AVRT: Audio-Visual Reasoning Transfer through Single-Modality Teachers

Edson Araujo[1] Saurabhchand Bhati[2] M. Jehanzeb Mirza[2] Brian Kingsbury[3]
Samuel Thomas[3] Rogerio Feris[3,4] James R. Glass[2] Hilde Kuehne[4,5]
[1] University of Tübingen, [2] MIT, [3] IBM Research, [4] MIT-IBM Watson AI Lab, [5] Tuebingen AI Center

Abstract

Recent advances in reasoning models have shown remarkable progress in text-based domains, but transferring those capabilities to multimodal settings, and allowing reasoning e.g. 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 an 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. Our evaluation shows that the proposed pipeline based on generated multimodal traces for SFT allows models to achieve superior performance on various datasets, i.a. OmniBench, Daily-Omni, and MMAR, comapred to RL alone, establishing a new pipeline for training of audio-visual reasoning models in particular and opening new ways to train reasoning models on multimodal data in general. [†]

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 [1, 4, 14, 18, 27].

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 [10] and DeepSeek-R1 [5]. These advances have been significantly driven by reinforcement learning techniques [19]. These reasoning capabilities have been successfully extended to inputs beyond text, such as vision-text models [6, 9] and audio-text models [8, 23, 25], 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 [7] or try to approach the problem by extending reinforcement learning formulations e.g. by improved credits assignment or by context summarization [11, 29].

This paper proposes a 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 a single cross-modal reasoning traces, as 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 multimodal reasoning that explicitly correlates information across audio and visual channels. By using an LLM as a merging interface for different the 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 convert it into a consistent target format. The resulting audio-visual traces are then in the first step 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 [19] objective.

We evaluate the proposed approach on four challenging

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

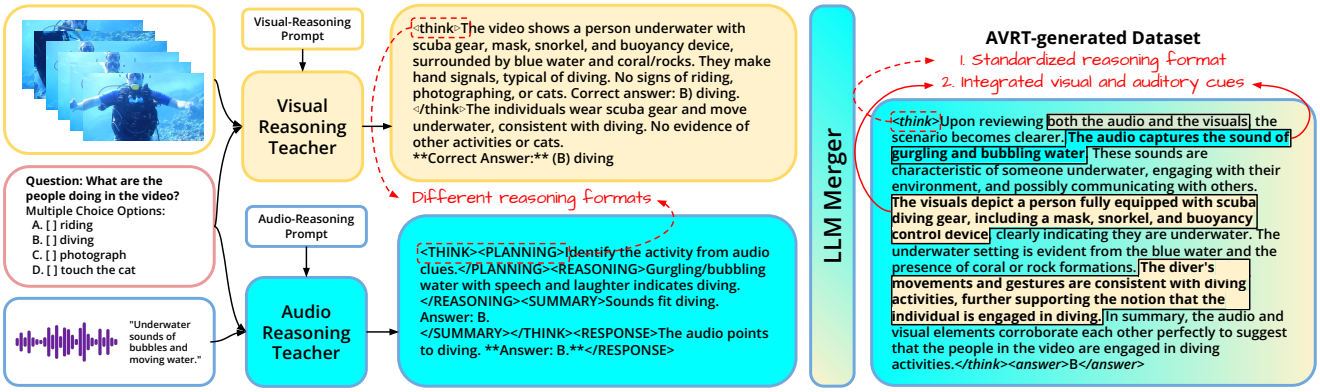

Figure 1. **Top:** Overview of the AVRT pipeline: We first generate reasoning chains from single-modality teacher models that are prompted in the format they were optimized for and, second, leverage an LLM merger as an interface between the teacher models and the resulting reasoning chain to aggregate 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.

datasets: First, a zero-shot vision-audio downstream datasets, we consider OmniBench [13] and DailyOmni [32], second, we evaluate on AVQA [28] as in-domain validation dataset, and finally on MMAR [17] as audio-only downstream task. Using only a 3B-parameter student, we achieve improved performance compared to other 3B audio-visual reasoning models and competitive performance to 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 to aggregate multiple single-modality reasoning traces into integrated multi-modal reasoning traces, creating high-quality audio-visual reasoning data without expensive annotation. 2) We are the first to train a multimodal reasoner based on this type of composed reasoning data, demonstrating that cross-modal understanding can emerge from the combination of single-modality expertise. 3) We conduct an extensive evaluation on audio-visual benchmarks, achieving state-of-the-art results that compete with larger models through our approach.

