WHEN VISION SPEAKS FOR SOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid progress in video-capable MLLMs, we find that their apparent audio understanding in videos is often vision-driven: models rely on visual cues to infer or hallucinate acoustic information, rather than verifying the audio stream. This issue appears across both state-of-the-art open-source omni models and leading closed-source models from providers such as Google and OpenAI. We characterize this failure mode as an audio-visual *Clever Hans* effect, in which models appear (falsely) audio-grounded, but actually exploit visual-acoustic correlations without verifying whether the audio and visual streams are truly aligned. To systematically study this behavior, we introduce THUD, an intervention-driven probing framework based on three counterfactual audio edits: *Shift*, which tests temporal synchronization; *Mute*, which tests sound existence; and *Swap*, which tests audio-visual consistency. Beyond diagnosis, we further study a two-stage alignment recipe: intervention-derived preference pairs teach audio verification, while event-level general video preferences regularize the model against over-specialization. Our best 10K-sample recipe improves average performance across the three intervention dimensions by 28 percentage points, while slightly improving performance on general video and audio-visual QA benchmarks.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have rapidly advanced video understanding (Lin et al., 2024; Maaz et al., 2024; Zhang et al., 2025b). Powered by foundation models such as GPT (OpenAI, 2026), Gemini (Google DeepMind, 2026), and Qwen-VL (Team, 2025d), recent Video-LLMs (Dai et al., 2023; Zhang et al., 2023; Li et al., 2023a; Team, 2025b) can interpret dynamic scenes (Fu et al., 2024; Ren et al., 2024), answer complex questions (Patraucean et al., 2023; Li et al., 2024b), and follow instructions (Jin et al., 2023; Wang et al., 2024b). Yet, in videos with both visual and acoustic signals, such capabilities can blur the boundary between genuine audio-visual grounding and visually driven narration. For example, when shown a skateboarder crashing onto concrete, a model may describe a heavy thud even when the audio evidence is absent or misaligned (Li et al., 2023b; Guan et al., 2024; Sung-Bin et al., 2025; Cai et al., 2025). Such behavior is often interpreted as multimodal perception, but it may instead reflect an illusion of audio-visual understanding: the model predicts what a video should sound like from what it sees. While static vision-language models are known to behave like "bags-of-words" driven by text priors (Thrush et al., 2022a; Yüksekgönül et al., 2023; Tong et al., 2024), analogous prediction shortcuts in dynamic audio-visual contexts remain underexplored. This raises a central question: *Are current video-capable multimodal models truly performing audio-visual grounding, or merely hallucinating acoustic events from visual-semantic shortcuts?*

Clever Hans Effect in Audio-Visual Grounding

A video-capable multimodal model exhibits a *Clever Hans effect* when it appears audio-grounded but produces sound-related outputs primarily from visual cues rather than verified audio evidence.

We find that current video-capable MLLMs are often visually dominated when reasoning about audio-related information in sounded videos. As illustrated in Fig. 1, this shortcut can lead models to produce nearly unchanged descriptions even when the audio track changes substantially. This behavior resembles the famous *Clever Hans effect* (Pfungst, 1911), where apparent competence arises from exploiting unintended but correlated cues rather than performing the intended task. Such semantic laziness (Geirhos et al., 2020) allows models to exploit visual-semantic shortcuts and language priors instead of fine-grained audio-visual grounding that checks whether the audio and visual streams are temporally and semantically consistent (Goyal et al., 2017; Yüksekgönül et al., 2023). This failure

