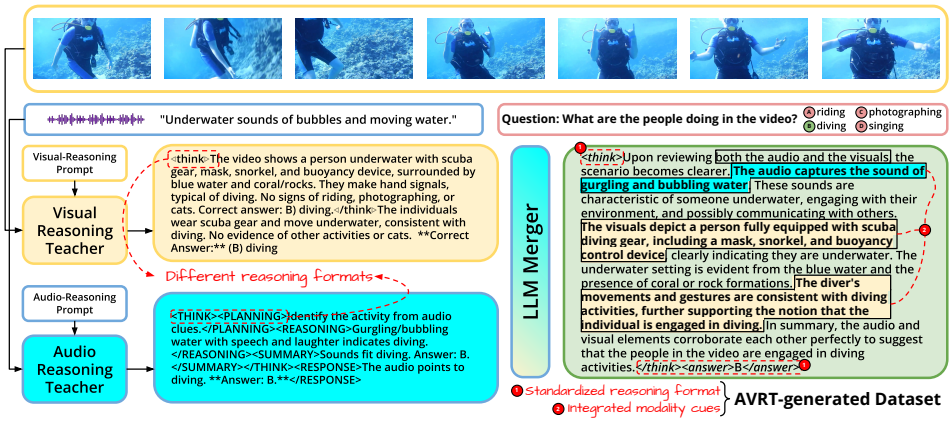


Figure 1: **Overview of the AVRT pipeline:** We first generate reasoning chains from single-modality teacher models that are prompted in the format they were optimized for and, second, leverage an LLM merger as an interface between the teacher models and the resulting reasoning chain to aggregate the information and put it into the target format. The resulting audio-visual traces are then used to train a student model in a cold-start manner via supervised fine-tuning (SFT) followed by GRPO fine-tuning for audio-visual question-answering.

multimodal reasoning that explicitly correlates information across audio and visual channels. By using an LLM as a merging interface for the different teacher models, every model can be prompted in the format that it was trained for, leading to high-quality modality-specific outputs. The following LLM merger then allows to aggregate the information, including different meta tags, and to convert it into a consistent target format. The resulting audio-visual traces are used to fine-tune a student audio-visual LLM in a cold-start manner via supervised fine-tuning (SFT) to learn the reasoning formatting and patterns, as well as to integrate multimodal reasoning into the student model. In a second step, the student model is trained following the GRPO (Shao et al., 2024) objective.

We evaluate the proposed approach on seven challenging datasets: First, as zero-shot vision-audio downstream datasets, we consider DailyOmni (Zhou et al., 2025), OmniBench (Li et al., 2024), Video-MME (Fu et al., 2025), and RivaBench (Sun et al., 2025); second, we evaluate on AVQA (Yang et al., 2022) as in-domain validation dataset; and finally on MMAR (Ma et al., 2025) and MMAU (Sakshi et al., 2024) as audio-only downstream task. With 3B- and 7B-parameter students, AVRT achieves state-of-the-art results among 3B models and on most benchmarks among 7B models. Our ablation shows that the combination of reasoning traces of two different teacher models by a language-only merger model improves audio-visual reasoning in the GRPO learning stage. Our contribution can be summarized as follows: 1) We propose a novel method that aggregates single-modality reasoning traces into coherent multimodal reasoning via a text-only merger, a recipe applicable to any modality combination lacking reasoning data, which we instantiate here for the audio-visual domain. 2) We are the first to train a multimodal reasoner based on composed single-modality reasoning data, demonstrating that cross-modal reasoning can emerge from composed single-modality expertise without requiring a natively multimodal reasoner as teacher. 3) We conduct an extensive evaluation on audio-visual benchmarks, achieving state-of-the-art results that compete with larger models. We show that traces composed from single-modality teachers through a text-only merger provide better supervision than either teacher's traces alone (see Table 5), and that training on audio-visual data improves single-modality performance (see Table 4), providing evidence of genuine cross-modal reasoning transfer.

2 Related Work

Audio-Visual Large Language Models. Audio-visual large language models have made constant progress in addressing challenging tasks in multimodal understanding. Early works like Meerkat (Chowdhury et al., 2024) focus on fine-grained spatial and temporal grounding on five audio-visual tasks, introducing optimal transport-based modality alignment and cross-attention modules for audio-visual consistency. AVicuna (Tang et al., 2025) proposes specifically targeting temporal

Table 1: **Comparison of audio-visual question answering datasets.** AVRT-20K provides reasoning traces that integrate audio (A) and visual (V) modalities, addressing a key limitation in AVQA datasets which focus on question-answer pairs without intermediate reasoning steps. All datasets use multiple-choice questions with 4 options. The number of QA pairs is reported in thousands (K).

Dataset	Modalities	Answer Format	# QA pairs (K)
AVQA	A+V	Video + question + 4-way answer	57.3
OmniInstruct-v1	A+I	Image + question + 4-way answer	96.1
MUSIC-AVQA	A+V	Video + question + 4-way answer (focused on music)	45.9
AVQA - R1 - 6K	A+I	Image + question + 4-way answer (subset of OmniInstruct)	6.4
AVRT-20K (ours)	A+V	Video + question + 4-way answer + Reasoning chains incorporating audio and visual data	19.2

referential dialogue in untrimmed videos, introducing the Audio-Visual Tokens Interleaver for temporal alignment. VideoLLaMA 2 (Cheng et al., 2024) advances spatial-temporal modeling through specialized Spatial-Temporal Convolution connectors and enhanced audio understanding via joint training, achieving state-of-the-art performance among open-source models. Rather than developing new models from scratch, PAVE (Liu et al., 2025a) introduces a lightweight adaptation framework that extends existing video LLMs to other modalities through efficient "patches" that add only 0.1% additional parameters. More recent work moves toward comprehensive omni-modal capabilities, with Qwen2.5-Omni (Xu et al., 2025a) enabling end-to-end streaming multimodal inputs and outputs through innovations like TMRoPE for synchronizing video with audio and the Thinker-Talker architecture for concurrent text and speech generation. Ola (Liu et al., 2025b) proposes progressive modality alignment training strategies that use video as a central bridge to connect modalities. Despite these advances, these existing approaches often struggle to effectively associate information across both modalities, lacking structured reasoning processes that can explicitly capture and use cross-modal dependencies and correlations for comprehensive multimodal understanding. AVRT addresses this gap by introducing explicit cross-modal reasoning traces generated from single-modality specialists, enabling structured reasoning that correlates information across audio and visual channels without requiring end-to-end multimodal pretraining.

Audio-Visual Reasoning. Audio-visual reasoning with large language models has seen rapid advancements. Daily-Omni (Zhou et al., 2025) introduces a dedicated Audio-Visual QA dataset with an agent leveraging visual, audio, and speech modalities. Recent work has increasingly leveraged reinforcement learning: EchoInk-R1 (Xing et al., 2025) uses GRPO for structured cross-modal reasoning, HumanOmniV2 (Yang et al., 2025) addresses shortcut problems through context summarization and logical rewards, Omni-R1 (Zhong et al., 2025) handles temporal-spatial trade-offs, AVATAR (Kulkarni & Fazli, 2026) introduces Temporal Advantage Shaping for credit assignment, and Video-RTS (Wang et al., 2025) combines data-efficient RL with test-time scaling. video-SALMONN-o1 (Sun et al., 2025) is the first open-source reasoning-enhanced audio-visual LLM with pDPO training and the RivaBench benchmark. Building on the success of Qwen2.5-Omni, Qwen3-Omni (Xu et al., 2025b) extends omni-modal capabilities to reasoning with its MoE architecture. More recently, AURELIA (Chowdhury et al., 2025) introduces a test-time reasoning distillation approach that uses three different forward passes to generate step-by-step reasoning, which is then provided as input alongside the audio-visual data and question to guide the model's reasoning process. Rather than relying on test-time adaptation strategies that require multiple inference passes we use single-modality specialist models as teachers to enhance a multimodal student model through knowledge distillation, constructing a dataset with explicit audio-visual reasoning chains. This allows for a supervised fine-tuning (SFT) approach, supplemented by RL, to produce a model that generates answers in a single forward pass, significantly reducing inference cost compared to multi-pass test-time strategies.

Audio-Visual Datasets. AVQA (Yang et al., 2022) can be considered one of the foundational audio-visual QA datasets with $57,335$ question-answer pairs from daily audio-visual activities requiring clues from both modalities. OmniInstruct (Li et al., 2024) develops a comprehensive tri-modal reasoning dataset combining visual, audio, and textual resources, while MUSIC-AVQA (Li et al., 2022) expands to musical performance with $45,867$ question-answer pairs across $9,288$ videos. AVQA-R1-6K (Xing et al., 2025) provided a manually curated subset of OmniInstruct focusing on questions that are more likely to require audio-visual reasoning. While these datasets have advanced the field significantly, they primarily focus on question-answer pairs without providing explicit reasoning traces that demonstrate how models should integrate cross-modal information. Our AVRT approach directly addresses this gap through structured reasoning chain generation, as compared in

Table 1. By providing explicit cross-modal reasoning traces rather than answer-only supervision, AVRT enables models to integrate audio and visual evidence.

