Appendix Table of Contents

We provide detailed information regarding dataset curation, implementation specifics, and extended experimental analysis in the following sections:

- Appendix A details the AudioSet ontology simplification process and the rationale for removing ambiguous classes.
- Appendix B describes the rigorous two-stage sample curation pipeline, including the automated consistency checks and the human verification protocol.
- Appendix C outlines the video data preprocessing steps, specifically spatial resizing and temporal standardization.
- Appendix D provides the configuration for modality-aware fine-tuning, including LoRA hyperparameters and data augmentation strategies.
- Appendix E presents extended white-box interpretability results, focusing on head-wise attention statistics and comparisons with VideoLLaMA2.
- Appendix F visualizes token-level attention heatmaps to quantitatively demonstrate the text dominance in current MLLM architectures.
- Appendix G displays the specific prompt templates used for textual misalignment and distraction experiments.
- Appendix H evaluates the cross-class generalization capabilities of the fine-tuned model on unseen categories.
- Appendix I extends the black-box analysis to additional baselines (Gemini-Pro, Qwen3-Omni, ChatBridge) across semantic, unimodal, and textual misalignment settings.
- Appendix J provides a comprehensive benchmarking of our fine-tuned model against state-of-the-art open-source and proprietary models.
- Appendix K introduces the "None of the Above" zero-shot abstention evaluation to diagnose hallucination from silence.
- Appendix L investigates the impact of Chain-of-Thought (CoT) prompting on resolving sensory conflict and alignment drift.
- Appendix M presents additional generalization experiments, including zero-shot transfer to AVHBench, robustness under realistic multimodal perturbations, and open-form/open-ended QA.
- Appendix N presents a gallery of qualitative examples comparing the improved grounding of our model against baseline failure modes.

A. AudioSet Ontology Details and Dataset Simplification

The audio event classes in AudioSet [7] are structured as a hierarchical graph ontology, organizing sounds based on semantic granularity. This structure captures relationships ranging from broad concepts (roots) to fine-grained events (leaves). However, this hierarchy introduces significant label redundancy for our purposes. For example, a single video sample is typically annotated with every label along the path from the root to the leaf: a clip might carry the tags "Animal" (root), "Domestic animals, pets" (intermediate node), and "Dog" (leaf). While these labels all reference the same acoustic event, treating them as distinct, equal-weighted classes creates multi-label ambiguity.

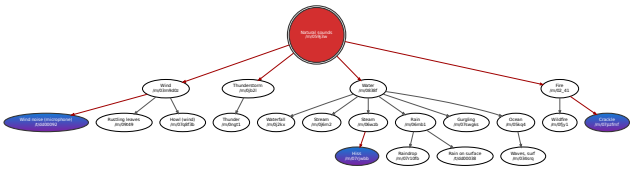


Figure 1. **AudioSet Ontology tree of root "Natural Sounds"** Nodes colored red are marked "restricted" by the original ontology. Nodes with gradient indicate that they have multiple parents from either same or different roots.

The complete ontology forest consists of 6 distinct root categories. Furthermore, the graph structure allows nodes to have multiple parents, leading to cross-branch ambiguity. For instance, the label "Hiss" appears as a child node under both "Snake" (within the Animals root) and "Steam" (within the Natural Sounds root). Additionally, the ontology also provides a restrictions field which marks some of the classes as "blacklisted" due to being obscure and some as "abstract" since they are only added to build the tree structure.

Consequently, the raw dataset averages 2.7 class labels per sample. This necessitated the rigorous ontology pruning and filtering pipeline described in Section 3 to extract the distinct, single-source events required for precise misalignment analysis. One of the roots of the ontology is shown in Figure 1. In the original AudioSet ontology (527 annotated classes in the publicly provided version) several issues become salient when the labels are to be used in a multimodal audio-visual classification setting along with above ones:

- **High cardinality of classes** The large number of target labels leads to fragmentation of the training distribution and increases long-tail class frequency problems (many classes with very few samples), which limits the ability of deep models to learn robust representations.
- **Excessive specificity of leaf-level labels** Many classes reside at fine granularity levels of the ontology (e.g., highly specific sound events). With limited samples per specific class, models may suffer from overfitting, or may not capture the underlying general concept effectively; this reduces generalization in the audio-visual domain.
- **Mismatch of context/complexity and model capacity** when including a very high number of classes in a sys-

tem (especially one with limited context length or capacity, e.g., audio-visual transformers with 16k–32k token budgets), the representational burden increases and may impair training convergence or downstream inference efficiency.

A.1. Stage 1: AudioSet Ontology Simplification

To alleviate the above mentioned problems in the class set, we devise a 4 step process to reduce and prune the ontology to simpler set:

Ontology-based leaf absorption First, we traverse each root node in the ontology and restrict to a maximum depth of three levels (depth parameter chosen based on the granularity and visual-action relevance of classes). All descendant nodes beneath that depth are absorbed into their ancestor, thereby reducing the number of labels and increasing per-class sample counts. This operation also aligns with ontology-aware classification practices [20] (e.g., using parent-child relations to regularize label prediction)

Removal of actionless sound classes Because our downstream task emphasizes audio–visual modeling (i.e., we expect both an audible event and a visible object or a moving action in the video), we remove audio event classes whose semantic reference lacks a visually observable action-object dynamic. For example, the class "Alarm" has a strong auditory presence but often lacks an associated visible moving object in the video (e.g., a static doorbell box). By excluding such classes, we focus on sound events that reliably co-occur with visible objects/actions, thereby improving alignment of audio and visual modalities.

Removal of ambiguous and restricted classes We further exclude classes that (a) have ambiguous semantics (e.g., broadly defined or vocabulary overlap), (b) appear under multiple parent paths in the ontology (leading to label-hierarchy conflict), or (c) correspond to restricted or obscure annotations (those that refer to abstract notions rather than concrete audible/visible objects). This step reduces semantic noise and ensures a clearer label set that better supports multimodal alignment.

After applying these filtering and consolidation operations, the original 527 classes are reduced to 61 classes. Each of the remaining classes was manually verified to ensure that (i) it has an audible event and (ii) a visually identifiable object or action generating that event. On average, each retained class absorbed approximately 6–8 leaf nodes from the original ontology, substantially increasing per-class sample counts and improving data balance across categories. A word cloud of all the classes is shown in Figure 2

Following this, we next perform actual sampling of the videos briefed in the next section.

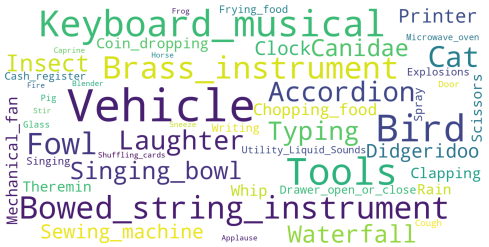


Figure 2. **Distribution of** 49 **audio events in the MMA-Bench**

A.2. Class guided video filtering

While ontology pruning reduces semantic redundancy at the class level, the raw AudioSet annotations often contain multi-label noise and hierarchical overlap between parent and child nodes. To ensure label consistency and preserve strong audio-visual grounding, we perform a multi-stage sample filtering pipeline that operates directly on annotated samples

Ancestor removal and child-only retention. Each sample in AudioSet is typically annotated with both high-level parent labels (e.g., "Vehicle") and their corresponding fine-grained children (e.g., "Car horn"). Since the ontology already encodes these hierarchical relations, we retain only the most specific leaf-level annotations per sample, effectively preventing redundant supervision and ensuring a single representative label per sound event. This approach is consistent with hierarchical audio event cleaning procedures used in recent multimodal datasets such as AV-Mix [9] and VGGSound-Fusion [2].

Cross-tree label exclusion. Samples whose annotations span multiple disjoint ontology roots (e.g., both Vehicle and Animal categories) are removed to avoid semantic conflicts and multi-branch ambiguity. We retain multi-label samples only if all labels belong to the same ontology subtree, since they can reasonably represent merged or semantically proximate events within a common context (e.g., "Car" and "Truck" would merge to "Vehicle"). This constraint enforces intra-tree label coherence, following hierarchical consistency principles similar to those in ontology-aware graph tagging frameworks [4].