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

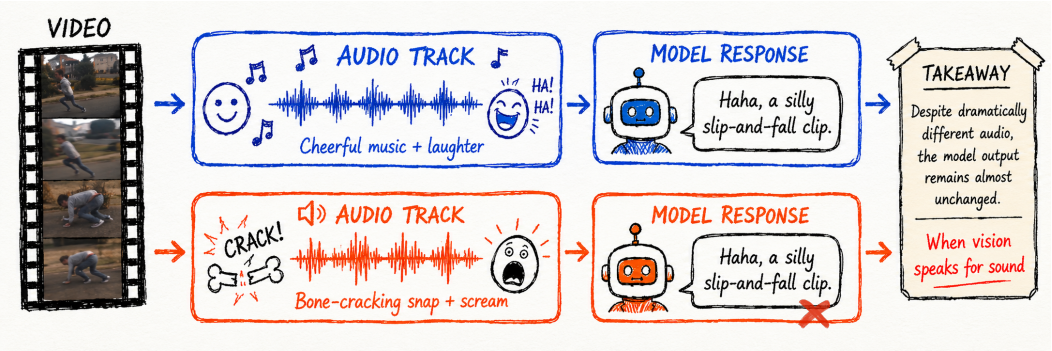


Figure 1: **When vision speaks for sound**. Given the same visual event but different audio tracks, current video-capable models produce nearly identical captions, suggesting visual-prior shortcutting rather than audio-grounded understanding.

often remains hidden because common audio-visual evaluations preserve the natural correlations that make such shortcuts effective (Gemmeke et al., 2017; Chen et al., 2020; Carreira & Zisserman, 2017): barking dogs produce barks, falling objects produce impacts, and speaking faces produce speech (Arandjelović & Zisserman, 2017; Owens & Efros, 2018). As a result, a model can appear grounded by recognizing the visual event and predicting its likely sound, without verifying whether that sound is actually present, synchronized, or physically consistent. This pseudo-alignment creates an illusion of multimodal understanding that current evaluations often fail to expose (Mangalam et al., 2023; Li et al., 2024a). To expose the Clever Hans effect, evaluation must move beyond naturally correlated videos and use controlled interventions that systematically break the audio-visual correspondences that allow visual-semantic shortcuts to succeed (Korbar et al., 2018; Morgado et al., 2020).

To this end, we introduce **THUD** (**T**emporal and **H**allucination **U**nmasking **D**iagnostics), an intervention-driven diagnostic protocol for probing audio-visual grounding in sounded videos. THUD constructs a dynamic probing space by counterfactually perturbing the audio-visual correspondences of natural videos across temporal synchronization, audio existence, and sound consistency, thereby neutralizing semantic shortcuts and exposing whether a model engages in genuinely grounded audio-visual reasoning or merely hallucinates from visual-semantic and language priors. Beyond diagnosis, we further study whether targeted post-training can mitigate these shortcuts through a family of alignment recipes that combine intervention-derived preference pairs with general video data. The best-performing recipe uses a 10K-sample mixture of counterfactual temporal preferences and event-level general video supervision, substantially improving the model's ability to detect temporal interventions, including out-of-distribution synchronization tests, while avoiding an alignment tax (Askell et al., 2021; Ouyang et al., 2022) on standard video understanding benchmarks. Additional targeted supervision on Mute and Swap further improves audio-existence and sound-consistency verification, showing that intervention-based training can be extended beyond temporal alignment. However, the same training yields only marginal gains without such targeted examples, suggesting that temporal synchronization, audio existence, and sound consistency are distinct failure modes of grounded audio-visual understanding rather than a single unified deficiency.

In summary, we make three contributions: 1) We identify and systematically expose a *Clever Hans* effect in current Video-LLMs, where models substitute genuine audio-visual grounding with visual-semantic shortcuts. Through controlled interventions, we quantify how strongly models rely on visual priors when answering sound-related questions. 2) We introduce THUD, a counterfactual diagnostic protocol that dismantles natural cross-modal correlations. By applying *Mute*, *Shift*, and *Swap* interventions, THUD audits existential, temporal, and material aspects of audio-visual grounding. 3) We evaluate preference-optimization recipes for mitigating audio-visual shortcuts. Our final 10K recipe improves average performance across ***Shift***, ***Mute***, and ***Swap*** interventions by 28 percentage points, while slightly improving general video and audio-visual understanding.