3 METHODOLOGY

In this paper, we derive audio-visual reasoning traces based on audio-visual question-answering pairs as e.g. provided by AVQA (Yang et al., 2022). We describe the generation of reasoning traces in Sec. 3.1 and our training procedure for audio-visual question answering in Sec. 3.2.

3.1 CROSS-MODAL REASONING TRACE GENERATION

Our data generation process consists of two main stages: a single-modality reasoning extraction step and a cross-modal aggregation step. Let an audio-visual question-answering data sample be denoted as (X, Q), where X is a video and Q is a question. The video X comprises both an audio stream A and a visual stream V, such that $X = (A, V)$.

Single-Modality Reasoning Extraction. First, we generate modality-specific reasoning. We select specialized teacher models for the audio (T_A) and visual (T_V) modalities. For a given sample (X, Q), we provide each teacher with the question and its corresponding modality. We use carefully crafted prompts, P_A and P_V, to elicit detailed reasoning traces. The audio reasoning trace is generated as $R_A = T_A(Q, A, P_A)$, and the visual reasoning trace is $R_V = T_V(Q, V, P_V)$. These traces capture the unique characteristics and patterns of each modality.

Cross-Modal Aggregation. In the second stage, we perform cross-modal aggregation. We use a text-only large language model, M_{agg}, to merge the reasoning outputs. This model takes the reasoning traces from both modalities and the original question to produce a unified, cross-modal reasoning output: $R_{agg} = M_{agg}(Q, R_A, R_V)$. This aggregation step transforms the diverse reasoning formats into a uniform structure, correlating characteristics from both modalities.

Filtering. To avoid introducing noise during the SFT, we consider only reasoning traces where both modality-specific teachers generate correct responses. This filtering strategy ensures high-quality training data by avoiding the propagation of erroneous reasoning patterns that could introduce noise during cross-modal aggregation (Turpin et al., 2023; Xie, 2024). As a result, we only keep a 20K subset of the original 40K AVQA data samples. The ablation in Section 4.5 shows that filtering improves OmniBench accuracy over training on the unfiltered data (56.3% vs. 53.6%), while the model still generalizes to questions that neither teacher answers correctly.

3.2 TRAINING

Stage 1: Supervised Fine-Tuning. We fine-tune the base model on the merged audio-visual reasoning traces using an autoregressive language loss. Given a training sample (X, Q, R_{agg}) where $X = (A, V)$ is the video with audio and visual streams, Q is the question, and R_{agg} is the aggregated reasoning trace, we optimize the cross-entropy loss:

$$\mathcal{L}_{SFT} = -\sum_{t=1}^{|R_{agg}|} \log p_\theta(r_t | X, Q, r_{<t}), \tag{1}$$

where r_t represents the t-th token in the reasoning trace R_{agg} and θ are the model parameters. The model learns to generate structured reasoning following the format established during cross-modal aggregation: `<think>...</think> <answer>...</answer>`, where the thinking section contains the multimodal reasoning and the answer section provides the final response.

Stage 2: Reinforcement Learning. In a second step, we employ Group Relative Policy Optimization (GRPO)(Shao et al., 2024). GRPO eliminates the need for explicit value function estimation by deriving advantage estimates through group-based comparisons of model outputs.

The GRPO training operates by sampling G distinct responses $\{o_1, o_2, \ldots, o_G\}$ for each input question q using the current policy $\pi_{\theta_{old}}$. Each response o_i receives a scalar reward r_i from our reward function. The advantage for o_i is computed by normalizing rewards within the group:

$$\hat{A}_{i,t} = \widetilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}, \tag{2}$$

Table 2: Statistics of the AVRT-20K dataset showing sample counts, formatting, token length as well as data properties for the training and validation split.

AVRT-20K Metrics	**Train / Val**
Total Samples	18,279 / 945
Reasoning Format Compliance	100.0% / 100.0%
Thinking Section Length (tokens)	165.5 ± 33.9 / 163.4 ± 32.5
Answer Section Length (tokens)	1.0 ± 0.0 / 1.0 ± 0.0
Video and Audio Duration (sec)	10.0 ± 0.1 / 10.0 ± 0.2
Primary Resolution	1280×720 (62% / 43%)

where this advantage $\widetilde{r}_i$ is applied uniformly across all tokens t in response o_i.

Our reward combines three terms to enforce correctness, formatting, and reasoning quality:

$$r_i = R_{\text{format}}(o_i) + R_{\text{acc}}(o_i) + R_{\text{length}}(o_i). \tag{3}$$

The three components are defined as:

(1) Format Reward (R_{format}): A binary reward that verifies adherence to our proposed reasoning format (`<think>...</think><answer>...</answer>`):

$$R_{\text{format}}(o_i) = \mathbb{I}(\text{format is correct}). \tag{4}$$

(2) Final Answer Accuracy (R_{acc}): A simple string matching evaluation that compares the model's predicted answer choice against the ground truth label:

$$R_{\text{acc}}(o_i) = \mathbb{I}(\text{answer is correct}). \tag{5}$$

(3) Reasoning Length Reward (R_{length}): A dense reward that encourages optimal reasoning trace length using a Gaussian-shaped distribution. The reward is:

$$R_{\text{length}}(o_i) = \min\Big(1.0,\ \exp\Big(-\frac{(w_i-\mu)^2}{2\sigma^2}\Big) + \mathbb{I}(w_{\min} \le w_i \le w_{\max}) \cdot b\Big), \tag{6}$$

where w_i is word count, μ is the optimal target, σ is the width, and b is a bonus in $[w_{min}, w_{max}]$.

4 Experiments

4.1 Training Datasets

AVRT-20K for Supervised fine-tuning. In the first phase of the training, we fine-tune the respective student model in a fully supervised way on pre-constructed audio-visual reasoning traces. To this end, we introduce the AVRT-20K dataset, which is constructed using our proposed AVRT method on a subset of the AVQA dataset. We use Kimi-VL-Thinking (Team et al., 2025) and Audio Flamingo 3 (*think*) (Ghosh et al., 2025) as the single-modality teachers T_V and T_A. These models were chosen due to their balance between achieving state-of-the-art results in their modalities, and generating descriptive reasoning traces. We use 10-second audio input and 8 uniformly-sampled video frames from each sample as the input for the audio and visual teacher, respectively. The full prompt templates used for each model can be found in the supplementary material.

The final collection comprises $18,279$ training and 945 validation samples from AVQA; Table 2 summarizes its statistics. The question-type distribution closely mirrors that of AVQA (App. B).

RL Training. In the second phase, the resulting model is trained on the full AVQA training set (Yang et al., 2022). AVQA comprises $57,335$ question-answer pairs across $57,015$ videos from daily activities. The dataset is split into $40,127$ training samples and $17,208$ validation samples.

4.2 Benchmark Datasets

We evaluate the resulting models on a diverse set of benchmarks spanning different modality combinations to assess cross-modal reasoning capabilities.

Audiovisual. *AVQA* (Yang et al., 2022) is a large-scale benchmark containing 57,335 question-answer pairs across 57,015 videos designed to evaluate models' ability to reason over both audio and visual content. Since our training data is drawn from AVQA, it is an *in-domain* evaluation.
DailyOmni (Zhou et al., 2025) evaluates models on real-life audio-visual scenarios that require joint reasoning across video, audio, and textual information. It contains 684 videos and 1,197 question-answer pairs (550 on 60-second, 647 on 30-second videos) from all 11 YouTube categories.
OmniBench (Li et al., 2024) evaluates models on integrating image, audio, and text inputs for cross-modal reasoning. The benchmark contains 1,139 question-answer pairs organized into 8 categories, with audio clips averaging 9.22 seconds in duration.
Video-MME (Fu et al., 2025) is a comprehensive multi-modal evaluation benchmark for video analysis, spanning 6 primary visual domains with 30 subfields, covering short-, medium-, and long-term videos ranging from 11 seconds to 1 hour, with 2,700 question-answer pairs across 900 videos. While commonly adopted as a general video understanding benchmark, Video-MME also integrates audio information, making it suitable for evaluating audio-visual capabilities. In our evaluation, all models receive both the video frames and the audio track as input.
RivaBench (Sun et al., 2025) extends the scope of complex video understanding with reasoning-intensive application scenarios, including academic presentations (Academic) and stand-up comedy (StandUp). It contains 1,912 and 2,128 five-way multiple-choice questions for the Academic and StandUp scenarios, respectively, on YouTube videos averaging 47.2s and 43.2s.

Audio-Only. *MMAR* (Ma et al., 2025) is an audio reasoning benchmark designed to evaluate models' ability to perform complex reasoning tasks on auditory information.
MMAU (Sakshi et al., 2024) evaluates multimodal audio understanding on tasks requiring expert-level knowledge and complex reasoning. It comprises 10k audio clips paired with human-annotated questions spanning speech, environmental sounds, and music, covering information extraction and reasoning tasks across 27 distinct skills. We include both audio benchmarks to assess how well the proposed cross-modal training approach transfers to single-modality audio reasoning scenarios.