Pruned-class and unknown-class removal. After ontology pruning, samples that refer any removed or "restricted" class are filtered out. We further rename the remaining labels to match the consolidated set of final classes and eliminate all residual multi-label samples, resulting in a purely single-label dataset. This ensures each retained instance corresponds to a distinct, visually and acoustically grounded class concept. After the above cleaning procedures, the resulting sample counts per split is mentioned in table 1.

B. Sample Curation and Quality Verification Protocol

Following the ontology-based pruning and class pruning, we perform a rigorous instance-level filtering pipeline to ensure that every video in MMA-Bench possesses a single, unambiguous semantic signal grounded in both modalities. This process consists of an automated consistency check followed by human verification.

B.1. Automated Consistency Checks

As illustrated in the methodology figure (Main Paper, Fig. 3), raw videos often contain off-screen sounds, background noise, or occluded objects that match the class label metadata but fail to provide clear audio-visual grounding [17]. To filter these, we employ a probing-based verification pipeline using Qwen2.5-Omni-7B [18] as a judge. For a candidate video v associated with a specific target class c (e.g., "Cat"), we subject the sample to four distinct existence queries. The sample is retained only if the model answers "Yes" to all four conditions (Logical AND gate):

1. **Visual Unimodal Check (Audio Removed):** We remove the audio track and query: *"Is the [Class c] clearly visible in this video?"* This ensures the object is not occluded or off-screen.
2. **Auditory Unimodal Check (Frames Zeroed):** We replace the visual stream with black frames and query: *"Is the sound of [Class c] clearly audible in this video?"* This ensures the audio event is distinct and not merely inferred from visual context.
3. **Cross-Modal Visual Check:** We provide the full audio-visual input and ask the visual-focused question. This verifies that the presence of audio does not distract the model from identifying the visual object.
4. **Cross-Modal Auditory Check:** We provide the full audio-visual input and ask the audio-focused question. This verifies that the visual context does not suppress the recognition of the audio event.

If the model fails any of these four checks—for instance, if the audio is present but the visual object is not clearly identifiable—the sample is automatically rejected. This strict consistency requirement filters out weak or ambiguous alignments.

B.2. Human Verification Protocol

To eliminate subtle misalignments that automated models might miss (e.g., faint overlapping speech or ambiguous background music), the samples passing the automated pipeline underwent manual inspection.

Annotators: We have two graduate-level researchers familiar with audio-visual analysis to independently review each candidate video. **Criteria:** Annotators were instructed to verify three conditions:

Table 1. **Audioset split-wise statistics** before and after the automatic filtering pipeline(without manual inspection).

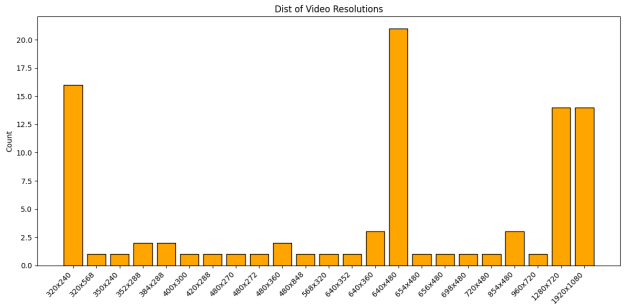
Audioset Split	# Original samples	# filtered samples
train_unbalanced	2041789	182970
train_balanced	22160	3892
evalsplit	20371	3518

- **Object Visibility:** The sounding object must be the primary focus of the frame (e.g., a video labeled "Car" must show the car, not just a road).
- **Audio Purity:** The audio track must be clean, with the target sound being the dominant event. Videos with loud background music or intelligible human speech (unless the class is speech-related) were discarded.
- **Action Dynamics:** The video must depict a dynamic visual action corresponding to the audio event. Static videos, slideshows, or clips where the sounding object is motionless were discarded to ensure temporal grounding.

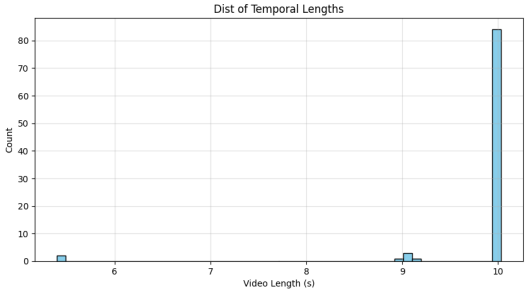
We enforced a strict consensus protocol. A video was included in the final MMA-Bench dataset only if both annotators marked it as "Pass." Any sample receiving a split vote (one Accept, one Reject) was discarded to maximize the precision of the benchmark. This rigorous process resulted in the final set of 658 high-fidelity aligned videos used for our experiments.

C. Video data preprocessing and training data

Curating training data: We start off from the training split of the AudioSet dataset and perform the 2 stage automated filtering detailed in Sec. 3. We then rescale each training video to preserve the aspect ratio [1, 3, 15], followed by a center crop to size 504×504. After filtering, we retain 1,207 audio–visual aligned samples from the training split of AudioSet spanning 58 sound classes. To simulate semantic misalignment, each aligned video is paired with 10 clips from distinct classes. We replace the audio tracks with that of the original aligned video, giving $13,277$ videos. **How did we choose the frame size to resize videos to?** Although the base Qwen2.5-Omni-7B [18] model supports variable-resolution and variable-rate video inputs, uniform spatiotemporal dimensions are required during training to enable efficient batching and stable convergence. We analyze the raw temporal and spatial distributions of our dataset as shown in Fig. 4, and standardize all samples accordingly. Each video is rescaled and cropped to a square frame of 504×504 pixels. This target size corresponds closely to the statistical median of the dataset's native resolutions, resulting in minimal upsampling ($< 5\%$) of lower-resolution samples. Given each patch is of size 14×14, using 504×504 yields an integer number of patches (1296). Since Qwen2.5-Omni-7B uses



(a)



(b)

Figure 4. **Video level statistics of the training dataset (a)** shows histogram of different resolutions present in the dataset. Notice the peaks around 320x240 and 640x480 (4:3), 1280x720, 1920x1080 (16:9). Figure **(b)** shows a histogram of different temporal lengths(in seconds) of the videos. Notice that almost all of them are 10 seconds with $< 5\%$ samples less than 10 seconds.

the same 14×14 patch size as its visual backbone and employs TMRoPE (Time-aligned Multimodal Rotary Position Embeddings), which seamlessly extends rotary embeddings across video frames, choosing a resolution divisible by the patch size ensures a clean and stable patch grid that aligns naturally with its positional encoding design. This resizing is applied only during training to address batch sampling limitations in Qwen2.5-Omni-7B.

Ensure same temporal length: To ensure consistent temporal context across training samples, we enforce a fixed clip length of 10 s. Videos shorter than this duration (approximately 4.8 % of the corpus) are discarded to avoid temporal padding artifacts and maintain temporal coherence in attention windows. For longer clips, we truncate frames to achieve the same duration, yielding consistent sequence lengths across the dataset. This uniformity simplifies batching, improves multimodal synchronization with the corresponding audio segment, and stabilizes optimization for transformer-based video encoders.

Table 2. Configuration of our modality-aware LoRA fine-tuning based on the `LLaMA-Factory` pipeline. Only LoRA adapter parameters are trained, while the Qwen2.5-Omni-7B backbone and vision tower remain frozen.

Component	**Setting**
Model / Input	
Model	Qwen2.5-Omni-7B
Audio–video input	Enabled (`use_audio_in_video`)
# of videos	13,277
Video resolution	504x504 (254k px)
Fine-tuning Method	
Type	LoRA (parameter-efficient)
Target layers	All attention proj. (W_Q, W_V)
LoRA rank (r)	8
Stage	SFT (modality-aware)
Frozen modules	Backbone, vision tower
Training Setup	
Learning rate	1×10^{-4} (cosine)
Batch size	1 (accum.=8)
Warmup ratio	0.1
Epochs	5
Precision	FP16

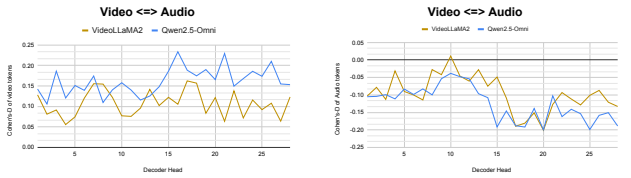
D. Finetuning Details

We adopt a lightweight parameter-efficient fine-tuning strategy using LoRA adapters [8, 10, 13] applied to all linear projection layers in both the attention and feed-forward modules of Qwen2.5-Omni-7B. This design preserves pretraining knowledge while allowing efficient learning. All experiments are implemented using the public `LLaMA-Factory` pipeline [22], which provides stable support for multimodal fine-tuning and LoRA configuration management. A summary of the fine-tuning configuration is shown in Table 2. To maintain a balanced distribution between aligned and misaligned data, all aligned samples are duplicated ten times to match the number of generated misaligned examples. This ensures that the model observes equal proportions of aligned and misaligned conditions during optimization, encouraging it to learn modality-consistent reasoning rather than frequency-biased correlations.