2 How Can We Align Models Beyond Visual Shortcuts?

Fig. 2 illustrates that even native multimodal models such as Gemini and Qwen3-Omni can produce plausible acoustic interpretation from visual actions alone, rather than verifying whether the corre-

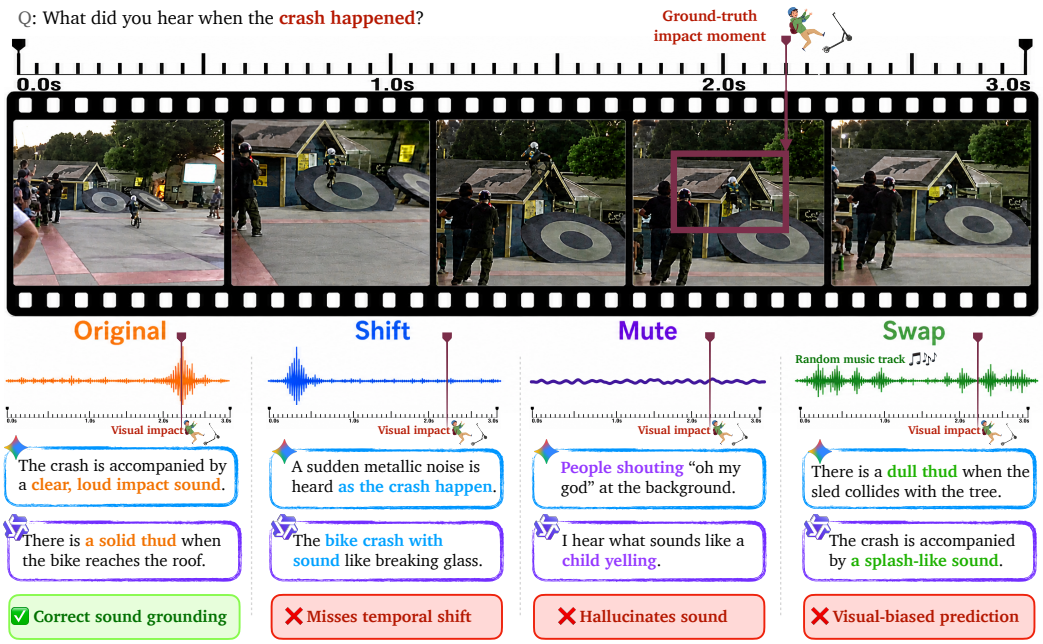


Figure 2: **Representative failure cases** under *Shift*, *Mute*, and *Swap* interventions. Gemini and Qwen3-Omni often rely on visual priors rather than verifying the audio stream, leading to missed temporal shifts, hallucinated sounds, and visually biased predictions.

sponding sound is present, temporally aligned, or consistent with its visual source. These failures motivate our intervention-driven diagnostic protocol, which deliberately breaks natural audio-visual correlations to expose models' reliance on visual-semantic shortcuts.

To align models beyond visual shortcuts, we construct training signals that task them to compare visible events against the actual audio stream rather than rely on visual priors. Our recipe turns physical audio-visual interventions into alignment data in three steps. First, we source videos with salient acoustic consequences and break natural correlations (§2.1). Second, we annotate event-time labels and construct chosen–rejected preference pairs (§2.2). And third, we combine intervention data with general video instruction data to preserve overall comprehension (§2.3).

2.1 Data Sourcing and Physical Interventions

To build intervention data for audio-visual grounding, we use the Oops dataset (Epstein et al., 2019), a collection of in-the-wild videos centered on unintentional human actions. As shown in Appx. §A.1, Oops contains many failure-centered events, such as slipping, skiing crashes, and objects breaking, that naturally induce strong expectations about the accompanying sound. This property makes it a suitable source for constructing Clever Hans-style cases: the visual content often suggests a plausible acoustic event, while the audio track determines whether that event is actually present, temporally aligned, and physically consistent with the observed action.