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 [2] 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 [20] proposes specifically targeting temporal referential dialogue in untrimmed videos, introducing the Audio-Visual Tokens Interleaver for temporal alignment. VideoLLaMA 2 [1] 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 [15] 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 [27] 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 [14] 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.

Audio-Visual Reasoning. Audio-visual reasoning with large language models has seen rapid advancements. Daily-Omni [32] introduces a dedicated Audio-Visual QA dataset, accompanied by Daily-Omni-Agent, a training-free agent that utilizes an open-source visual language model (VLM), an audio language model (ALM) and an automatic speech recognition (ASR) model to establish a baseline for this benchmark. Recent work has increasingly leveraged reinforcement learning: EchoInk-R1 [26] proposes a framework using GRPO for structured cross-modal reasoning, demonstrating emergent self-corrective reasoning behaviors. HumanOmniV2 [29] addresses shortcut problems by requiring explicit context summarization before reasoning, introducing context and logical rewards alongside IntentBench for understanding human intentions. Omni-R1 [31] tackles the trade-off between temporal coverage and spatial resolution through a two-system architecture with hierarchical rewards. AVATAR [11] presents an off-policy RL framework with

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	Video + 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

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

Temporal Advantage Shaping (TAS) for improved credit assignment during planning and synthesis stages. More recently, AURELIA [3] 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 present a more efficient paradigm, 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 a for 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 [28] can be considered one of the foundational audio-visual QA datsets with $57,335$ question-answer pairs from daily audio-visual activities requiring clues from both modalities. OmniInstruct [13] develops a comprehensive tri-modal reasoning dataset combining visual, audio, and textual resources, while MUSIC-AVQA [12] expands to musical performance with $45,867$ question-answer pairs across $9,288$ videos. AVQA-R1-6K [26] 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.

3. Methodology

In this paper, we derive audio-visual reasoning traces based on existing audio-visual question-answering pairs as e.g. provided by AVQA [28]. In this section, we first discuss the generation of audio-visual reasoning traces in Sec. 3.1 and our training procedure to leverage this data to perform 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. In the first stage, 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 and incorporating cross-modal dependencies.

Filtering. To make sure that we only use correct reasoning traces and do not introduce any 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 [22, 24]. As a result, we only keep a 20K subset of the original 40K AVQA data samples.

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)[19]. 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:

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

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

Our reward combines multiple components to enforce correctness, formatting, and reasoning quality.

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

The three components are defined as:

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

$$R_{format}(o_i) = \begin{cases} 1, & \text{if format is correct} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

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

$$R_{acc}(o_i) = \begin{cases} 1, & \text{if answer is correct,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

(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 $[w_{min}, w_{max}]$.

We optimize our reward function using Group Relative Policy Optimization (GRPO) [19], which computes advantage estimates through group-based comparisons of model outputs and applies policy gradient optimization with clipping and KL regularization.

Table 2. Statistics of the AVRT dataset showing sample counts, quality metrics, and distribution of question types and answer options across training and validation splits.

Metric	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%)

Question Type Distribution

Question Type	Train / Val (%)	Question Type	Train / Val (%)
Which	45.2 / 45.7	Where	8.0 / 9.7
Come From	30.9 / 29.8	Why	0.2 / 0.4
Happening	15.5 / 14.1	Others	0.2 / 0.3

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 [21] and Audio Flamingo 3 (*think*) [8] 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. Table 2 presents statistics for our AVRT-20K dataset. The final collection comprises $18,279$ training samples and 945 validation samples, all extracted from the original AVQA dataset. All samples achieve 100% reasoning format com-

pliance, ensuring consistent structure across the dataset. The thinking sections contain an average of 165.5 ± 33.9 tokens in the training set and 163.4 ± 32.5 tokens in the validation set, while answer sections are consistently single tokens, corresponding to the (A, B, C, D) options format. Videos maintain uniform duration of $\approx$10 seconds, with the main resolution being 1280×720.