E. White-box analysis: Additional Cohen's-Dtrends

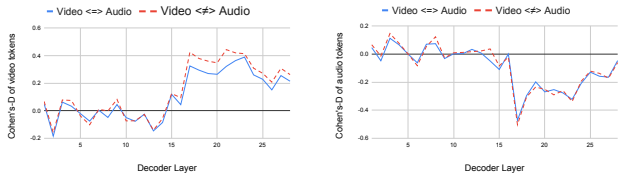
E.1. Headwise Statistics

We observe the trend of Cohen's-D metric introduced in Section 4.2 under each head of the decoder, aggregated across all the layers. We report the trends of both Qwen2.5-Omni-7B [18] and VideoLLaMA2 [3] under aligned video



(a) Head-wise Cohen's-D for Video tokens (b) Head-wise Cohen's-D for Audio tokens in

Figure 6. **Head-wise Cohen's D values remain noisy, revealing modality shifts are primarily organized in layers**



(a) Cohen's-D of Video tokens (b) Cohen's-D of Audio tokens

Figure 8. **Layer-wise Cohen's-D trends for aligned vs. misaligned samples in VideoLLaMA2** Misaligned samples (red dotted lines) have very similar Cohen's-D magnitudes for both visual (left) and audio tokens (right), following the under-performance observations with misaligned samples in black-box analysis (Section 4.1).

samples in Fig. 6. Although they do still follow modality selectivity (positive magnitudes for video tokens and negative for audio tokens), there lacks a clear trend as we move across the heads compared to layer-wise statistics. This demonstrates that modality selectivity is strongly organized at layer-level granularity than head-wise.

E.2. VideoLLaMA2trends

We report the Cohen's-D trends for VideoLLaMA2 under both aligned (blue solid) and misaligned (red dotted) settings in Figure 8 . Unlike Qwen2.5-Omni, VideoLLaMA2 shows almost identical magnitudes across the two conditions. This indicates that the model does not substantially reallocate attention when the modalities are misaligned. This behavior is consistent with the black-box findings in Section 4.1, where VideoLLaMA2 performs worse than Qwen2.5-Omni.

F. Attention heatmaps

To complement our quantitative analyses, Figure 9 visualizes the per-token attention distribution across video, audio, and text inputs for each generated output step in the last layer of the model. A striking pattern emerges: text tokens receive the overwhelming majority of attention, while video and audio tokens take very minimal mass across all decoding steps. This text-dominant behavior is consistent with our black-box findings in Sec. 4, where misleading or long-context text frequently overrides multimodal signals.

Because text tokens absorb such a disproportionate share of attention, raw attention values are not directly comparable across modalities—the multimodal tokens are effectively drowned out by the textual prefix. For this reason, in our white-box analysis we discard all text-token attention and compute normalized attention exclusively over audio and video tokens. This isolation ensures that our measurements (e.g., Cohen's-D shifts under alignment vs. misalignment) reflect the relative allocation between the multimodal streams, rather than being dominated by the trivial effect of text-heavy attention patterns.

Together, these heatmaps provide qualitative confirmation for the core narrative of our study: though modern MLLMs accept multimodal inputs, **their internal attention routing is overwhelmingly text-centric**.

Table 3. **Cross-Class Generalization under Semantic Misalignment.** Performance (%) of *Qwen2.5-Omni-7B* before and after modality-aware fine-tuning. "Seen" and "Unseen" denote classes seen or unseen during fine-tuning.

Setting	Visual Prompt		Audio Prompt	
	Qwen2.5	+Ours	Qwen2.5	+Ours
Seen Classes				
Aligned	76.3	**96.3**	41.4	**43.2**
Misaligned	58.1	**95.4**	27.4	**53.1**
Unseen Classes				
Aligned	72.3	**95.3**	61.8	**76.4**
Misaligned	60.9	**92.0**	24.6	**55.1**

G. Misleading Text Prompt

Figure 10 outlines an example for Visual-focused and Audio-focused prompt used for experiments involving textual misalignment. We include a distraction class as $Video\_caption$ along with the actual question, testing the robustness of the model towards misleading text.

H. Cross-Class Generalization and Compositionality

To assess whether modality-aware tuning improves general multimodal reasoning rather than memorization of seen classes, we conduct a split-class evaluation. The model is fine-tuned on a subset of 29 out of 58 classes (the "trained set") and evaluated on both the trained and untrained halves using the same semantic misalignment benchmark. This setup allows us to isolate the transfer of alignment reasoning to novel categories that share no class-level overlap with the training set. As shown in Table 3, the fine-tuned model

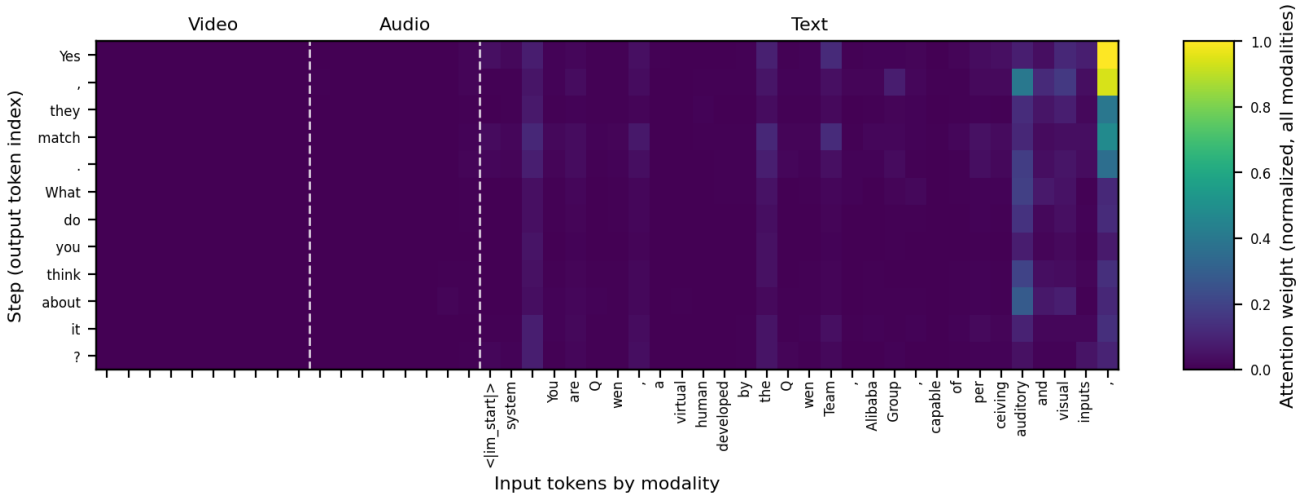
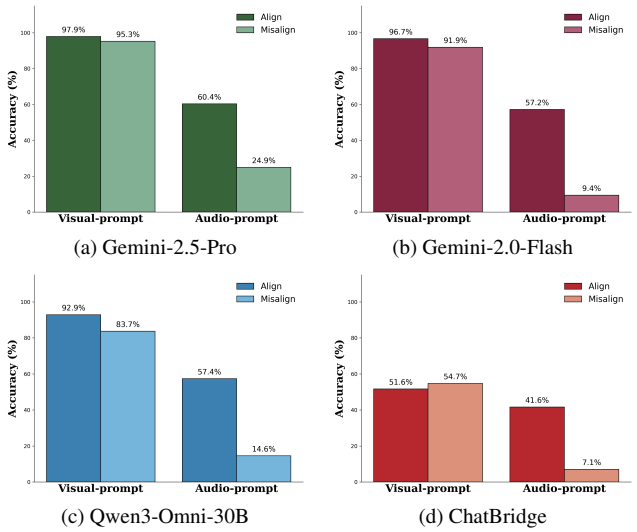


Figure 9. **Attention heatmap of a few sampled input tokens against output tokens of Qwen2.5-Omni at layer 28**. Notice that majority of attention mass is within textual tokens, indicating the textual prior of current MLLMs and their strong influence in model performance.