4.3 Implementation Details

For all experiments, we use Qwen2.5-Omni (Xu et al., 2025a) as the base student models with frozen vision and audio modules. Supervised fine-tuning is conducted on $18,279$ samples over 1 epoch with an effective batch size of 32 (1 sample per device $\times$ 8 gradient accumulation steps $\times$ 4 H100 GPUs). We use a learning rate of $2e-6$ with cosine scheduling, AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$), weight decay of 0.01, and 100 warmup steps. Training employs DeepSpeed ZeRO Stage 2 optimization with CPU offloading and bfloat16 precision. For reinforcement learning, we use identical infrastructure with GRPO-specific hyperparameters: group size $G = 4$, clipping parameter $\epsilon = 0.2$, KL regularization coefficient $\beta = 0.01$, and temperature 1. For the reasoning length reward, the optimal target length is set to $\mu = 100$ words with a SD of $\sigma = 20$ words, and the target range for the bonus to $w_{min} = 100$ and $w_{max} = 200$ words.

4.4 Comparison to State-of-the-Art

Table 3 shows the performance of the proposed AVRT-trained model across seven benchmark datasets spanning different modality combinations. We include Qwen3-Omni (Xu et al., 2025b) (30B parameters) as an upper bound to contextualize improvements, and report the single-modality teachers as reference points for our distillation process. Our 3B- and 7B-parameter models achieve strong performance both in absolute accuracy and relative improvement over base models, frequently surpassing the single-modality teachers they learn from.

Audio-Visual Benchmarks. On all datasets, the 3B model significantly outperforms the base model, surpassing even some 7B models, e.g., on DailyOmni with 49.2% (+6.1) (improvements wrt. Qwen2.5 Omni baselines) and OmniBench with 56.3% (+6.1) accuracy respectively. The 7B variant achieves 54.4% (+2.9) on DailyOmni and 57.1% (+6.4) on OmniBench, outperforming most baselines with the exception of video-SALMONN-o1, which achieves 64.0% on DailyOmni. This gap reflects a difference in specialization as video-SALMONN-o1 is trained on speech-heavy data, which aligns well with DailyOmni's dialogue and narration tasks, whereas AVRT targets general sound reasoning. On benchmarks beyond speech, AVRT consistently outperforms video-SALMONN-o1 as for OmniBench (57.1 vs. 40.5), MMAR (59.1 vs. 51.2), and MMAU (75.4 vs. 53.8). For RivaBench, the

Table 3: Comparison of audio-visual reasoning models on benchmark datasets. AVQA is considered in-domain since the training dataset is derived from its training set, and therefore excluded from the average. All other models are tested in zero-shot mode without finetuning. Reproduced baseline results marked with *. 🧠 stands for reasoning.

Model	🧠	*Audio-Visual*							*Audio*		
		AVQA†	**DailyOmni**	**OmniBench**	**Video-MME**	**Riva-S**	**Riva-A**	*AV Avg.*	**MMAR**	**MMAU**	*A Avg.*
30B Audio-Visual Models											
Qwen3-Omni* (Xu et al., 2025b)	✓	93.8	57.5	63.6	57.5	81.7	56.2	63.3	66.4	76.5	71.5
Modality-Specific Teachers											
Kimi-VL-Thinking* (Team et al., 2025)	✓	81.4	47.0	33.5	58.9	72.3	49.6	52.3	*N/A*	*N/A*	*N/A*
AF3 *(think)** (Ghosh et al., 2025)	✓	64.3	52.5	28.9	45.0	65.5	41.7	46.7	60.1	73.3	66.7
3B Audio-Visual Models											
AVATAR (Kulkarni & Fazli, 2026)	✓	-	44.7	45.8	-	-	-	-	-	-	-
Qwen2.5 Omni* (Xu et al., 2025a)	×	88.3	43.1	50.2	55.4	62.4	38.4	49.9	53.7	61.1	57.4
AVRT (Ours)	✓	**91.1** (+2.8)	**49.2** (+6.1)	**56.3** (+6.1)	**62.6** (+7.2)	**71.3** (+8.9)	**49.3** (+10.9)	**57.7** (+7.8)	**57.3** (+3.6)	**70.0** (+8.9)	**63.7** (+6.3)
7B Audio-Visual Models											
EchoInk-R1 (Xing et al., 2025)	✓	-	46.2	46.5	-	-	-	-	-	-	-
Omni-R1 (Zhong et al., 2025)	✓	-	46.8	46.9	60.7	-	-	-	-	-	-
HumanOmni (Zhao et al., 2025)	✓	-	47.6	44.9	-	-	-	-	-	-	-
Ola-7B (Liu et al., 2025b)	×	-	52.3	45.3	**68.4**	-	-	-	-	-	-
AV-Reasoner (Lu et al., 2025)	✓	-	53.8	48.3	-	-	-	-	-	-	-
AVATAR (Kulkarni & Fazli, 2026)	✓	-	47.0	49.1	-	-	-	-	-	-	-
V-RTS* (Wang et al., 2025)	✓	86.6	47.8	35.3	63.0	70.3	48.9	53.1	*N/A*	*N/A*	*N/A*
v-SALMONN-o1* (Sun et al., 2025)	✓	84.8	**64.0**	40.5	41.3	**76.7**	48.3	54.2	51.2	53.8	52.5
Qwen2.5 Omni* (Xu et al., 2025a)	×	84.9	51.5	50.7	58.1	74.9	41.3	55.3	56.5	71.9	64.2
AVRT (Ours)	✓	**90.4** (+5.5)	54.4 (+2.9)	**57.1** (+6.4)	64.1 (+6.0)	75.3 (+0.4)	**50.7** (+9.4)	**60.3** (+5.0)	**59.1** (+2.6)	**75.4** (+3.5)	**67.3** (+3.1)

3B model achieves 71.3% (+8.9) on StandUp (Riva-S) and 49.3% (+10.9) on Academic (Riva-A). The 7B variant achieves 75.3% (+0.4) on StandUp and 50.7% (+9.4) on Academic, showing that performance improvements scale across model sizes. The large gains on Academic demonstrate that the reasoning approach is particularly effective on expert-level content requiring deeper understanding of technical material. On the in-domain AVQA, the 3B and 7B models reach 91.1% (+2.8) and 90.4% (+5.5).

Audio Benchmarks. On audio-only reasoning, the model shows strong transfer capabilities. For MMAR, the 3B model achieves 57.3% (+3.6) and approaches the specialized audio teacher AF3 *(think)* (60.1%), while the 7B model reaches 59.1% (+2.6). On MMAU, AVRT 3B achieves 70.0% (+8.9) and the 7B model achieves 75.4% (+3.5), demonstrating that audio reasoning capabilities learned through multimodal training transfer effectively to complex single-modality tasks. Overall, AVRT 3B achieves an audio-visual average of 57.7% (+7.8) and an audio average of 63.7% (+6.3), while AVRT 7B reaches 60.3% (+5.0) and 67.3% (+3.1) respectively. The 7B model shows smaller gains than the 3B model (AV Avg. +5.0 vs. +7.8), as expected given its stronger baseline; this matches distillation literature, where weaker students benefit more from teacher guidance (Busbridge et al., 2025). Importantly, the 7B model still improves on all benchmarks.

4.5 Ablation Studies

Evaluation of SFT fine-tuning. We first assess the impact of the supervised fine-tuning step on the generated reasoning traces compared to the Qwen2.5-Omni 3B baseline, as well as to the same model trained only with an RL objective. As shown in Table 4, simply training the model with an RL objective leads to an improvement of 2.1% (52.3% vs. 50.2%) on the audio-visual setting. The proposed 2-stage training with a SFT cold-start based on the generated reasoning traces further improves performance to 56.3%, an additional gain of 4.0 points over the RL-only baseline. We hypothesize that SFT provides two key benefits for the subsequent RL stage. First, *format learning*: SFT teaches the model the structured reasoning format (thinking section followed by answer), ensuring high R_{format} and R_{length} rewards from the start of RL training. Second, *reasoning priors*: The model learns cross-modal correlation patterns from the distilled traces, which guide RL exploration toward productive reasoning strategies rather than exploring randomly.

Single Modality Performance. To further investigate the impact of training on multimodal data on the respective single-modality performance, we evaluate the model using one modality at a time. As shown in Table 4, the results demonstrate that multimodal training with the proposed reasoning trace aggregation approach not only leads to improvements on audio-visual settings, but also on the single-modality performance. This suggests reasoning transfer, as both AVRT and the RL-only baseline are trained solely on audio-visual inputs, without modality dropout or single-modality augmentation.

Reasoning trace types. To validate the effectiveness of the audio-visual reasoning trace aggregation, we compare the approach against alternative supervision strategies. As shown in Table 5,

Table 4: Ablation of training stages on OmniBench for only audio(A), only vision(V), and audio-visual input(AV).