The distribution of question types closely mirrors that of the original 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%) questions. This similarity demonstrates that our random sampling successfully encompasses the distributional characteristics of the original dataset, ensuring our subset maintains representativeness across different reasoning types and question categories.

RL Training. In the second phase, we train the resulting model on the full AVQA training set [28]. AVQA comprises $57,335$ question-answer pairs across $45,867$ unique audio-visual samples from daily activities. The dataset is split into $40,127$ training samples and $17,208$ validation samples. For the RL phase, we use the full training set to further improve the model's audio-visual reasoning capabilities through reinforcement learning, building upon the reasoning foundations established during the SFT phase.

4.2. Benchmark Datasets

We evaluate our model on four representative datasets that span different modality combinations to comprehensively assess cross-modal reasoning capabilities: DailyOmni (video+audio), OmniBench (image+audio), AVQA (video+audio, in-domain) and MMAR (audio-only) to examine potential overfitting on our training distribution.

DailyOmni [32] is a benchmark for evaluating multimodal large language models on real-life audio-visual scenarios that require joint reasoning across video, audio, and textual information. The dataset contains 684 videos and $1,197$ question-answer pairs (550 from 60-second videos, 647 from 30-second videos) covering all 11 YouTube categories to ensure diversity of topics, styles, and acoustic environments. The questions are deliberately designed to force integration of modalities, moving beyond simple perception to complex reasoning tasks that require understanding of concurrent multimodal phenomena including speech, music, and environmental sounds.

OmniBench [13] was designed to evaluate large language models' ability to integrate image, audio, and text inputs for cross-modal reasoning. The benchmark contains $1,142$ question-answer pairs organized into 8 task categories: Action & Activity, Story Description, Plot Inference, Object Identification & Description, Contextual & Environmental, Identity & Relationship, Text & Symbols, and Count & Quantity. Each sample includes multiple-choice questions with corresponding image and audio content, with audio clips averaging 9.22 seconds in duration.

AVQA [28] is a large-scale benchmark containing $57,015$ question-answer pairs across $45,867$ videos designed to evaluate models' ability to reason over both audio and visual content. The dataset features high-quality manual annotations and questions that specifically require integration of both modalities, making it well-suited for evaluating genuine cross-modal reasoning capabilities rather than single-modality shortcuts.

MMAR [17] is an audio-only reasoning benchmark designed to evaluate models' ability to perform complex reasoning tasks using solely auditory information. We include MMAR to assess how well our cross-modal training approach transfers to single-modality audio reasoning scenarios. The benchmark provides a controlled evaluation environment to understand whether the multimodal reasoning capabilities developed through our teacher aggregation methodology can effectively generalize to audio-only tasks.

4.3. Implementation Details

For all experiments, we use Qwen2.5-Omni-3B [27] as the base student model with frozen vision and audio modules. Fully 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, we set the optimal target length to $\mu = 100$ words with a standard deviation of $\sigma = 20$ words, and the target range for the bonus to $w_{min} = 200$ and $w_{max} = 300$ words.

4.4. Comparison to State-of-the-Art

Table 3 shows the performance of the proposed model in comparison to existing audio-visual reasoning models across four benchmark datasets. The 3B parameter model achieves strong performance, both in terms of absolute accuracy and relative improvement over our reproduced baseline. For the case oof zero-shot performance on DailyOmni and OmniBench, the model achieves 49.2% accuracy on DailyOmni, and 56.3% accuracy on OmniBench, outperforming 3B baselines and remaining competitive with 7B models. It further shows that the resulting model improves significantly by $+6.1$ percentage points (56.3% vs. 50.2%) over our reproduced baseline, whereas AVATAR-3B's reported gain over their baseline is $+3.4$ points (45.8% vs. 42.4%). Note that