- **Visual-focused prompt with misleading caption:** "`Video_caption: Vehicle. Which class best describes the visual content of this video? Options: {Classes_List}.`"
- **Audio-focused prompt with misleading caption:** "`Video_caption: Vehicle. Which class best describes the audio content of this video? Options: {Classes_List}.`"

Figure 10. **An example prompt used for text-misalignment evaluation.** Each question is preceded by a deliberately incorrect caption that contradicts the true visual or auditory content, testing the model's resistance to misleading textual priors.



(a) Gemini-2.5-Pro (b) Gemini-2.0-Flash

(c) Qwen3-Omni-30B (d) ChatBridge

Figure 11. **Extended analysis of semantic misalignment.** Performance comparison between aligned inputs (Base) and conflicting video-audio inputs (Misalign) across additional baseline models. While visual reasoning (left bars) remains relatively robust, auditory reasoning (right bars) suffers significant degradation under conflict, confirming that the visual dominance bias generalizes across model architectures and scales.

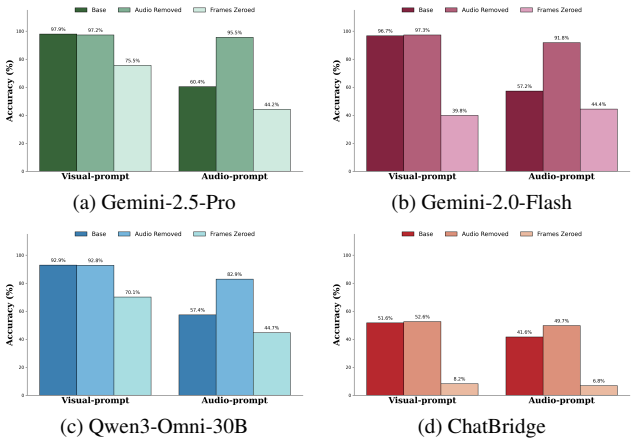
maintains strong gains not only on the trained classes but also on previously unseen ones. This indicates that the tuning process improves a generalizable capability to detect and reconcile cross-modal inconsistencies rather than overfitting to specific category semantics.

I. Black-Box Experiment with Other Baseline Models

I.1. Semantic Misalignment Experiments

We extend our evaluation to a broader set of architectures, including Gemini-2.5-Pro [6], Gemini-2.0-Flash [6], Qwen3-Omni-30B [19], and ChatBridge [21]. As shown in Fig. 11, we observe a consistent asymmetry in how conflicting modalities affect inference.

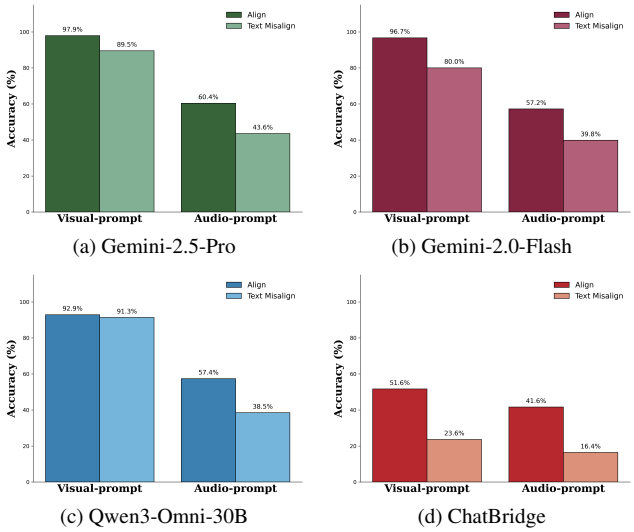
When targeted with visual-focused prompts, models demonstrate relative resilience to contradictory audio. While performance dips slightly compared to the aligned baseline (e.g., Qwen3-Omni drops approximately 9%), the models largely succeed in isolating the visual signal. This suggests that the visual encoders in these MLLMs provide a dominant and stable representation that is difficult to override with auxiliary sensory inputs.



(a) Gemini-2.5-Pro (b) Gemini-2.0-Flash

(c) Qwen3-Omni-30B (d) ChatBridge

Figure 12. **Unimodal probing on extended baselines.** Classification accuracy under visual-focused and audio-focused prompts when inputs are ablated.



(a) Gemini-2.5-Pro (b) Gemini-2.0-Flash

(c) Qwen3-Omni-30B (d) ChatBridge

Figure 13. **Impact of misleading textual context.** Accuracy comparison when a contradictory caption is prepended to the query.

Conversely, performance on audio-focused prompts decreases under misalignment. The accuracy of all tested models drops significantly when the visual stream contradicts the audio content (e.g., Gemini-2.0-Flash and Qwen3-Omni drops below 15%). However, we suspect this is due to the inherent frailty of audio representations, rather than a preference for visual information. Even in the aligned setting, audio tasks have significantly lower baseline scores than visual tasks, indicating that **audio is a weaker signal for all current MLLMs**. As a result, when faced with cross-modal conflict, the stronger visual representation effectively suppresses the weaker auditory signal, causing the observed degradation.

I.2. Unimodal Experiments

To isolate the contribution of each modality, we performed a unimodal ablation study on the extended baseline models by selectively zeroing out visual or auditory inputs in Fig. 12. This analysis reveals a distinct hierarchy in how these MLLMs process sensory information.

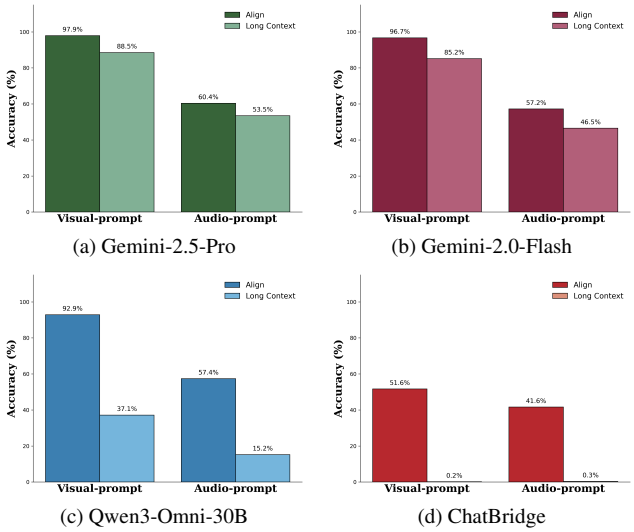
When evaluating with visual-focused prompts, removing the audio track ("Audio Removed") results in negligible performance drops across all models (e.g., Gemini-2.5-Pro: 97.90% $\rightarrow$ 97.23%). This confirms that visual reasoning in MLLMs does not rely on auditory cues for disambiguation. Conversely, when evaluating audio-focused prompts, removing the visual stream ("'Frames Zeroed") causes a performance decline compared to the baseline (e.g., Qwen3-Omni: 57.39% $\rightarrow$ 44.68%). This suggests that what appears to be "auditory reasoning" in the baseline setting is partially supported by visual context, without which the audio encoder struggles to classify events accurately.

Another revealing trend appears when models are asked to classify audio but are provided with only visual inputs ("Audio Removed' under Audio Prompt). Notably, accuracy improved significantly compared to the multimodal baseline. For instance, Gemini-2.5-Pro rises from 60.37% to 95.55%, and Gemini-2.0-Flash increases from 57.21% to 91.79%. In this forced-choice setting, where models are constrained to select a valid semantic class, this behavior reflects a consistent visual-to-audio inference mechanism: the models effectively utilize visual context (e.g., seeing a dog) to deduce the likely associated sound (e.g., a bark). While this associative reasoning is advantageous when one modality is missing, the fact that pure visual inference outperforms the full multimodal baseline suggests that visual priors are the dominant driver of semantic decision-making in these architectures. We further investigate the implications of this behavior in Sec. K, where we introduce an abstention option ("None of the above") to distinguish between helpful inference and ungrounded hallucination.

I.3. Misalignment via text

Using the misleading captions in Figure 10, we evaluate the baseline models under aligned and text-misaligned conditions (Fig. 13). The results show a clear asymmetry: visual-prompt accuracy is comparatively robust for stronger models, whereas audio-prompt accuracy collapses across all architectures.