Model	SFT	RL	Mod.	OmniBench
Qwen2.5-Omni 3B (Baseline)	×	×	A	39.4
	×	×	V	42.7
	×	×	AV	50.2
Qwen2.5-Omni 3B + RL (Baseline+RL)	×	✓	A	41.1 (+1.7)
	×	✓	V	43.2 (+0.5)
	×	✓	AV	52.3 (+2.1)
AVRT (Ours)	✓	✓	A	**41.9** (+2.5)
	✓	✓	V	**45.8** (+3.1)
	✓	✓	AV	**56.3** (+6.1)

Table 5: Ablation on reasoning chain types for SFT. Models are fine-tuned on audio-only resp. video-only reasoning chains, then trained with AV-RL. Our audio-visual aggregation outperforms single-modality chains.

Reasoning Chain Modality	OmniBench
Qw2.5-O-3B + RL	52.3
Qw2.5-O-3B + SFT Audio-Only + RL	51.0 (-1.3)
Qw2.5-O-3B + SFT Video-Only + RL	52.1 (-0.2)
Audio-Visual (Ours)	**56.3** (+4.0)

Table 6: Ablation on merger models. An LLM merger from the same model family as the student enhances performance.

LLM Merger	Student Model	OmniBench
Gemma3-12B-It	Qwen2.5-Omni-3B	48.5
Qwen2.5-14B-Instruct	Qwen2.5-Omni-3B	**56.3**

Table 7: Ablation on reasoning chain length reward. Including a length reward consistently improves performance.

Reward Components	DailyOmni	OmniBench
Baseline	43.1	50.2
$R_{\text{acc}} + R_{\text{format}}$	45.5 (+2.4)	54.7 (+4.5)
$R_{\text{acc}} + R_{\text{format}} + R_{\text{length}}$	**49.2** (+6.1)	**56.3** (+6.1)

we compare SFT on audio-only, video-only, and audio-visual aggregated traces, each followed by RL, against RL alone (52.3%): audio-only traces lower accuracy (-1.3), video-only traces leave it nearly unchanged (-0.2), and audio-visual traces improve it ($+4.0$). The audio-visual reasoning traces achieve the best performance at 56.3%, substantially outperforming single-modality alternatives. This demonstrates that aggregating diverse reasoning perspectives from both modalities provides more effective supervision than training on single-modality traces, which may introduce modality-specific biases that hinder cross-modal integration.

Different merger models. We investigate the impact of using different teacher models for cross-modal aggregation in our pipeline. As shown in Table 6, we compare two merger models: Gemma3-12B-It and Qwen2.5-14B-Instruct. While Qwen2.5-14B-Instruct improves over the Qwen2.5-Omni 3B baseline by +6.1 points, Gemma3-12B-It decreases it by 1.7 points. The 8-point gap reveals the importance of architectural alignment. During training, Qwen2.5-14B-Instruct converged significantly faster to the MCQ format than Gemma3, which we attribute to tokenization alignment. Recent work on reasoning transfer (Bousselham et al., 2026) shows that shared architecture between teacher and student is critical for effective distillation. Here, using a merger with the same backbone as the student enables aligned token distributions and faster convergence, whereas the Gemma3 merger's different token patterns force the student to adapt to a different distribution.

Reasoning length reward. Next, we consider the impact of incorporating a reward function that encourages optimal reasoning trace length. As shown in Table 7, compared to the Qwen2.5-Omni 3B baseline (DailyOmni 43.1%, OmniBench 50.2%), the format and accuracy rewards alone achieve 45.5% and 54.7% respectively. Adding the length reward further improves to 49.2% (+6.1 points on DailyOmni) and 56.3% (+6.1 points on OmniBench). Notably, the Gaussian-shaped reward penalizes both too-short and too-long responses, discouraging runaway verbosity.

Dual-Teacher Filtering Strategy. Our approach keeps only samples where both teachers answer correctly (Sec. 3.1), halving the data from 40K to 20K. A potential concern is that this filtering might introduce bias toward "easy" samples. To validate that filtering does not limit generalization, we compare filtered vs. unfiltered training on OmniBench. Training with unfiltered data (40K samples, matched batch sizes) yields 53.6% accuracy, while our filtered approach achieves 56.3% (-2.7 points for unfiltered). Filtering thus improves performance by selecting high-quality reasoning traces, and the hard OmniBench questions (54% of the test set) confirm that the model generalizes beyond the training distribution. Question types where both teachers systematically fail would be absent from SFT supervision. However, since the two teachers operate on different modalities, their failure modes are largely decorrelated: the audio teacher fails on visually-dependent questions while the visual teacher fails on audio-dependent ones, limiting shared blind spots.

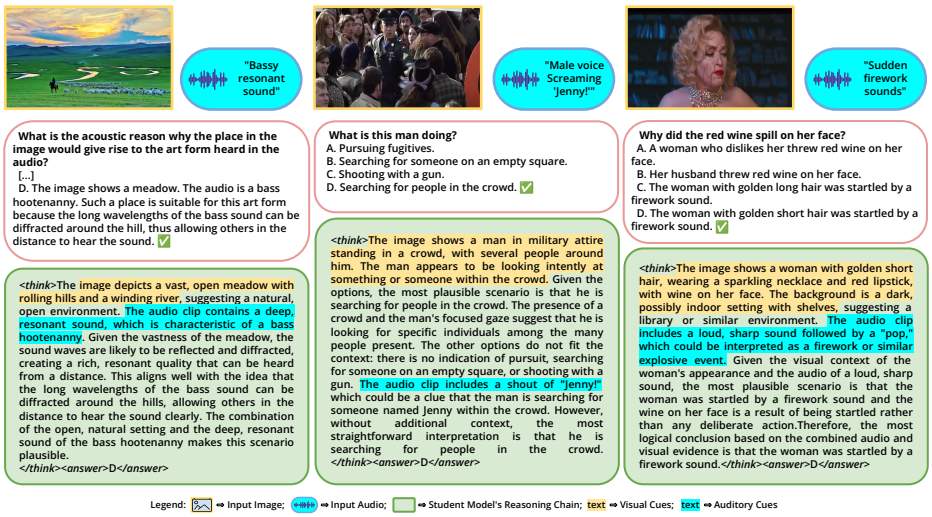


Figure 2: Qualitative results of the AVRT-trained model on OmniBench. The model retrieves cues from both the audio and the image and combines them in its reasoning chain to answer the question.

OmniBench Subsets. To check whether filtering biases the model toward easy samples, we split OmniBench questions by teacher performance: *easy* (both teachers correct, 64 questions), *medium* (one teacher correct, 456 questions), and *hard* (neither correct, 619 questions). The model outperforms the Qwen2.5-Omni baseline on all difficulty levels: easy (76.9% vs. 70.5%), medium (59.2% vs. 53.5%), and hard (51.8% vs. 45.3%). The +6.5 point gain on hard questions, where neither teacher was correct, shows that the student generalizes beyond the filtered training distribution.

Format Learning vs. Cross-Modal Reasoning. To test whether the gains stem from the output format or from reasoning, we analyze error types on OmniBench. We classify each incorrect prediction as an *Instruction Format Error* (IFE), where the response does not follow the required answer format (*e.g.* missing or malformed answer tags), or a *Logic Error*, where it is correctly formatted but wrong, indicating a reasoning failure. Compared to the baseline, AVRT reduces logic errors by 5.4 points (48.9% $\rightarrow$ 43.5%) while reducing IFE by only 0.7 points (0.9% $\rightarrow$ 0.2%). Thus, the gains come predominantly from cross-modal reasoning rather than format learning.

4.6 Qualitative Results

Fig. 2 shows reasoning traces of the AVRT-trained student on OmniBench. In the first example, the model links the acoustic properties of a bass "hootenanny" to the open meadow, reasoning about how sound propagates in open spaces. In the second, it uses visual cues (the people surrounding the main character) to interpret the audio (a male voice calling "Jenny!") and correctly identifies a search for people in a crowd. In the third, it connects a temporal audio event (the firework sound) with visual evidence (the wine spill on the woman's face). Notably, the model generalizes from its training domain (AVQA videos with 8 frames) to single static images paired with audio.

5 Conclusion

We introduced AVRT, a framework that generates high-quality audio-visual reasoning traces by merging outputs from specialized single-modality teachers via a text-only LLM, and uses them as cold-start in a two-stage training pipeline. Evaluated on seven benchmarks, the 3B model achieves state-of-the-art results among comparable-size models: 56.3% on OmniBench, 49.2% on Daily-Omni, and 57.3% on MMAR, with the 7B model further improving to 57.1%, 54.4%, and 59.1%, respectively. Beyond audio-visual tasks, cross-modal training also improves single-modality reasoning, providing evidence of genuine reasoning transfer. While instantiated here for the audio-visual domain, the approach is applicable to any modality combination lacking reasoning data.