Formalizing interventions. Let a video be represented as $v = (x_{1:T}, a_{1:T})$, where $x_{1:T}$ denotes the visual stream and $a_{1:T}$ denotes the audio track. We construct intervened videos by applying one of three operators:

$$\tilde{v} = \mathcal{I}_k(v), \quad k \in \{\textit{Shift}, \textit{Mute}, \textit{Swap}\}. \tag{1}$$

For ***Shift***, the audio track is displaced by a temporal offset Δ:

$$\mathcal{I}_{\text{SHIFT}}(v; \Delta) = (x_{1:T}, a_{1:T}^{+\Delta}), \quad \Delta \in [-\Delta_{\max}, \Delta_{\max}]. \tag{2}$$

Here, $\Delta < 0$ corresponds to an *early* audio event, while $\Delta > 0$ corresponds to a *delayed* audio event. This intervention requires the model to compare the timing of the visible event with the timing of its acoustic consequence.

For ***Mute***, the audio signal is replaced with silence:

$$\mathcal{I}_{\text{MUTE}}(v) = (x_{1:T}, \varnothing). \tag{3}$$

For ***Swap***, the original audio is replaced with an audio track $a'_{1:T}$ from another video:

$$\mathcal{I}_{\text{SWAP}}(v, v') = (x_{1:T}, a'_{1:T}), \qquad v' = (x'_{1:T}, a'_{1:T}). \tag{4}$$

The substituted audio is acoustically plausible but physically inconsistent with the visible event, forcing the model to verify audio-visual consistency rather than rely on the most likely sound implied by vision alone. Overall, these interventions convert naturally correlated videos into controlled counterfactual cases that target temporal synchronization, sound presence, and physical consistency; a detailed summary is provided in Appx. §A.2.

2.2 Annotation and Preference Pair Construction

We annotate each source video with event-time labels used to evaluate audio-visual interventions:

$$z_i = (e_i^v, t_i^v, e_i^a, t_i^a), \tag{5}$$

where e_i^v and t_i^v denote the visual event and its timestamp, e_i^a and t_i^a denote the corresponding acoustic event and timestamp. These fields correspond to the visual event, visual time, audio event, and audio time labels in Fig. 8 (Appx. §A.1).

Cross-model verification. We use Gemini to generate initial event-time annotations because it supports direct video ingestion and can inspect both visual and audio streams. For visual timestamps, we further verify Gemini's annotations with GPT and Claude by decomposing each video into N temporally ordered frame units and asking the models to locate the visual event within the frame sequence. For audio timestamps, which require access to the acoustic stream, we cross-verify Gemini's predictions with human inspection.

Let $\mathcal{M}_v$ denote the set of visual annotator models and let $\mathcal{M}_a = \{\text{Gemini}, \text{Human}\}$ denote the audio verification sources.

$$z_i^{(m)} = (e_i^{v,m}, t_i^{v,m}, e_i^{a,m}, t_i^{a,m}), \tag{6}$$

where visual fields are available for $m \in \mathcal{M}_v$ and audio fields are available for $m \in \mathcal{M}_a$. A sample is automatically retained when both visual and acoustic timestamps agree within strict tolerances:

$$\max_{m,m' \in \mathcal{M}_v} \left| t_i^{v,m} - t_i^{v,m'} \right| \le \epsilon_v, \qquad \max_{m,m' \in \mathcal{M}_a} \left| t_i^{a,m} - t_i^{a,m'} \right| \le \epsilon_a. \tag{7}$$

Here, ϵ_v and ϵ_a denote the tolerance thresholds for visual and acoustic timestamps, respectively. Cases with model disagreement are manually inspected and corrected to ensure reliable event-time labels. We provide the annotation prompts, frame-unit construction details, agreement criteria, and manual verification protocol in Appx. §B.