Model	# Params	Reasoning	DailyOmni *(Video+Audio)*	OmniBench *(Image+Audio)*	AVQA† *(Video+Audio)*	MMAR *(Audio-Only)*
7B Audio-Visual Models						
Qwen2.5 Omni [27]	7B	×	44.0	44.2	-	56.7
Qwen2.5 Omni* [27]	7B	×	51.5	50.7	84.9	56.5
Echolnk [26]	7B	✓	46.2	46.5	-	-
Omni-R1 [31]	7B	✓	46.8	46.9	-	-
HumanOmni [30]	7B	✓	47.6	44.9	-	-
Ola-7B [14]	7B	×	52.3	45.3	-	-
AV-Reasoner [16]	7B	✓	**53.8**	48.3	-	-
AVATAR [11]	7B	✓	47.0	**49.1**	-	-
Modality-Specific Teachers						
Kimi-VL-Thinking [21]	-	✓	-	33.5	-	*N/A*
AF3 *(think)* [8]	-	✓	-	28.9	-	60.1
3B Audio-Visual Models						
Qwen2.5 Omni [27][11]	3B	×	42.9	42.4	-	-
AVATAR [11]	3B	✓	44.7 (+1.8)	45.8 (+3.4)	-	-
Qwen2.5 Omni* [27]	3B	×	43.1	50.2	88.3	53.7
AVRT (Ours)	3B	✓	**49.2** (+6.1)	**56.3** (+6.1)	**91.1** (+2.8)	**57.3** (+3.6)

Table 3. Comparison of audio-visual reasoning models on benchmark datasets. DailyOmni, OmniBench, and MMAR are tested in zero-shot mode without further finetuning. AVQA results are considered fine-tuned since the training dataset is derived from AVQA's training set. We report reproduced baseline results for Qwen2.5 Omni marked with *.

we could not replicate [11] directly, so we report gains relative to the respective Qwen2.5-Omni-3B baselines.

Next, we consider AVQA as a case of in domain-validation. The model here achieves 91.1% accuracy, showing a +2.8 percentage point improvement over the reproduced Qwen2.5-Omni-3B baseline (91.1% vs. 88.3%). While this improvement is more modest compared to other benchmarks, it is important to note that AVQA represents a fine-tuned evaluation scenario since the training dataset is derived from AVQA's training set.

For MMAR, an audio-only reasoning benchmark, the model achieves 57.3% accuracy, outperforming the baseline by +3.6 percentage points (57.3% vs. 53.7%). The improvement on MMAR validates that the audio reasoning capabilities developed through the teacher aggregation methodology generalize effectively beyond the multimodal training domain approaching that of the specialized audio teacher AF3 *(think)* (60.1%). This indicates that training with aggregated multimodal reasoning chains can also lead to strong single-modality capabilities as result of audio-visual learning.

Overall, the results show that reasoning-capable multimodal models are able to outperform their non-reasoning counterparts across parameter sizes, validating the importance of structured reasoning in audio-visual tasks.

4.5. Ablation Studies

Evaluation of SFT fine-tuning. We first asses 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 can be considered as an indication for reasoning transfer learning as the SFT dataset is composed mainly of questions that require both audio and vision (99.0%) (see supplementary material) and as both our supervised fine-tuning (SFT) with reasoning traces plus RL and the RL-only

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 4. Impact of training stages with audio-visual training data evaluated on single modality as well as on joint audio-visual performance. We evaluate using only audio (A), only vision (V), and both modalities (AV) on OmniBench.

Reasoning Chain Modality	OmniBench
No SFT (Baseline+RL)	52.3
Audio-Only SFT	52.5
Video-Only SFT	50.8
Audio-Visual SFT (Ours)	**56.3**

Table 5. Ablation on reasoning chain types for SFT. Models are fine-tuned based on the different reasoning chains, followed by AV-RL training. Our audio-visual reasoning chain aggregation outperforms single-modality chains.