AI USE STATEMENT

Generative models are part of the proposed method (Sec. 3.1) and evaluation (App. A.1). Otherwise, generative AI was used only to polish the English; the authors reviewed all edits and take full responsibility for the content.

REFERENCES

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. In *CVPR*, 2026.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In *ICML*, 2025.

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

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

Sanjoy Chowdhury, Hanan Gani, Nishit Anand, Sayan Nag, Ruohan Gao, Mohamed Elhoseiny, Salman Khan, and Dinesh Manocha. Aurelia: Test-time reasoning distillation in audio-visual llms. In *ICCV*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *CVPR*, 2025.

Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *CVPR*, 2025.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *NeurIPS*, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Yogesh Kulkarni and Pooyan Fazli. Avatar: Reinforcement learning to see, hear, and reason over video. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. *CVPR*, 2022.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.

Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. Pave: Patching and adapting video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025a.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025b.

Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. *arXiv preprint arXiv:2506.05328*, 2025.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Proc. NeurIPS*, 2025.

OpenAI. Hello gpt-4o, 2024. URL `https://openai.com/index/hello-gpt-4o`.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN-o1: Reasoning-enhanced audio-visual large language model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2023.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

Zikai Xie. Order matters in hallucination: Reasoning order as benchmark and reflexive prompting for large-language-models. *arXiv preprint arXiv:2408.05093*, 2024.

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

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*, 2022.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*, 2025.

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

SUPPLEMENTARY MATERIAL

This supplementary material is organized as follows. Section A presents additional analyses that further investigate the source and nature of AVRT's improvements: Section A.1 evaluates the quality and groundedness of the generated reasoning traces against a native multimodal reasoner (Qwen3-Omni-Thinking), including a hallucination source attribution analysis; Section A.2 examines model behavior under missing modalities; and Section A.3 presents control experiments that systematically ablate the SFT data composition to isolate the contribution of reasoning trace content from output formatting and answer exposure. Section B provides the complete dataset construction prompts and additional dataset statistics. Section C discusses limitations and future work.

A ADDITIONAL ANALYSES

A.1 REASONING TRACE QUALITY ANALYSIS

To assess the quality of our generated reasoning traces, we conduct an automated evaluation using Gemini-3.1-Flash-Lite as an external judge, comparing AVRT pipeline traces against those produced by Qwen3-Omni-Thinking (Xu et al., 2025b), a native multimodal reasoning model. We perform two types of evaluation over $\sim$600 samples from the AVRT-20K validation set: (1) *individual evaluation*, where the judge independently classifies each model's trace into one of three categories: *Accurate* (observations are factually correct and reasoning is sound), *Missing Info* (correct answer but omits important evidence), or *Hallucinating* (contains fabricated observations). The judge also reports per-modality grounding flags; and (2) *head-to-head comparison*, where the judge receives both traces simultaneously and selects a winner based on factual accuracy, completeness, logical soundness, and groundedness. The comparison prompt is shown in Figure A.4.

Table A.1 and Figure A.1 summarize the results. In head-to-head comparison, AVRT traces are preferred in **51.2%** of cases versus 38.5% for Qwen3-Omni-Thinking (10.3% ties). This is particularly notable because AVRT traces are generated by composing single-modality teachers, none of which has multimodal understanding on its own, while Qwen3-Omni-Thinking is a natively multimodal reasoning model with end-to-end audio-visual processing. In individual evaluation, both models achieve similar accuracy ($\sim$88%) and hallucination rates ($\sim$11%), confirming that AVRT's composed traces are on par with a native multimodal reasoning model. The most notable difference is in audio grounding: AVRT traces are grounded in audio content 93.4% of the time, compared to 81.3% for Qwen3-Omni-Thinking, suggesting that our explicit audio teacher stage produces more faithful acoustic descriptions. Qwen3-Omni-Thinking is slightly better at visual grounding (94.6% vs. 93.4%), but it exhibits a higher rate of *Missing Info* (2.0% vs. 0.3%), indicating that it sometimes omits modality-specific evidence.

Hallucination Source Analysis. To understand the origins of hallucinations in the AVRT pipeline, we conduct a follow-up analysis on the 66 traces (11.1%) flagged as hallucinating. Using Gemini-3.1-Flash-Lite as a judge, we present the original video and audio alongside the individual teacher traces and the final merged trace, and ask the model to attribute each hallucinated claim to its source: the audio teacher, the vision teacher, the merger model, or both teachers jointly (see Figure A.5 for the full prompt). The majority of hallucinations originate from both teachers contributing false claims (6.7% of all samples), indicating that cross-modal consistency issues in the teacher outputs are the primary driver. Audio-only and vision-only teacher hallucinations account for 1.7% and 1.5% respectively. Notably, the merger introduces novel hallucinations in only 1.2% of cases, confirming that it is largely faithful to its inputs and that improving individual teacher accuracy would address most hallucination issues. Figure A.2 presents qualitative examples from each hallucination source category.

Figure A.3 presents qualitative examples from each category. The *Accurate* examples demonstrate genuine cross-modal reasoning: the sewing example integrates the rhythmic machine sounds with the visual identification of a Hello Kitty sewing machine, while the hail example combines visual observation of white objects on the ground with the characteristic audio of hail striking surfaces. The *Hallucinating* examples illustrate typical reasoning errors: the kayak trace inaccurately describes "two paddles" and "oars" instead of a single double-bladed paddle, while the forest trace fabricates a "distant waterfall" sound despite correctly identifying the woodland setting visually. The *Missing*

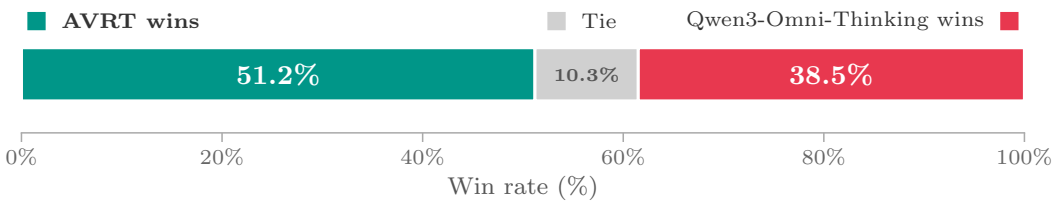


Figure A.1: **Head-to-head reasoning trace comparison on** ∼600 **AVRT-20K validation samples.** Gemini-3.1-Flash-Lite judges AVRT traces as better in **51.2%** of matchups vs. 38.5% for Qwen3-Omni-Thinking, based on factual accuracy, completeness, and groundedness. The evaluation prompt is provided in Figure A.4.

Table A.1: **Individual trace evaluation on** ∼600 **AVRT-20K validation samples.** Both models achieve comparable accuracy (∼88%) and hallucination rates (∼11%). AVRT achieves substantially higher audio grounding (93.4% vs. 81.3%) and lower missing-info rate (0.3% vs. 2.0%). Best results in **bold**. The gray sub-rows decompose AVRT's 11.1% hallucination rate by source; the merger introduces novel hallucinations in only 1.2% of cases.

Metric	AVRT	Qwen3-Omni
Accurate	**88.5%**	87.1%
Hallucinating	11.1%	10.9%
↪ *Hallucination source (AVRT):*		
Both teachers	6.7%	
Audio teacher only	1.7%	
Vision teacher only	1.5%	
Merger introduced	1.2%	
Missing Info	**0.3%**	2.0%
Audio Grounded	**93.4%**	81.3%
Visual Grounded	93.4%	**94.6%**

Info example shows the model correctly identifying a civil defense alarm from visual context (a Russian news broadcast) but failing to describe the actual alarm sound in its reasoning.

A.2 Model Behavior Under Missing Modalities

To investigate whether AVRT hallucinates modality-specific information when one modality is unavailable, we evaluate the model with masked inputs. We test the model in three conditions: (1) both audio and visual inputs provided (standard setting), (2) only video frames with silent audio, and (3) only audio with blank video frames. This experiment examines whether the model acknowledges missing information or fabricates details about the absent modality. The evaluation uses the same OmniBench test set.

Table 4 in the main paper shows that performance degrades by 14.4 points when only audio is available and by 10.5 points when only video is available, demonstrating that AVRT appropriately relies on both modalities. Figure A.6 provides a qualitative example comparing reasoning traces across three conditions. When both modalities are provided, the model explicitly integrates evidence using phrases like "From the audio..." and "From the visual inspection..." to arrive at the correct answer. When only video is available, the model constrains reasoning to visual observations ("Upon reviewing the image...") without fabricating audio content, though this leads to an incorrect prediction due to missing acoustic cues. Similarly, with audio-only input, the model focuses exclusively on sounds without hallucinating visual details like clothing or positions. While encouraging, this behavior could be sensitive to prompting strategies. The model may occasionally reference missing modalities due to strong pretraining priors.