Preference pair construction. The annotated intervention cases are converted into chosen–rejected preference pairs:

$$\mathcal{D}_{\text{pref}} = \left\{ \left(\tilde{v}_i, q_i, y_i^+, y_i^- \right) \right\}_{i=1}^{N}, \tag{8}$$

where $\tilde{v}_i$ is the intervened video, q_i is the diagnostic prompt, y_i^+ is the chosen response, and y_i^- is the rejected response. The chosen response explicitly verifies the audio-visual relation, while the rejected response is visually plausible but inconsistent with the audio evidence, approximating the shortcut behavior we aim to suppress. The overall annotation and intervention pipeline is summarized in Fig. 8 (Appx. §A.1).

For ***Shift***, chosen responses detect early or delayed audio, while rejected responses claim synchronization or the wrong temporal direction. For ***Mute***, chosen responses identify silence, while rejected responses hallucinate expected sounds. For ***Swap***, chosen responses flag audio-visual source inconsistency, while rejected responses accept the mismatched sound. These pairs train the model to verify audio evidence rather than follow visually plausible shortcuts. Examples are provided in Appx. §D.

Table 1: Paired diagnostic accuracy (%) of video-capable multimodal models. **Orig.** denotes naturally correlated controls, while ***Shift***, ***Mute***, and ***Swap*** denote counterfactual interventions. **Avg Gap** is the average accuracy drop, reflecting shortcut reliance.

Model	Size	Temporal Sync.		Audio Existence		Sound Consistency		Avg Gap
		Orig.	*Shift*	Orig.	*Mute*	Orig.	*Swap*	
Gemini	N/A	54.9	46.5	100.0	13.4	93.6	18.3	56.8
MiniCPM-o-4.5	9B	83.8	13.7	100.0	19.0	95.8	4.9	80.7
Nemotron-3-Omni	30B	35.9	26.8	66.2	4.2	88.7	19.9	46.6
Qwen3-Omni	30B	100.0*	1.4	95.1	0.0	75.4	37.3	77.3
Ming-Omni-2.0	100B	54.2	20.1	95.7	54.9	90.1	15.5	49.8
MiMo-V2.5	311B	73.9	9.9	99.3	2.1	89.4	15.3	78.4

2.3 Two-Stage Alignment with General Video Data

Intervention data provides targeted supervision for detecting ***Shift***, ***Mute***, and ***Swap*** failures, but may over-specialize the model to counterfactual cases. We therefore mix it with general video instruction data, whose temporally segmented annotations expose ordinary audio-visual correspondences at the event level. Appx. §A.4 summarizes this two-stage alignment pipeline.

We use FineVideo (Farré et al., 2024) as the source of general video data because its annotations are organized around time segments, describing what occurs from one timestamp range to the next. We re-annotate selected FineVideo clips with Gemini and apply human agreement checks, enriching the original segment annotations with both visual and audible event-level information. The resulting annotations are used to construct four instruction types summarized in Appx. §E.

Our training follows the standard post-training recipe of Supervised Fine Tuning (SFT) followed by preference alignment (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022). We use SFT warm-up on intervention-derived data to establish audio-aware response patterns, and then apply DPO on intervention preference pairs mixed with general video data to favor audio-verified responses over visually plausible shortcuts. The general video mixture is included to reduce over-specialization to intervention cases and preserve broad video understanding. The overall two-stage alignment pipeline is summarized in Fig. 9 (Appx. §A.4).

3 Experiments

3.1 Experimental Setup

Evaluation conditions and metrics. We evaluate audio-visual grounding under four conditions: *Original*, ***Shift***, ***Mute*** and ***Swap***. Original videos serve as positive controls with natural audio-visual correspondence, while the interventions probe audio existence, temporal synchronization, and sound consistency. We report paired accuracy for each grounding dimension.