baseline are trained solely on audio-visual inputs, without any 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, we evaluate models trained with: (1) no SFT (RL only, +2.1), (2) SFT on audio-only reasoning traces (+0.8), (3) SFT on video-only reasoning traces (+1.9), and (4) audio-visual aggregated traces (+6.1), all compared to the Qwen2.5-Omni 3B baseline (50.2%). 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. Both models improve over the Qwen2.5-Omni 3B baseline (50.2%), with Qwen2.5-14B-Instruct achieving 56.3% (+6.1 points) compared to 48.5% (-1.7 points) for Gemma3-12B-It. Notably, during training, we observe that the model fine-tuned with reasoning traces generated by Qwen2.5-14B-Instruct converged significantly faster to the multiple-choice

LLM Merger Model	Student Model	OmniBench
Gemma3-12B-It	Qwen2.5-Omni-3B	49.5
Qwen2.5-14B-Instruct	Qwen2.5-Omni-3B	**56.3**

Table 6. Ablation on merger models. Using an LLM merger with the same backbone as the student model enhances performance.

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

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

Model	Easy	Medium	Hard
Baseline	70.5	53.5	45.3
Ours	**76.9**	**59.2**	**51.8**

Table 8. Performance on different OmniBench difficulty subsets. Our model outperforms the base model on all difficulty levels.

question (MCQ) format compared to the Gemma3-based merger. This suggests that using a teacher model from the same architectural family as the student model facilitates more efficient knowledge transfer, as the student model does not need to adapt to a substantially different token distribution during fine-tuning.

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). This consistent improvement demonstrates that maintaining an sufficient reasoning trace length is effective for audio-visual reasoning.

Omnibench Subsets. To finally analyse the impact of the reasoning capabilities, we categorize OmniBench questions into difficulty subsets based on teacher-model performance: *easy* (both teachers correct, 64 questions), *medium* (one teacher correct, 456 questions), and *hard* (neither correct, 619 questions). As shown in Table 8, our model outperforms the Qwen2.5-Omni baseline across all difficulty levels: easy (76.9% vs. 70.5%, +6.4 points), medium (59.2% vs. 53.5%, +5.7 points), and hard (51.8% vs. 45.3%, +6.5 points). The *easy* subset shows the largest improvement but exhibits greater variability due to its smaller sample size. Overall, the predominance of hard questions (54%) further demonstrates that OmniBench is well-suited for evaluating cross-modal reasoning.