A.3 Control Experiments: Role of Reasoning Content

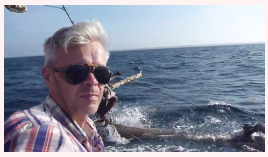
To validate that AVRT's improvements stem from the actual reasoning content in the composed traces, rather than from additional training data, learning the output format, or simply seeing correct

Audio Teacher: "What are the people doing?" GT: A



Pred: A (take honey) ✓
Audio: "The audio features the buzzing of bees and honeycomb being handled."
Vision: Correctly identifies hands scraping honey from a honeycomb into a container.
Merged: "The audio reveals buzzing bees. . . the dripping sound corresponds with the visual of honey flowing."
Judge: No bee buzzing or dripping in the audio. Vision correct.

Vision Teacher: "Source of sound?" GT: B (wind)



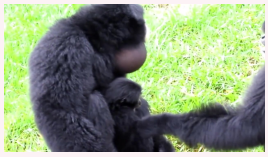
Pred: B ✓
Audio: Correctly identifies the sound of wind.
Vision: "A person navigating an outrigger canoe. . . absence of sails. . . the individual using a paddle."
Merged: "The individual using a paddle. . . the absence of a sail makes it less likely to hear wind. . . "
Judge: Video shows a sail and the boat under wind power. Vision misidentified the scene; merger incorporated false observations.

Both Teachers: "What are the people doing?" GT: D



Pred: D (eat fruit) ✓
Audio: "The audio features the sound of someone eating fruit."
Vision: "The child is holding a green, round object. . . consuming fruit."
Merged: "Audio: eating fruit. . . Visual: child holding a green object. . . eating fruit."
Judge: Child is eating a raw onion, not fruit. Audio fabricated eating sounds; vision misidentified onion as fruit.

Merger: "What are you calling?" GT: A (gibbon)



Pred: A ✓
Audio: Correctly identifies gibbon vocalizations.
Vision: Identifies a primate with dark fur (but incorrectly claims a prehensile tail).
Merged: "This creature is seen swinging from branches. . . the calls are unmistakably those of a gibbon."
Judge: Gibbon is sitting on grass, never swinging. Merger fabricated this—neither teacher mentioned it.

Figure A.2: Qualitative examples of hallucination sources in the AVRT pipeline. Each box represents a different source category: **Audio Teacher** (top-left), the audio teacher fabricated bee buzzing sounds not present in the audio; **Vision Teacher** (top-right), the vision teacher missed a visible sail and claimed the person was paddling; **Both Teachers** (bottom-left), the audio teacher fabricated eating sounds and the vision teacher misidentified a raw onion as fruit; **Merger** (bottom-right), the merger invented branch-swinging behavior absent from both teacher traces. Hallucinated claims are underlined. Despite hallucinations, all examples reach the correct answer.

answers, we conduct a series of control experiments. Each control ablates the SFT data composition while keeping all other training hyperparameters fixed, isolating the contribution of specific pipeline components.

We evaluate three controls against the baseline and the full AVRT pipeline:

1. *Answer-only SFT*: the model is fine-tuned on the AVRT-20K training samples containing only the correct answer letter (*e.g.* "A"), without reasoning traces and without subsequent RL.
2. *Answer-only SFT + RL*: same answer-only SFT followed by RL, testing whether RL can compensate for the lack of reasoning supervision.
3. *Format-only SFT + RL*: SFT on traces with the correct `<think>...</think>` `<answer>...</answer>` structure and the correct answer letter, but generic, non-informative filler text (*e.g.* "Let me analyze this question carefully. . . ") instead of actual cross-modal reasoning, followed by RL.

Accurate: "What are the people doing?" GT: D (sew clothes)



Pred: D ✓

The audio clearly captures the rhythmic sounds of a sewing machine in operation. Visually, the video displays a Brother Innov-is sewing machine with Hello Kitty branding... the needle is visible piercing through fabric held by the presser foot...

Accurate: "What happened in the video?" GT: D (hail)



Pred: D ✓

The visual component shows small white round objects on the ground... The audio captures heavy rainfall along with distinct sounds of hail hitting surfaces... intermittent shouts in the background might be people reacting to the hailstorm...

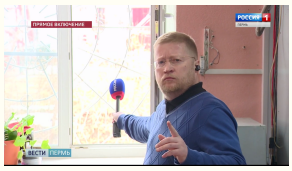
Accurate: "What's driving in the video?" GT: D (train)



Pred: D ✓

The video displays a large, multi-sectioned vehicle that moves along railroad tracks with people boarding through a staircase. The audio clearly emits the sounds associated with a train, such as the rhythmic clacking of wheels and the deep rumble of its engine...

Missing Info: "What is the main source of sound?" GT: D (civil defense alarm)



Pred: D ✓

The image depicts a news report with the logo of "Rossiya 1" indicating a live broadcast... the most plausible sound for a news broadcast discussing civil defense is the civil defense alarm...

Judge: Correct answer, but fails to describe the actual alarm sound heard in the audio.

Hallucinating: "What are the people doing?" GT: B (row-boat)



Pred: B ✓

The audio has clear sounds of water being displaced by oars... visuals showing a person maneuvering a small boat using two paddles...

Judge: Video shows a kayak with a single double-bladed paddle, not oars or two paddles.

Hallucinating: "Where does this video take place?" GT: D (in the woods)



Pred: D ✓

The video showcases tall, thin deciduous trees that appear bare... the audio provides a significant clue: there is a noticeable sound of a distant waterfall...

Judge: Correctly identifies forest visually, but hallucinates the waterfall sound.

Figure A.3: Qualitative examples from the automated reasoning trace quality evaluation. **Green**: accurate traces with correct cross-modal reasoning. **Red**: traces with hallucinated details (underlined), though still reaching the correct answer. **Yellow**: correct answer but missing explicit audio analysis. Color highlights in traces: audio and visual evidence.

Table A.2 summarizes the results across three benchmarks.

Answer-only SFT. The answer-only SFT model produces clean single-letter outputs (99.6% well-formed on OmniBench) but accuracy degrades versus the baseline on OmniBench and MMAR (-4.2 and -15.7 points), despite a gain on DailyOmni ($+4.0$). The model learns to map inputs directly to answer letters in the AVQA domain, bypassing any reasoning process, which leads to superficial pattern matching that does not consistently generalize to other benchmarks.

Answer-only SFT + RL: output collapse. Adding RL on top of the answer-only SFT leads to severe output collapse. On OmniBench, accuracy drops to 12.5%, with 48% of outputs being invalid. This collapse arises from a conflict between the SFT initialization and the RL objective: the SFT

Head-to-Head Trace Comparison Prompt (Gemini-3.1-Flash-Lite)

```
You are an expert judge comparing two audio-visual reasoning
traces.
You will be given: 1. A video file (with its original audio
track) 2. A separate audio file extracted from the same video 3.
A multiple-choice question about the video 4. The answer choices
5. The ground-truth correct answer 6. Two reasoning traces (Trace
A and Trace B) generated by different AI systems
Your task is to carefully watch the video, listen to the audio,
read both reasoning traces, and determine which trace is better
based on these criteria:
Factual Accuracy --- Are the observations about the video and audio
content correct?
Completeness --- Does the trace address both audio and visual
evidence?
Logical Soundness --- Is the reasoning chain coherent and
well-structured?
Groundedness --- Are claims about audio and visual content
supported by the actual media?
[Question, Answer Choices, Ground-Truth Answer, Trace A, and Trace
B are provided]
Respond with a JSON object with fields: ``winner'' (``A'',
``B'', or ``Tie''), ``confidence'' (0.0--1.0), ``explanation'',
``trace_a_verdict'' (ACCURATE / MISSING_INFO / HALLUCINATING),
``trace_b_verdict'', ``trace_a_audio_grounded'' (boolean),
``trace_a_visual_grounded'' (boolean), ``trace_b_audio_grounded''
(boolean), ``trace_b_visual_grounded'' (boolean).
```

Figure A.4: The prompt used for head-to-head comparison of reasoning traces. The judge model (Gemini-3.1-Flash-Lite) receives the original video, audio, and two reasoning traces from different systems (AVRT and Qwen3-Omni-Thinking, with randomized assignment to A/B positions). It selects a winner based on factual accuracy, completeness, logical soundness, and groundedness, while also returning per-trace quality verdicts and per-modality grounding flags.

Table A.2: **Control experiments isolating the role of reasoning trace content.** All models use Qwen2.5-Omni-3B as the base. "Answer-only SFT" trains on 20K samples with correct answer letters only (no reasoning traces). †"Format-only" uses `<think>`generic filler`</think><answer>`correct letter`</answer>` traces that match AVRT's output structure but contain no informative reasoning. Deltas are relative to the baseline. Neither answer-only nor format-only training approaches AVRT, confirming that the actual reasoning content in the composed traces is the key driver of performance.