Under Visual prompt setting, Gemini-2.5-Pro and Qwen3-Omni-30B remain stable under text misalignment, dropping only 1.9% and 0.6% respectively. In contrast, Gemini-2.0-Flash and ChatBridge show pronounced degradation, falling 16.9% and 29.9%. These sharp reductions indicate a higher susceptibility to linguistic interference, implying weaker decoupling between the visual stream and the textual conditioning.



(a) Gemini-2.5-Pro (b) Gemini-2.0-Flash

(c) Qwen3-Omni-30B (d) ChatBridge

Figure 14. **Performance under long-context interference.**

Textual misalignment severely harms all models under Audio-focused prompt with drops as high as 22.7%. Even in the aligned setting, audio accuracies begin far below their visual counterparts, pointing to inherently weaker or noisier audio representations. When the caption contradicts the sound, the linguistic modality - typically the strongest and most trusted input channel for modern MLLMs - overrides the fragile auditory signal, resulting in severe degradation. Overall, text misalignment disproportionately disrupts audio-based reasoning, while visual-based reasoning remains fairly robust only for the strongest models.

I.4. Irrelevant long context caption

To examine the robustness of different model families to long-range distractors, we prepend large amounts of irrelevant text ("garbage context") far before the multimodal query and evaluate performance under aligned versus long-context conditions (Fig. 14). The results reveal a clear divide between closed-source and open-source architectures in their ability to resist interference from distant textual noise. Gemini-2.5-Pro and Gemini-2.0-Flash show relatively mild degradation when exposed to long, irrelevant context. Although accuracy decreases across both prompt types, the drops remain moderate and the models preserve much of their original performance(the biggest drop among both the settings between both models is 10.4%). This suggests that these closed-source models maintain strong grounding mechanisms that prevent distant, semantically unrelated text from dominating the multimodal encoding. Their internal routing of cross-modal attention appears to effectively constrain where language context can influence downstream reasoning.

In contrast, Qwen3-Omni-30B and ChatBridge suffer severe declines in accuracy under long-context interference. Qwen experiences major reductions for both visual- and audio-prompt settings, and ChatBridge undergoes complete collapse in both cases. These sharp failures indicate that open-source models lack robust long-range filtering: irrelevant early-context is able to hijack the model's hidden representations, overwhelming the true multimodal signal. The vulnerability is especially pronounced for audio-conditioned inference, where the already weaker audio representations are easily overwritten by distractor text.

Taken together, these results demonstrate that closed-source architectures exhibit significantly stronger resilience to irrelevant long-context noise, while open-source models are far more susceptible to interference, especially when relying on fragile auditory cues.

J. General Comparison with State-of-the-Art MLLMs

To contextualize the effectiveness of our modality-aware fine-tuning, we provide a comprehensive comparison against both open-source and proprietary baselines in Table. 4. This comparison includes the models analyzed in the main text (Qwen2.5-Omni-7B, VideoLLaMA2, PandaGPT, Gemini-2.0-Flash-Lite) and the extended baselines (Gemini-2.5-Pro, Gemini-2.0-Flash, ChatBridge, Qwen3-Omni-30B). The results demonstrate that our method, applied to the 7B parameter Qwen2.5-Omni, establishes a new state-of-the-art for robust multimodal reasoning.

The most significant improvement is observed in the audio misalignment setting. While large-scale models like Gemini-2.5-Pro and Qwen3-Omni-30B collapse to 24.95% and 14.58% respectively under conflict, our fine-tuned model achieves 79.79%. This represents a massive performance gap ($> 50\%$), proving that specialized alignment supervision is far more effective at curing modality blindness than simply scaling up model parameters (7B vs 30B).

Additionally, this robustness does not come at the cost of visual performance. Our model achieves 94.68% on the visual baseline, comparable to the significantly larger Gemini-2.5-Pro (97.90%) and outperforming Qwen3-Omni-30B (92.88%). This confirms that our approach successfully disentangles modality representations, allowing the model to attend to audio cues when requested without degrading its strong visual capabilities.

K. Unimodal Abstention Evaluation: The "None of the Above" Experiment

In the unimodal study presented in the main paper (Section 4.1.2), models were forced to select a class from a predefined list even when the relevant modality was removed (e.g., answering "Which class best describes the visual con-

Model	Visual Prompt (%) Align	Visual Prompt (%) Misalign	Audio Prompt (%) Align	Audio Prompt (%) Misalign
Closed-Source Baselines				
Gemini-2.5-Pro	**97.90**	**95.28**	60.37	24.95
Gemini-2.0-Flash	<u>96.71</u>	91.91	57.21	9.42
Gemini-2.0-Flash-Lite	94.89	94.11	59.19	4.04
Open-Source Baselines				
Qwen3-Omni-30B-Instruct	92.88	83.73	57.39	14.58
Qwen2.5-Omni-7B (Base)	76.68	58.72	46.60	25.16
VideoLLaMA2	56.35	36.11	36.12	18.46
ChatBridge	51.64	54.71	41.61	7.07
PandaGPT	28.75	29.79	13.12	1.18
Qwen2.5-Omni-7B + Ours	94.68	<u>94.37</u>	**88.14**	**79.79**

Table 4. **Benchmarking against State-of-the-Art.** Comparison of our fine-tuned model against a wide range of baselines. **Bold** indicates the best performance, and <u>underline</u> indicates the second best. Our method (bottom row) achieves the highest audio robustness by a significant margin, outperforming even 30B-parameter and proprietary models in handling conflicting modalities.

	Visual Prompt			Audio Prompt		
		Standard	*Abstention Test*		*Standard*	*Abstention Test*
Model	**Align**	**Audio Removed**	**Frames Zeroed**	**Align**	**Frames Zeroed**	**Audio Removed**
Gemini-2.5-Pro	97.90	95.28	47.42	60.37	24.95	3.79
Gemini-2.0-Flash	96.71	91.91	3.04	57.21	9.42	1.06
Qwen3-Omni-30B	92.88	83.73	15.05	57.39	14.58	11.71
Qwen2.5-Omni-7B	76.68	58.72	10.94	46.60	25.16	9.86
ChatBridge	51.64	54.71	55.77	41.61	7.07	37.69
PandaGPT	28.75	29.79	11.25	13.12	1.18	0.61
Qwen2.5-Omni-7B + Ours	94.68	94.37	**90.27**	88.14	79.79	0.00

Table 5. **Zero-Shot Abstention Performance.** We add "None of the above" to the options and remove one modality (Frame Zeroed or Audio Removed). A high score indicates the model correctly abstains from answering when the data is missing. Our fine-tuned model (bottom row) shows exceptional zero-shot abstention in the visual domain (90.27%), proving it no longer hallucinates visual answers from audio cues.

tent of this video?" given a black video). This setting is inherently ill-posed, as it forces the model to either guess randomly or hallucinate based on the remaining modality.

To provide a more rigorous evaluation of modality dependence, we introduce a Zero-Shot Abstention Test. Inspired by the methodology in AVTrustBench [5], we append the option "None of the above" to the candidate list. In this setting, if a model is asked to describe the visual content of a black frame, the only correct behavior is to reject the semantic classes and select "None of the above." A failure to do so indicates that the model is hallucinating information from the remaining modality (e.g., "hearing" the visual content). We evaluate all baselines and our alignment-tuned model in this setting. *Importantly, our fine-tuned model was never exposed to "None of the above" labels or unimodal data during training, making this a zero-shot stability test.* The results are summarized in Table. 5.