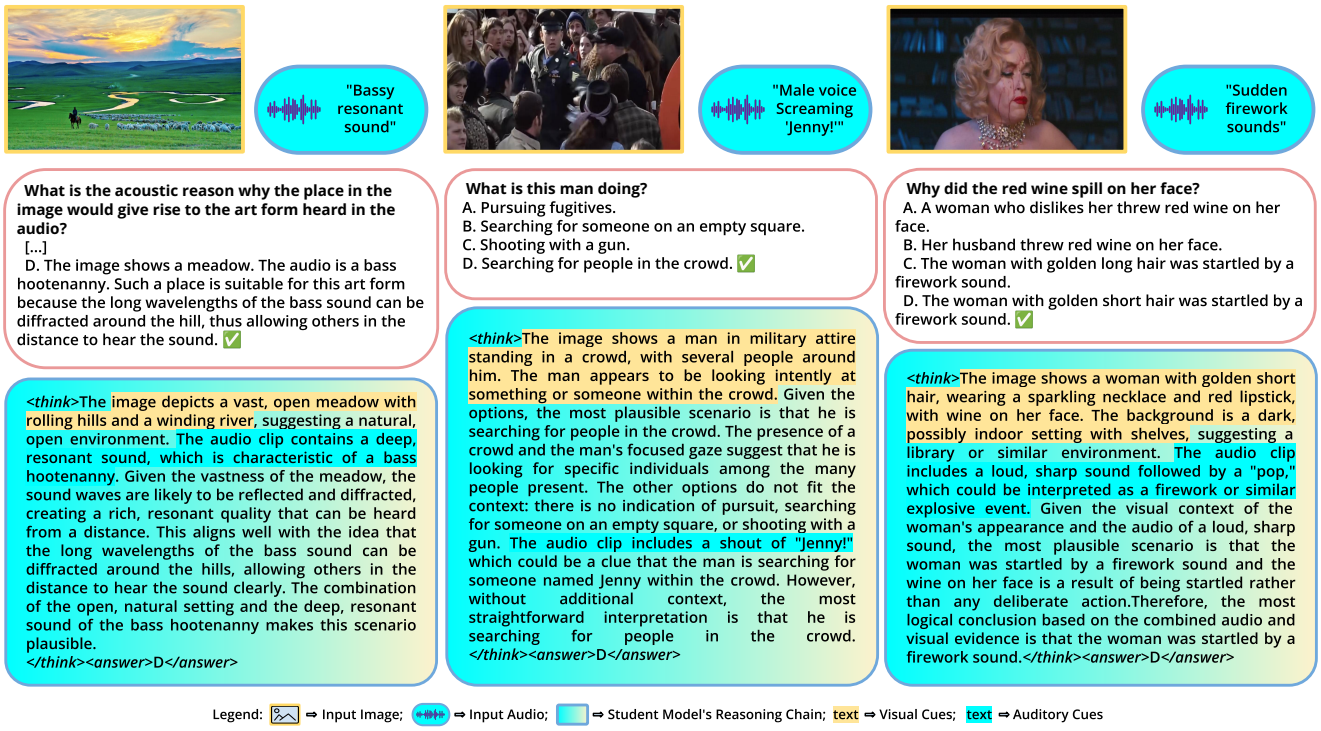

Figure 2. Qualitative results of the AVRT-trained model on OmniBench: It shows that the model trained on the respective AVRT-20K data is able to retrieve audio and visual information to answer the question, to combine the two sources of information, and to generate high-quality reasoning chains based on different cues in both modalities. *Best viewed in color and zoomed in.*

4.6. Qualitative Results

We finally provide qualitative results of reasoning traces produced by the student model after fine-tuning on the proposed AVRT reasoning traces on OmniBench in Fig. 2. The results show that with only SFT on the generated reasoning traces, the model is able to reason about the image and audio content to answer the question. In the first example, the model correctly associates the acoustic properties of a bass "hootenanny" with the open meadow environment, demonstrating understanding of how sound propagates differently in open versus enclosed spaces. The second example showcases more sophisticated multi-modal reasoning, where the model uses visual cues (people around the main character) to contextualize the audio (male voice calling "Jenny!") and correctly identifies the scenario as searching for people in a crowd among multiple plausible options. The third example illustrates the model's ability to connect temporal audio events (firework sounds) with visual evidence (wine spill on face). As shown in the figure, the model learns to incorporate both visual and auditory cues to arrive at correct answers. These results are on OmniBench, which is a particularly challenging dataset where 54% of questions fall into the "hard" category (neither teacher correct). The model's reasoning traces demonstrate cross-modal integration rather than relying on single-modality shortcuts. Notably, the model successfully generalizes from its training domain (AVQA videos with 8 frames and audio) to a different evaluation domain (single static images paired with audio), suggesting that the reasoning capabilities learned through our teacher aggregation approach transfer effectively across different input formats and temporal structures.

5. Conclusion

We introduced AVRT, a novel framework that generates high-quality audio-visual reasoning data by distilling knowledge from specialized single-modality teachers, enabling effective supervised fine-tuning of student models with minimal reinforcement learning post-training. To this end, the pipeline uses two specialized reasoning teachers, one for audio and one for vision, to extract expert reasoning traces of each modality separately. The resulting traces are then merged and formatted by an LLM merger model into a single, multimodal reasoning trace. We then use those reasoning traces as cold-start in a two stage training pipeline. The resulting 3B parameter model achieves state-of-the-art performance: 56.3% on OmniBench, 49.2% on DailyOmni, and 57.3% on MMAR, establishing new benchmarks for efficient audio-visual understanding.