Models. We group evaluated models by access mode. The API-tested models include Gemini-3.1-Pro (Google DeepMind, 2026), MiMo-V2.5 (Xiaomi MiMo Team, 2026), and Nemotron-3-Nano-Omni (Team, 2026a). We also query GPT-5.5 (OpenAI, 2026), but omit it from Tab. 1 because its tested interface does not support direct audio input for video; its outputs are provided in Appx. §F. The locally evaluated models include MiniCPM-o-4.5 (Cui et al., 2026), Qwen3-Omni (Team, 2025c), and Ming-flash-omni-2.0 (Team, 2025a).

Training and general capability evaluation. For controlled training experiments, we use Qwen3-Omni-30B as the trainable backbone and compare checkpoints trained with different combinations of intervention data and general video data. To test whether intervention training incurs an alignment tax, we evaluate these checkpoints on Video-MME (Fu et al., 2025), LVBench (Wang et al., 2024a), DailyOmni (Zhou et al., 2025), and WorldSense (Hong et al., 2025), which measure general video and omni-modal understanding beyond our intervention distribution. We further evaluate on VGGSoundSync (Chen et al., 2021) to test out-of-distribution temporal synchronization beyond our constructed intervention set.

3.2 Do Video-Capable Multimodal Models Rely on Visual Shortcuts?

We examine whether video-capable multimodal models verify the audio stream or infer plausible sounds from visual context. Tab. 1 reports paired diagnostic accuracy under naturally correlated *Original* controls and counterfactual interventions. Original videos serve as positive controls, while drops under ***Shift***, ***Mute***, or ***Swap*** reveal failures when natural audio-visual correlations are broken. **Avg Gap** measures the average accuracy drop from *Original* to intervention conditions, with larger values indicating a larger performance collapse under counterfactual interventions. Its formula and the LLM-judge protocol for free-form outputs are provided in Appx. §G.

Overall, most models show large drops from *Original* to intervention settings, indicating that strong performance on naturally correlated videos is fragile. MiniCPM-o-4.5 and MiMo-V2.5 have the largest gaps, 80.7% and 78.4%. Qwen3-Omni is diagnostic: its perfect original temporal-sync accuracy drops to 1.4% under ***Shift***, suggesting a synchronized-default prior rather than true temporal grounding. These results suggest that current models often rely on visual-semantic priors instead of verifying audio presence, timing, and source consistency.



Figure 3: **Failure-mode heatmap**. Red indicates higher failure; audio hallucination dominates, while temporal failures are model-specific.

Fig. 3 reveals widespread audio hallucination: Mute Hallucination ranges from 0.39 to 0.99, while Swap False-Match ranges from 0.63 to 0.95. These rates are consistently higher than the corresponding False Silence and Swap False-Mismatch rates, indicating a bias toward inventing or accepting audio content rather than rejecting real audio. Temporal perception is worse. Qwen3-Omni misses 98% of ± 2 s offsets; MiniCPM and MiMo miss roughly three quarters; and even when an offset *is* flagged, the delay/early sign is wrong about half the time, close to a random label. Definitions for each axis are given in Appx. §H.

Fig. 4 decomposes each model's predictions on the three intervention tasks. On Mute and Swap, almost all errors collapse onto Hallucinated synced, with four of six models fabricating matching audio on over 80% of muted clips and the mismatched class recovered at most 37% of the time. Hallucinated shift is negligible everywhere, indicating that models hold a strong *synced* prior and rarely entertain temporal alternatives. The Shift panel makes the consequence concrete: Qwen3-Omni