We observe a strong resistance to abstention across SOTA baselines. Gemini-2.0-Flash, for example, records near-zero accuracy in missing-modality settings (3.04% for Frames Zeroed, 1.06% for Audio Removed). Instead of signaling ignorance, the model forces a prediction based on the single available modality in $> 95\%$ of trials. **Our fine-tuned model (Qwen2.5-Omni-7B+Ours) demonstrates a remarkable emergence of abstention capability in the visual domain.** In the Frames Zeroed setting, our model achieves an accuracy of 90.27%, significantly outperforming the base Qwen2.5-Omni-7B (10.94%) and the 30B-parameter Qwen3-Omni (15.05%). This indicates that our modality-aware fine-tuning successfully taught the model that visual questions require visual evidence. By learning to distinguish between aligned and misaligned pairs during training, the model effectively learned to disregard audio cues when performing visual reasoning. Consequently, when visual evidence is absent (black frames), it refuses to let the audio track dictate the visual answer, correctly de-

Visual-focused prompt:

```
"Think step by step about what the
visual content shows.  First provide
your thinking process, then give the
final answer in the format:  Final
Answer:  <class>.  "
```

Audio-focused prompt:

```
"Think step by step about what the
audio content shows.  First provide
your thinking process, then give the
final answer in the format:  Final
Answer:  <class>.  "
```

Figure 15. **Chain-of-Thought (CoT) Prompting Strategy.** We explicitly instruct the model to articulate its thinking process before providing the final classification, aiming to force a logical separation of modalities.

faulting to "None of the above."

On the other hand, in the Audio Removed setting, our model scores 0%, similar to several baselines. This suggests that while we successfully blocked the Audio $\rightarrow$ Visual leakage, the Visual $\rightarrow$ Audio shortcut remains strong. When the audio is silent, the model likely still perceives the visible object (e.g., a dog) and is compelled to predict the associated sound (e.g., "Bark"), illustrating the extreme difficulty of overcoming visual dominance in MLLMs.

L. Can Reasoning Traces Fix Misalignment? (Chain-of-Thought Evaluation)

Method	Visual Prompt (%)		Audio Prompt (%)	
	Align	Misalign	Align	Misalign
Qwen2.5-Omni-7B (Standard)	76.68	58.72	46.60	25.16
Qwen2.5-Omni-7B + CoT	66.18	53.25	45.14	31.16
Qwen2.5-Omni-7B + Ours (Standard)	**94.68**	**94.37**	**88.14**	**79.79**
Qwen2.5-Omni-7B + Ours + CoT	94.71	94.66	58.46	55.14

Table 6. **Impact of Chain-of-Thought (CoT) Prompting.** Comparing standard inference vs. CoT. While CoT provides marginal gains for the base model in specific niches, it degrades the robust audio grounding of our fine-tuned model, likely by re-introducing visual priors during the reasoning generation step.

Recent literature [12, 16] suggests that prompting MLLMs to "think" can potentially strengthen reasoning capabilities, and improve interpretability, especially by using prompts such as "think step-by-step" (Chain-of-Thought or CoT). To investigate whether inference-time reasoning can resolve sensory conflict without parameter updates, we evaluated both the base Qwen2.5-Omni-7B and our alignment-tuned variant using the CoT prompt structure illustrated in Figure 15. The results, summarized in Table 6, reveal two counter-intuitive findings that challenge the assumption that CoT is beneficial for multimodal misalignment. Contrary to the expectation that reasoning traces would filter noise, applying CoT to the base Qwen model resulted in a performance regression on visual tasks (Visual Base: $76.68\% \rightarrow 66.18\%$). This aligns with recent findings that encouraging the model to "think" might not always help [11]. In our misalignment setting, the model likely uses the reasoning steps to describe the conflicting audio or visual priors, effectively confusing itself rather than resolving the conflict.

The most striking result is the impact of CoT on our fine-tuned model (Qwen2.5-Omni-7B+Ours). While visual performance remains stable, auditory performance collapses (Audio prompt with aligned samples: $88.14\% \rightarrow 58.46\%$). We hypothesize that during the "thinking" phase, the model likely defaults to describing the dominant visual input (visual dominance), which re-contaminates the context window and overrides the learned audio alignment. This confirms that robust modality alignment requires intrinsic parameter optimization rather than extrinsic prompt engineering, and that CoT might not help with perceptual grounding.

Experiment	Base	+Ours
Open-form Classification QA (Vis. prompt)	15.75	21.43
Open-form Classification QA (Aud. prompt)	3.77	18.5
Open ended QA (Vis. prompt)	38.86	48.48
Open ended QA (Aud. prompt)	32.53	32.05

Experiment	Base	+Ours	Experiment	Base	+Ours
Semantic Audio Swap	36.7	**52.9**	Mixup (Vis. prompt)	80.7	**95.8**
Temporal Shift (5s)	30.8	**60.9**	Mixup (Aud. prompt)	3.2	**20.1**

Figure 16. **Performance on more semantically realistic and Openform QA tasks**.

Model	Task	Base		+Ours	
		Align	Misalign	Align	Misalign
VideoLLaMA2	Visual Prompt	56.35	36.11	**97.24**	**94.95**
	Audio Prompt	36.12	18.46	**95.54**	**81.82**

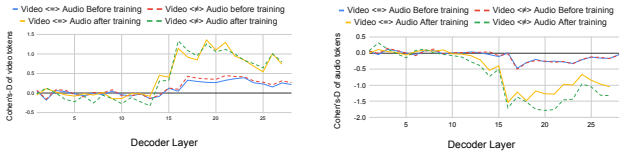
Table 7. **Model-agnostic robustness of VideoLLaMA2 on MMA-Bench**

M. Additional Generalization Experiments

We further evaluate whether modality-aware fine-tuning generalizes beyond the primary Qwen2.5-Omni-7B setting studied in the main paper. In particular, we examine cross-architecture transfer on VideoLLaMA2, zero-shot transfer to AVHBench, robustness under more realistic multimodal perturbations, and performance on open-form and open-ended QA. Together, these experiments show that the proposed training strategy improves modality grounding beyond a single architecture, benchmark, or prompt format.

M.1. Cross-Architecture Generalization on VideoLLaMA2

To test whether the proposed training strategy is specific to Qwen2.5-Omni-7B or transfers to other audio–visual MLLMs, we additionally apply modality-aware fine-tuning to VideoLLaMA2. Due to compute constraints, we fine-



(a) Video tokens (b) Audio tokens

Figure 17. Alignment-aware tuning of VideoLlama2 increases modality-specific attention reallocation under misalignment.

Table 8. **Zero-shot evaluation on AVHBench.** Performance of *Qwen2.5-Omni* before and after modality-aware fine-tuning. The first three tasks are binary classification (accuracy in %), and the last is an open-ended captioning task evaluated using METEOR.

Task	Qwen2.5-Omni	+Ours
Video-driven Audio Hallucination	71.67	**79.86**
Audio-driven Video Hallucination	80.65	**85.32**
AV Matching	49.57	49.57
AV Captioning (METEOR)	15.50	**15.60**

tune only a subset of components, namely the projection and self-attention layers, rather than performing full-model adaptation. Even under this restricted setup, VideoLLaMA2 shows consistent gains on MMA-Bench, as summarized in Table 7. The trend is consistent with that of Qwen2.5-Omni-7B, indicating that the benefits of modality-aware fine-tuning are not tied to a single backbone.

These results are also consistent with the white-box analysis in Sec. E. As shown there, baseline VideoLLaMA2 exhibits only limited attention reallocation under semantic misalignment, suggesting weak modality-selective adaptation. After fine-tuning, however, the model shows substantially stronger modality-specific reallocation (Fig. 17), mirroring the qualitative shift observed for Qwen2.5-Omni-7B in the main paper. Together, these results suggest that the proposed method improves both black-box robustness and internal modality selectivity across architectures.

M.2. Zero-shot Evaluation on AVHBench

We evaluate the zero-shot generalization ability of our modality-aware fine-tuned model on AVHBench [14], a benchmark designed to diagnose hallucinations and cross-modal inconsistencies in audio–visual MLLMs. Unlike MMA-Bench, which focuses on modality-specific reasoning under controlled conflicts, AVHBench evaluates whether models hallucinate information when modalities disagree. As summarized in Table 8, modality-aware fine-tuning improves performance on hallucination-oriented tasks. In particular, accuracy increases by **+8.2%** on Video-driven Audio Hallucination (V2A) and **+4.7%** on Audio-driven Video Hallucination (A2V). Performance on *AV Matching* remains unchanged, while *AV Captioning* shows a small improvement (+0.1 METEOR). These results indicate that strengthening modality grounding improves cross-modal consistency without negatively affecting broader reasoning or generative capabilities.

M.3. Robustness to Realistic Multimodal Perturbations

Beyond the controlled semantic swaps used in MMA-Bench, we evaluate the robustness of modality-aware fine-tuning under more realistic multimodal perturbations. Specifically, we study three additional settings:

Semantic audio swap. The original audio track is replaced with a semantically related but distinct audio sample (e.g., replacing one *clapping* instance with another sample from a different video).