References

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

[2] 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. 2

[3] 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. *arXiv preprint arXiv:2503.23219*, 2025. 3

[4] 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. 1

[5] 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, 2025. 1

[6] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072, 2025. 1

[7] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18804–18814, 2025. 1

[8] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025. 1, 4, 6

[9] 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. 1

[10] 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. 1

[11] Yogesh Kulkarni and Pooyan Fazli. Avatar: Reinforcement learning to see, hear, and reason over video, 2025. 1, 2, 6

[12] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[13] 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, 2024. 2, 3, 5

[14] 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*, 2025. 1, 2, 6

[15] Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. Pave: Patching and adapting video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3306–3317, 2025. 2

[16] Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms, 2025. 6

[17] 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. *arXiv preprint arXiv:2505.13032*, 2025. 2, 5

[18] OpenAI. Hello gpt-4o, 2024. 1

[19] 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. 1, 4

[20] 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*, pages 7293–7301, 2025. 2

[21] 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, 2025. 4, 6

[22] 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*, 36:74952–74965, 2023. 3

[23] 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. 1

[24] Zikai Xie. Order matters in hallucination: Reasoning order as benchmark and reflexive prompting for large-language-models. *arXiv preprint arXiv:2408.05093*, 2024. 3

[25] 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. 1

[26] 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, 2025. 2, 3, 6

[27] 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*, 2025. 1, 2, 5, 6

[28] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491, 2022. 2, 3, 5

[29] 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. 1, 2

[30] 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. 6

[31] 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, 2025. 2, 6

[32] Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2025. 2, 5

AVRT: Audio-Visual Reasoning Transfer through Single-Modality Teachers

Supplementary Material

This supplementary material provides additional dataset statistics, analysis of design decisions in our data pipeline, experiments with masked modalities, results on scaling to 7B models, and the complete prompts for reproducibility.

5.1. Additional Dataset Statistics

Table 9. Additional statistics for the AVRT-20K dataset showing answer option distribution and question modality requirements.

Answer Option	Train / Val (%)
A	24.5 / 27.1
B	25.2 / 27.3
C	24.9 / 23.3
D	25.4 / 22.3

Question Relations	Train / Val (%)
Both (Audio + Visual)	99.0 / 98.9
Sound Only	0.7 / 0.7
Visual Only	0.3 / 0.3

The AVRT-20K dataset has balanced answer choices with each option (A-D) appearing roughly 25% of the time. Nearly all questions (99%) require both audio and visual information to answer correctly, with less than 1% being answerable using only one modality.

5.2. Data Generation and Filtering Strategy

Filtering Strategy: Both Teachers Must Be Correct. In this work, we adopt a filtering strategy that retains only samples where both single-modality teacher models can correctly answer. A potential concern with this filtering approach is that it might bias the dataset toward "easy" samples, removing challenging questions that could improve model robustness. However, our evaluation on OmniBench difficulty subsets (Table 8 in the main paper) directly addresses this concern. We categorize OmniBench questions based on teacher performance: easy (both teachers correct), medium (one teacher correct), and hard (neither teacher correct). Our AVRT model outperforms the Qwen2.5-Omni baseline across *all* difficulty levels, including a +6.5 point improvement on hard questions where neither teacher was correct during their evaluation. This demonstrates that training on high-quality filtered data does not create a harmful bias toward easy samples.

Role of Ground-Truth Answers in Training Data Generation. Another important design choice in AVRT is that the merger model receives the correct answer during training data generation, as shown in Figure 6. We emphasize that this ground-truth answer is provided *only* to the merger LLM during the data generation phase, and the student model *never* sees correct answers during either supervised fine-tuning or inference. The merger's role is to synthesize coherent reasoning traces that integrate audio and visual evidence, and the student model must learn to reproduce this reasoning process without access to labels.

5.3. Model's Behavior with Missing Modality

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/masked audio, and (3) only audio with blank/masked 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, and we analyze both quantitative performance and qualitative behavior of the model.

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 3 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 this behavior is encouraging, it could be sensitive to prompting strategies, and the model may occasionally reference missing modalities due to strong pre-training priors.