Configuration	RL	OmniBench	DailyOmni	MMAR
Baseline (Qwen2.5-Omni-3B)	×	50.2	43.1	53.7
Answer-only SFT	×	46.0 (−4.2)	47.1 (+4.0)	38.0 (−15.7)
Answer-only SFT + RL	✓	12.5 (−37.7)	28.9 (−14.2)	45.2 (−8.5)
Format-only SFT + RL†	✓	21.3 (−28.9)	29.9 (−13.2)	22.1 (−31.6)
AVRT (Ours)	✓	**56.3** (+6.1)	**49.2** (+6.1)	**57.3** (+3.6)

stage trains the model to output bare answer letters (*e.g.* "B"), but the RL format reward pushes toward structured `<think>...</think><answer>...</answer>` outputs. The model attempts to adopt the tag structure but, lacking any prior exposure to reasoning traces, cannot produce coherent content within it. Among the 52% of valid outputs, most are empty thinking blocks (*e.g.* `<think></think><answer>D</answer>`), where the model learned the tags but has nothing to put inside them. The remaining 48% are broken fragments: 67% of invalid outputs consist

Hallucination Source Analysis Prompt (Gemini-3.1-Flash-Lite)

```
You are an expert judge performing a detailed hallucination source
analysis on an audio-visual reasoning trace.
An AI system produced a final reasoning trace by merging two
teacher traces:  (1) an audio teacher that analyzed only the audio,
and (2) a vision teacher that analyzed only the video frames.  A
merger model then combined these into a single unified reasoning
trace.  A previous evaluation flagged the final merged trace as
HALLUCINATING. Your job is to determine WHERE the hallucination
originates.
You are given:  a video file (with audio), a separate audio file,
a multiple-choice question with answer choices, the ground-truth
answer, the audio teacher's reasoning, the vision teacher's
reasoning, the final merged trace, and the previous judge's
explanation.

Categories for hallucination source:

AUDIO_TEACHER --- The hallucinated content traces back to a false
claim in the audio teacher's reasoning.

VISION_TEACHER --- The hallucinated content traces back to a false
claim in the vision teacher's reasoning.

MERGER --- The hallucinated content was NOT present in either
teacher trace.  The merger model invented or distorted information.

BOTH_TEACHERS --- Both teachers contributed hallucinated content
that ended up in the final trace.


[Question, Answer Choices, Ground-Truth Answer, Audio Teacher's
Reasoning, Vision Teacher's Reasoning, Final Merged Trace, and
Previous Evaluation Context are provided]
Respond with a JSON object with fields:  ``hallucination_source''
(AUDIO_TEACHER, VISION_TEACHER, MERGER, or BOTH_TEACHERS),
``confidence'' (0.0--1.0), ``hallucinated_claims'' (list of
false claims), ``source_per_claim'' (source for each claim),
``audio_teacher_accurate'' (boolean), ``vision_teacher_accurate''
(boolean), ``merger_faithful'' (boolean), ``explanation'' (2--4
sentences).
```

Figure A.5: The prompt used for hallucination source analysis. The judge model (Gemini-3.1-Flash-Lite) receives the original video, audio, both teacher traces, and the merged trace flagged as hallucinating. It attributes each hallucinated claim to its source: the audio teacher, vision teacher, merger, or both teachers.

of an orphaned </answer> closing tag with no content, 29% start with <think> but never close properly, and 4% are sequences of closing tags only (*e.g.* </think></answer>). Even among valid outputs, accuracy is poor because the model never learned cross-modal reasoning patterns. The model receives conflicting training signals: SFT encourages bare answer letters while RL rewards structured tag formatting.

Format-only SFT + RL: template reasoning. The format-only control exhibits a different but similarly detrimental failure mode. Unlike the answer-only variant, this model produces well-structured, plausible-looking outputs: 85% of OmniBench responses are validly formatted with numbered reasoning steps and modality-by-modality analysis. However, the content is generic template filler rather than genuine observation. A typical output follows the pattern: *"The audio does not provide any specific information about the men's actions. The video shows two men sitting at a table. Based on the available evidence…"* The model learned the *form* of reasoning perfectly (structured steps, modality labels, concluding statements) but applies the same template regardless of the ac-

Question: What are the men doing?
Options: (A) The man in jeans is taking notes from the newspaper. (B) The man in purple is reading the newspaper. (C) The man in jeans is playing a crossword puzzle. (D) The man on the table is doing a crossword puzzle.
Ground Truth: C

Both Audio + Video

Prediction: C
Reasoning:
From the audio, there is a clear sound of someone speaking, which suggests that the men are engaged in an activity that involves verbal communication...
From the visual inspection, I observe two men sitting at a table. One of them is wearing jeans and the other is in purple. Both appear to be focused on a piece of paper in front of them. The presence of a newspaper on the table further supports the idea...
Option C mentions the man in jeans is playing a crossword puzzle...

Video Only

Prediction: B
Reasoning:
Upon reviewing the image, I notice that there are two men present. The man in purple is holding a newspaper and appears to be reading it. The man in jeans is sitting next to him, but his attention is not clearly focused on the newspaper.
Given the options, the most plausible scenario is that the man in purple is reading the newspaper, while the man in jeans is engaged in another activity...

Audio Only

Prediction: B
Reasoning:
Upon listening to the audio, I noticed a clear and distinct sound of a newspaper being handled, specifically the sound of pages being turned. This sound is consistent with someone reading a newspaper.
Additionally, there was a brief mention of a crossword puzzle, but it was not the main focus of the audio. The primary activity described was reading a newspaper, which aligns with option B...

Figure A.6: Example of model reasoning traces when different modalities are masked. With both modalities (left), the model integrates audio and visual evidence. With video only (center), reasoning focuses on visual observations without hallucinating audio content. With audio only (right), the model constrains itself to acoustic features without fabricating visual details. Color highlights indicate modality-specific reasoning: audio and visual.

tual audio-visual content. The 15% of invalid outputs follow the same template structure but are truncated before the closing `</answer>` tag, suggesting the model generates increasingly verbose filler that exceeds the maximum output length.

This template reasoning explains the progressive performance collapse during RL training. Using the same experimental setup as the main paper, the accuracy reward signal is too noisy to learn from (since the model is not actually reasoning about the content), while the format and length rewards provide consistent positive signal for producing longer, well-structured text. The model thus over-optimizes for surface-level reward components while accuracy degrades to 21.3% on OmniBench and 22.1% on MMAR.

Implications. AVRT achieves 56.3% on OmniBench, 49.2% on DailyOmni, and 57.3% on MMAR, outperforming all controls by a large margin. The contrast between AVRT and the format-only control is particularly informative: both use the same output format, but only AVRT contains genuine cross-modal reasoning. None of the alternative SFT configurations could replicate the performance gains observed with AVRT, suggesting that the actual reasoning content in the composed traces is the key driver. In our setting, RL alone was not sufficient to compensate for the absence of reasoning supervision during SFT: without that foundation, RL either collapsed the model's outputs (answer-only) or optimized surface features while degrading accuracy (format-only).

B AVRT Additional Information

B.1 AVRT Dataset Construction Prompts

In this section, we provide the prompts used for both teacher models (Kimi-VL-Thinking and Audio Flamingo 3 (*think*)), and the merger model (Qwen2.5-14B-Instruct).

Visual Teacher Prompt. We design the visual teacher prompt to elicit detailed temporal reasoning from video frames. As shown in Figure B.1, the prompt instructs Kimi-VL-Thinking to analyze 8 evenly distributed frames from each video, providing both a comprehensive visual description and explicit reasoning that considers temporal progression. The prompt includes a concrete example that demonstrates the expected response format, encouraging the model to describe what it

Visual Teacher (Kimi-VL-Thinking) Prompt

```
You are an intelligent vision agent.  I will provide you with 8
representative frames from a video (evenly distributed across
the video duration) and a question about the video content in
MCQ format.  You need to first provide a thorough description
of what you're seeing across these video frames, then add
Chain-of-Thought-type reasoning to analyze the visual content, and
finally provide your answer.  Here is an example:
Input Question:  What type of activity is happening in this video?
Choose one among the following options:(A) Crime thriller scene (B)
Documentary narration\n(C) Romantic comedy scene\n(D) Action movie
or racing scene\n
Expected response format:\n
Visual Description:  Across these video frames, I can see a
progression of high-speed chase scenes with vehicles moving
rapidly through an urban environment.  The frames show consistent
dynamic motion, intense lighting, and what appears to be an ongoing
action sequence with cars and possibly motorcycles.  The temporal
progression across frames reveals the continuous high-energy nature
of the content.\n
Reasoning:  Based on the consistent high-speed vehicle movement
visible across multiple frames, the sustained dynamic camera
work, intense lighting throughout the sequence, and the overall
action-oriented visual elements that persist across the video
timeline, this content would be most suitable for action-focused
scenarios that require high-energy sequences.  The visual elements
strongly suggest this is an action movie or racing scene rather
than other genres like crime thriller, documentary, or romantic
comedy.\n
Answer:  (D) Action movie or racing scene\n
Follow this format:  provide a detailed visual description
analyzing the temporal progression across frames, then your
reasoning considering the full video context and evaluating each
option, then the final answer.  For answers that do not require
complex reasoning (e.g., for a question like "What color is the
object?" or "How many people are in the image?" where the answer
is direct), still provide the visual description but keep the
reasoning brief.\n
Here is the input question:
```

Figure B.1: The prompt used for the visual teacher model (Kimi-VL-Thinking). The prompt guides the model to provide temporal visual analysis across video frames, followed by explicit reasoning and a final answer.

observes across frames, reason about the visual evidence, and arrive at the correct answer. For questions requiring straightforward visual identification, the prompt allows for brief reasoning while still maintaining the descriptive component.