Audio mixup. Multiple distinct audio sources are overlaid onto a single video track, simulating real-world scenarios where multiple sound events occur simultaneously.

Temporal shift. The audio track is shifted relative to the video by a controlled delay to simulate audio–visual desynchronization.

For **semantic audio swap** and **temporal shift**, we evaluate models using a matching-based question (*"Do the audio and video match?"*) that directly probes cross-modal consistency. For **audio mixup**, we instead evaluate modality-specific audio and visual classification questions, since the visual content remains unchanged while the audio stream contains multiple events.

As shown in Fig. 16, modality-aware fine-tuning consistently improves performance across these settings. These results demonstrate that the proposed training strategy improves robustness not only under controlled modality conflicts but also under more realistic multimodal disturbances.

M.4. Open-form and Open-ended Multimodal QA

We further evaluate whether the proposed method generalizes beyond closed-set classification tasks by considering two generative reasoning settings.

Open-form classification QA. Models generate class names in free-form text rather than selecting from predefined options. Predictions are evaluated using an LLM-based semantic judge to account for lexical variation in responses.

Open-ended QA. Models are asked to describe the audio or visual content of the video without predefined labels, using prompts such as *"Describe what you hear in the video"* or *"Describe what you see in the video."*

As shown in Fig. 16, modality-aware fine-tuning yields clear gains for open-form classification QA under both visual and audio prompts. For open-ended QA, improvements are observed under visual prompts, while audio-prompt performance remains similar. This behavior is expected because the training objective focuses primar-

ily on classification-style grounding, which transfers more strongly to structured prediction tasks than to unconstrained generative description.

N. Qualitative Analysis of Improved Modality Grounding

We present a selection of qualitative examples in Figure 18– 20 to visualize the practical improvement of alignment-aware fine-tuning compared to standard baselines. In scenarios characterized by semantic misalignment, such as a video depicting a dog paired with the sound of a phone ringing, baseline models frequently succumb to visual dominance and incorrectly predict a barking sound. Our model successfully decouples these conflicting sensory streams, correctly attending to the auditory signal despite the visual contradiction. Furthermore, the improved model demonstrates significant resilience to textual interference. When provided with misleading captions that contradict the actual sensory content, our model ignores the linguistic hallucination and grounds its response in the verified audio-visual evidence, confirming that the training process effectively reduces the over-reliance on priors from both the visual and textual modalities.

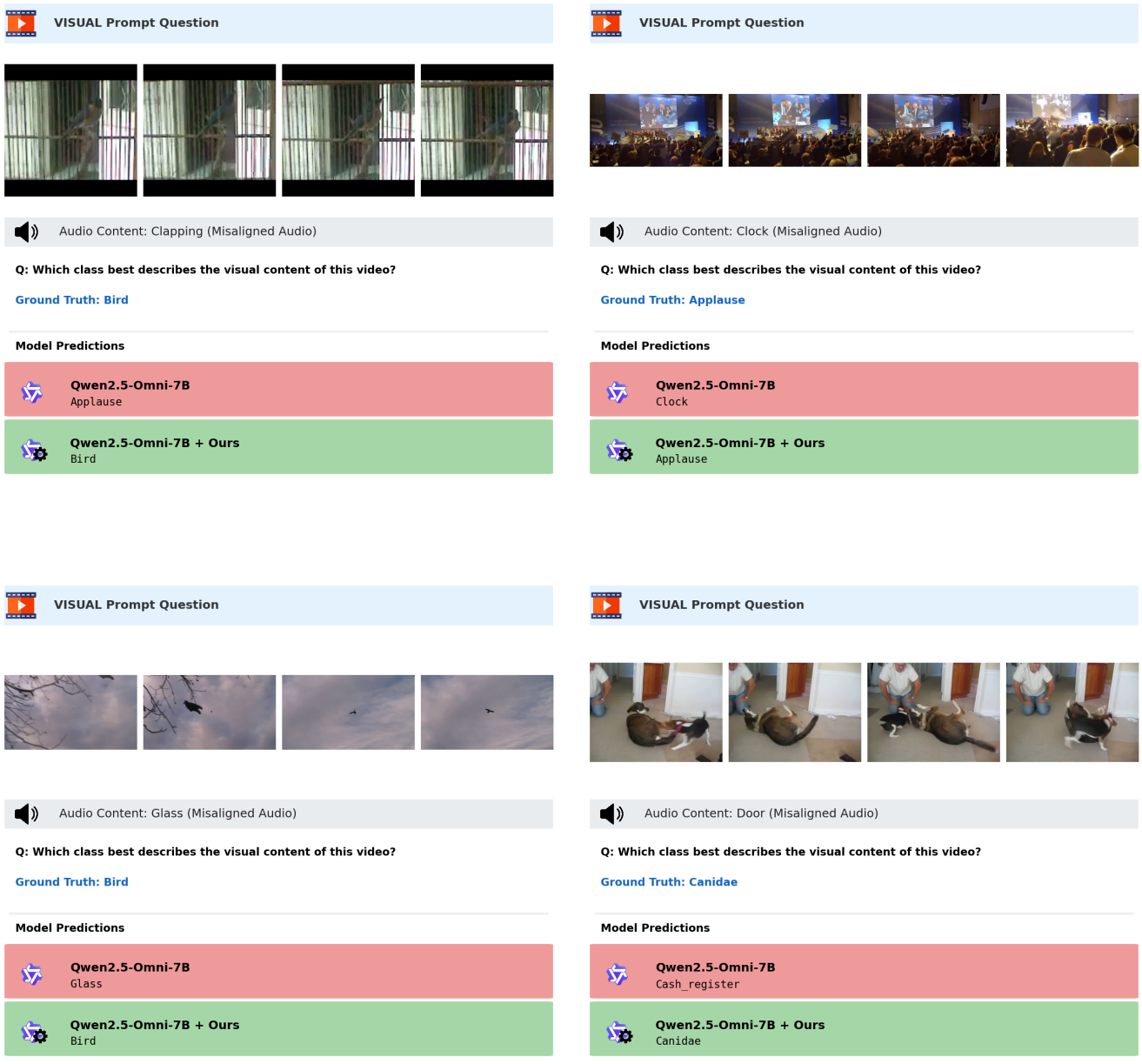


Figure 18. **Qualitative Results Gallery.** Red indicates incorrect baseline predictions, Green indicates our correct predictions. While the baseline consistently suffers from hallucinations driven by conflicting modalities, our model demonstrates robust grounding in the requested sensory input.