5.4. Experiments with Qwen2.5-Omni 7B

To demonstrate that the AVRT pipeline is not limited to 3B models, we apply the same methodology to the larger Qwen2.5-Omni 7B base model. Table 10 compares the performance of 3B and 7B models with and without AVRT training. The results show that AVRT 3B already outperforms the vanilla Qwen2.5-Omni 7B baseline on OmniBench (56.3% vs 50.7%), demonstrating the efficiency gains from audio-

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

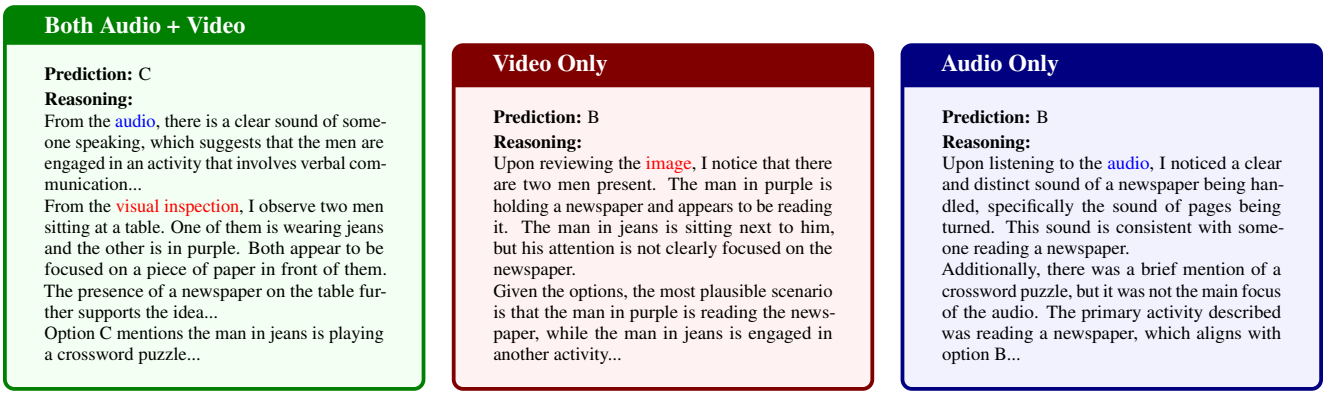

Figure 3. Example of model reasoning traces when different modalities are masked. With both modalities (left), the model integrates audio and visual evidence. With video only (center), reasoning focuses on visual observations without hallucinating audio content. With audio only (right), the model constrains itself to acoustic features without fabricating visual details. Color highlights indicate modality-specific reasoning: audio and visual.

visual reasoning trace aggregation. The AVRT methodology is model-agnostic and can be applied to any base audio-visual language model, with the gains from merging single-modality expertise expected to transfer to larger model sizes.

Table 10. Performance comparison across model sizes and training approaches. AVRT 3B outperforms the 7B baseline on OmniBench, demonstrating efficiency gains from reasoning trace aggregation.

Model	DailyOmni	OmniBench
Qwen2.5-Omni 3B	43.1	50.2
AVRT 3B (Ours)	**49.2**	**56.3**
Qwen2.5-Omni 7B	51.5	50.7
AVRT 7B (Ours)	**54.4**	**57.1**

5.5. 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).

5.5.1. Visual Teacher Prompt

We design the visual teacher prompt to elicit detailed temporal reasoning from video frames. As shown in Figure 4, 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 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.

5.5.2. 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 5, 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.

5.5.3. Merger Prompt

The merger model plays a critical role in combining the separate audio and visual reasoning traces into a unified multimodal analysis. Figure 6 shows the prompt used for Qwen2.5-14B-Instruct, which receives the question along with both the audio and visual analyses generated by the respective teacher models. Importantly, the prompt instructs the merger to integrate evidence from both modalities naturally, as if discovering the answer through its own reasoning process rather than acknowledging that it already has access to the correct answer. The merged reasoning is formatted within <think>tags, followed by the final answer in <answer>tags, creating training data that teaches the student model how to jointly reason about audio and visual inputs.

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 4. 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.

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 5. 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.

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}
Correct Answer: {correct_answer}
Audio Analysis: {audio_reasoning}
Visual Analysis: {vision_reasoning}
Instructions:
- Don't acknowledge that you already know the answer!
- Act as if you generated the reasoning and then you came across the right answer by yourself!
- Write plain English, but this time, 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 succint

Combined Analysis:

Figure 6. 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.