Audio Teacher Prompt. For the audio teacher, we adopt the prompts from the Audio Flamingo 3 paper to ensure optimal performance and maintain consistency with the original model's training methodology. As illustrated in Figure B.2, the audio prompt follows a similar structure to the visual prompt, instructing the model to provide a thorough audio description, evaluate each answer option with explicit reasoning, and then provide the final answer. The prompt includes an example that demonstrates how to analyze audio characteristics (e.g., beat, vocals, production style) and map them to appropriate contexts. This approach ensures that the audio teacher generates reasoning traces that complement the visual analysis.

Merger Prompt. The merger model guides the combination of reasoning traces into a unified multimodal analysis. Figure B.3 shows the prompt used for Qwen2.5-14B-Instruct, which receives the question along with both the audio and visual analyses generated by the teacher models. The merger does not receive the ground-truth answer; it must derive the correct answer solely from the

Audio Teacher (Audio Flamingo 3 *think*) Prompt

```
You are an intelligent audio agent.  I will provide you with an
audio and a question about the audio in MCQ format.  You need to
first provide a thorough description of what you're hearing in
the audio, then add Chain-of-Thought-type reasoning to analyze the
audio content and evaluate each option, and finally provide your
answer.  Here is an example:
Input Question:  What type of soundtrack would this piece be most
suitable for?  Choose one among the following options:(A) Crime
thriller movie (B) Documentary narration\n(C) Romantic comedy
movie\n(D) Futuristic movie or car racing video game\n
Expected response format:\n
Audio Description:  This audio features a high-energy electronic
track with a driving beat, synthesized sounds, and confident
rap vocals.  The lyrics mention themes of speed and success,
including phrases like 'living automatic' and references to new
cars.  The production has a modern, polished sound with heavy use
of electronic elements.\n
Reasoning:  Based on the driving beat, confident rap vocals,
mentions of speed and success, and overall high-energy modern
production with electronic elements, this piece would be most
suitable for high-octane, modern scenarios that require energetic
background music.  Evaluating the options:  (A) Crime thriller
movies typically use more suspenseful, darker soundtracks; (B)
Documentary narration usually requires more neutral, informative
background music; (C) Romantic comedy movies generally feature
lighter, more melodic soundtracks; (D) Futuristic movies or
car racing video games would benefit from exactly this type of
high-energy electronic music with themes of speed and technology.\n
Answer:  (D) Futuristic movie or car racing video game\n
Follow this format:  provide a detailed audio description first,
then your reasoning that evaluates each option, then the final
answer.  For answers that do not require complex reasoning (e.g.,
for a question like "Who performs the vocals in this song?" or
"What primary instrument is featured in this piece?" where the
answer is direct), still provide the audio description but keep
the reasoning brief.\n
Here is the input question:
```

Figure B.2: The prompt used for the audio teacher model (Audio Flamingo 3 *think*). The prompt instructs the model to describe audio characteristics and provide explicit reasoning that evaluates each answer option.

teacher traces. The merged reasoning is formatted within <think> tags, followed by the final answer in <answer> tags.

B.2 Additional Dataset Statistics

Figure B.4 provides additional statistics for the AVRT-20K dataset. The question type distribution and modality requirements are derived directly from the original AVQA annotations. The question type distribution closely mirrors that of the full AVQA dataset, with "Which" questions being most prevalent (45.2% in training), followed by "Come From" (30.9%), "Happening" (15.5%), and "Where" (8.0%), confirming that our filtered subset preserves the original distribution. The dataset has balanced answer choices with each option (A–D) appearing roughly 25% of the time, which is critical for mitigating selection bias in LLM training data. According to the AVQA modality annotations, 99% of the questions are labeled as requiring both audio and visual information, and only 1% as answerable from a single modality. Note that our filter keeps only samples that both single-modality teachers answer correctly, so the retained questions were answerable by these models from either modality alone; the annotation therefore reflects the intended cross-modal nature of the questions rather than strict necessity.

Merger (Qwen2.5-14B-Instruct) Prompt

```
You are an intelligent multimodal agent. I will provide you with
a question in MCQ format, along with separate audio and visual
analyses from specialized models. Your task is to merge these
analyses into a coherent reasoning chain that integrates both
modalities to arrive at the correct answer.
Question: {question}{formatted_choices}
Audio Analysis: {audio_reasoning}
Visual Analysis: {vision_reasoning}
Instructions:
- Format your merged reasoning inside <think>... </think>
- At the end, output your final answer (just the letter, e.g., A,
B, C, or D) inside <answer>... </answer>
- Write sentences that integrate both audio and visual evidence
- Explain how the audio and visual clues work together to lead you
to the conclusion
- Make the explanation thorough but succinct

Combined Analysis:
```

Figure B.3: The prompt used for the merger model (Qwen2.5-14B-Instruct). The prompt guides the model to integrate audio and visual analyses into a coherent multimodal reasoning trace formatted with <think> and <answer> tags.

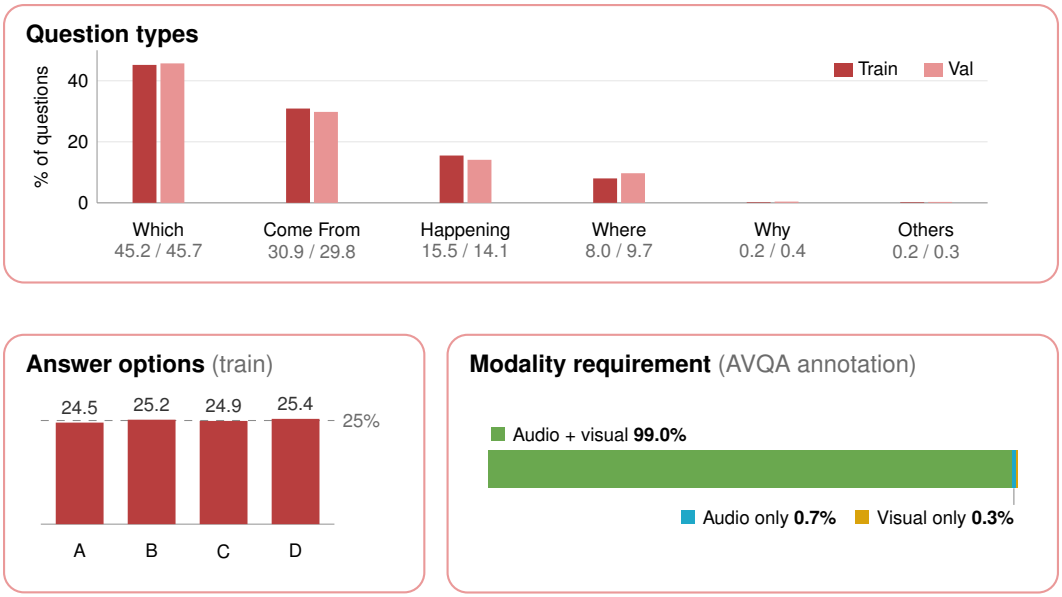


Figure B.4: Additional statistics for the AVRT-20K dataset, derived from the original AVQA annotations. **Top:** Question type distribution across training and validation splits (train / val percentages below each type) closely mirrors the full AVQA dataset, with "Which" (45.2%) and "Come From" (30.9%) dominating, confirming that our dual-teacher filtering preserves the original distribution. **Bottom left:** Answer option distribution (training split) is balanced at $\sim$25% per option (A–D, dashed line), mitigating selection bias. **Bottom right:** Modality requirements according to the AVQA annotations, which label 99% of questions as requiring both audio and visual information.

C Limitations and Future Work

One limitation of AVRT is that the quality of the generated multimodal traces is bounded by the factual reliability of the individual single-modality teachers. Our hallucination-source analysis suggests that most trace errors originate in the teachers rather than in the merger, and the pipeline may underperform if stronger unimodal teachers are not available for a given modality. Additionally, our

SFT data is derived from a single source domain (AVQA), which may constrain the diversity of reasoning patterns seen during training. The generality of the approach to other modality combinations also remains to be established experimentally. Future work could address the teacher-quality bottleneck through uncertainty-aware merging and explicit trace verification, apply the pipeline to more diverse datasets to characterize SFT data scaling, and extend the framework to additional modality combinations, such as incorporating a speech-specific teacher for three-way merging. Iterative self-improvement, where the trained student serves as a teacher for subsequent rounds of trace generation, could also bootstrap higher-quality traces, since the student already outperforms individual teachers on several benchmarks.