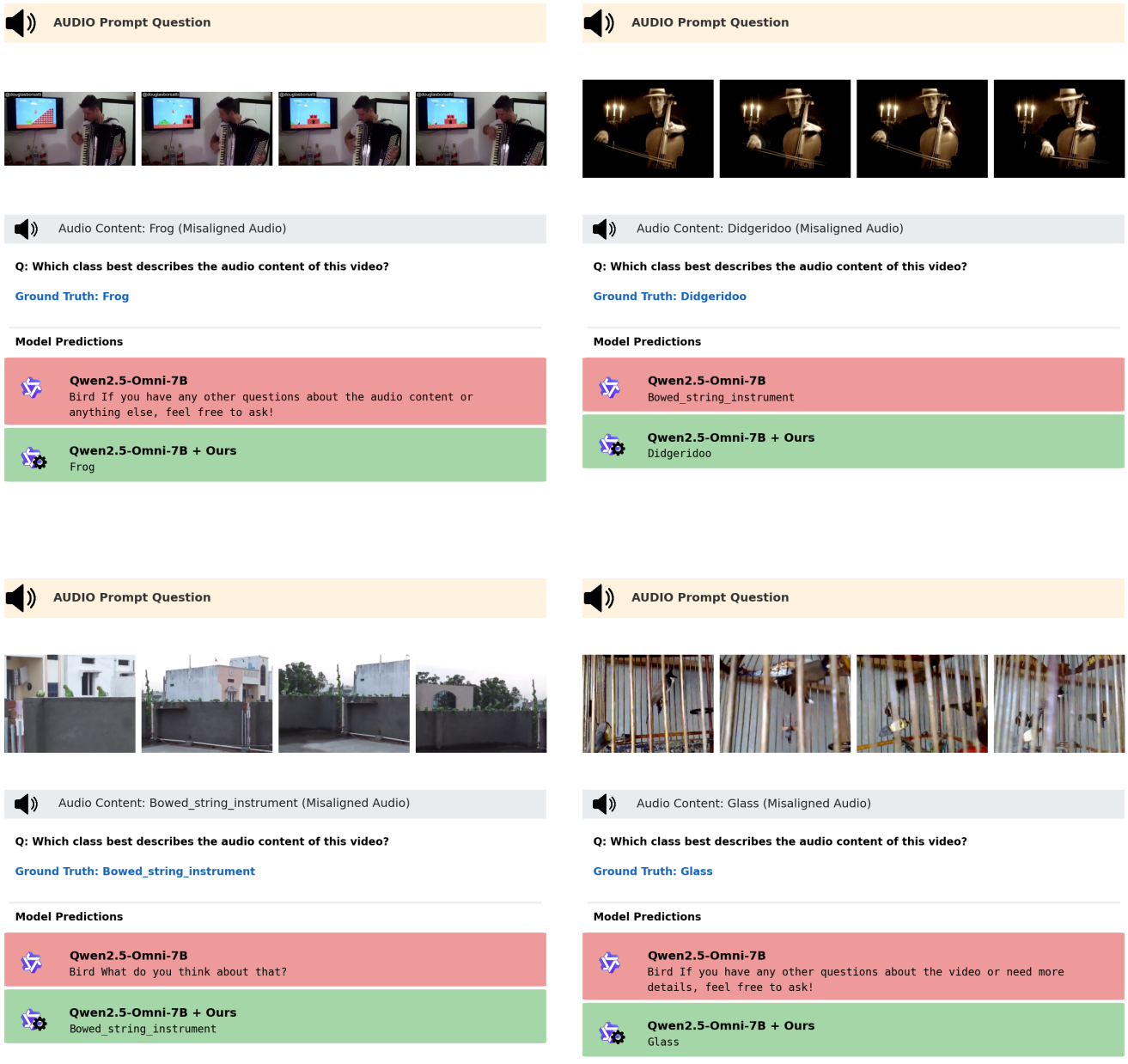


Figure 19. **Qualitative Results Gallery.** Red indicates incorrect baseline predictions, Green indicates our correct predictions. While the baseline consistently suffers from hallucinations driven by conflicting modalities, our model demonstrates robust grounding in the requested sensory input.

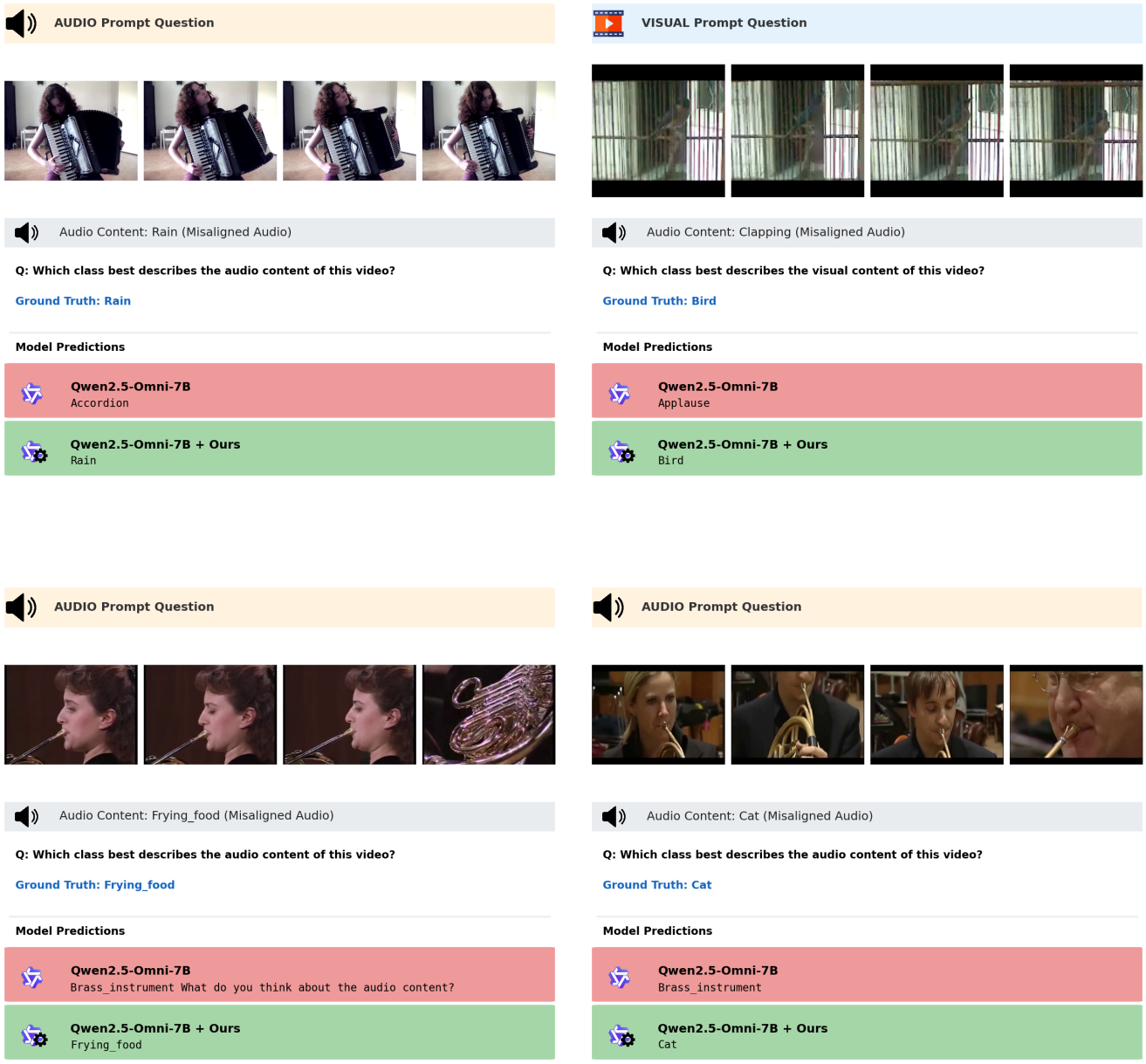


Figure 20. **Qualitative Results Gallery.** Red indicates incorrect baseline predictions, Green indicates our correct predictions. While the baseline consistently suffers from hallucinations driven by conflicting modalities, our model demonstrates robust grounding in the requested sensory input.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221, 2021. 3

[2] Yue Chen, Ziliang Zhao, and Andrew Zisserman. Vggsound-fusion: A benchmark for audio-visual learning with hierarchical labels. In *ICCV Workshops*, 2023. 2

[3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 3, 4

[4] Sheng-Yi Chou and Hung-yi Lee. Ontology-aware audio tagging with hierarchical graph attention networks. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. 2

[5] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. *arXiv preprint arXiv:2501.02135*, 2025. 9

[6] 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. 6

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 1

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[9] Jiahui Li, Bowen Zhou, and Xin Tang. Av-mix: Balanced audio-visual event dataset with ontology-guided class consolidation. In *CVPR*, 2024. 2

[10] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14394–14410, 2024. 4

[11] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. To think or not to think: A study of thinking in rule-based visual reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 10

[12] Ming Li, Jike Zhong, Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Yuxiang Lai, Chen Wei, Konstantinos Psounis, and Kaipeng Zhang. Tir-bench: A comprehensive benchmark for agentic thinking-with-images reasoning. *arXiv preprint arXiv:2511.01833*, 2025. 10

[13] Jie Mu, Wei Wang, Wenqi Liu, Tiantian Yan, and Guanglu Wang. Multimodal large language model with lora fine-tuning for multimodal sentiment analysis. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–23, 2025. 4

[14] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024. 11

[15] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 3

[16] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025. 10

[17] Huang Xie and Tuomas Virtanen. Zero-shot audio classification via semantic embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1233–1242, 2021. 3

[18] 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. 3, 4

[19] 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*, 2025. 6

[20] Sean Yang, Bernease Herman, and Bill Howe. Ontologue: Declarative benchmark construction for ontological multi-label classification. *Advances in Neural Information Processing Systems*, 35:22463–22476, 2022. 2

[21] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023. 6

[22] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 4