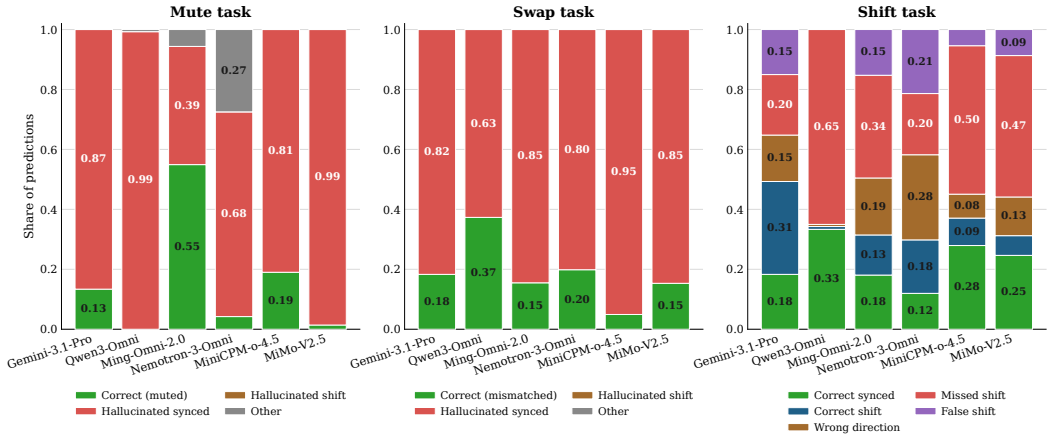


Figure 4: **Prediction breakdown** per model on the three intervention tasks. Errors cluster around a synced default, evidencing shortcut reliance over genuine audio-video alignment.

Table 2: Accuracy (%) under different alignment recipes on temporal synchronization, general video and audio-visual understanding benchmarks. We evaluate temporal grounding on **Sync** and **VGGSync**, video understanding on **V-MME** and **LVB**, audio-visual understanding on **WS** and **DO**. **Avg.** is the six-benchmark average. All DPO recipes are initialized from the SFT w/ OP checkpoint.

Recipe	Sync	VGGSync	V-MME	LVB	WS	DO	Avg.
Qwen3-Omni-30B	34.3	36.8	69.2	49.1	50.3	68.2	51.3
SFT w/ OP	73.9	46.0	68.1	46.4	49.7	67.6	58.6
SFT w/ CTP + FV-D + FV-AL	76.1	46.7	43.8	40.8	48.2	66.9	53.8
DPO w/ SP	75.4	55.7	69.3	50.9	49.8	**69.0**	61.7
DPO w/ OP + SP	76.5	56.4	69.9	47.7	49.7	68.5	61.5
DPO w/ SP + FV-D	82.2	55.4	69.1	51.5	49.8	68.0	62.7
DPO w/ OP + FV-D + LV-MCQA	83.0	**56.6**	69.2	50.4	49.9	67.6	62.8
DPO w/ CTP + FV-D	81.2	55.8	69.6	51.4	49.5	68.0	62.6
DPO w/ CTP + FV-D + LV-MCQA	82.2	55.7	69.2	51.1	49.8	67.8	62.6
DPO w/ CTP + FV-D + FV-A	82.6	55.9	69.1	50.8	49.9	67.3	62.6
Ours	**83.1**	56.4	**70.1**	**52.1**	**50.3**	67.9	**63.3**

OP: initial original-sync preference data; SP: SFT-policy negatives; CTP: counterfactual temporal preferences; FV-* and LV-MCQA denote general video preference data.

answers *synced* on 98% of inputs, while Gemini-3.1-Pro, Nemotron-3-Omni, and Ming-Omni-2.0 lose 19 to 22% of predictions to Wrong direction, showing partial sensitivity to offsets without reliable sign recovery. Errors are systematically biased toward the synced prior rather than randomly distributed, indicating that current models rely on shortcut consistency rather than genuine cross-modal alignment.

3.3 Targeted Alignment Improves Temporal Grounding Without Alignment Tax

We compare alignment recipes for Qwen3-Omni-30B to improve temporal grounding while preserving general capabilities. **Ours** uses 10,383 preference pairs combining THUD interventions (***Shift***, ***Mute***, and ***Swap***), FineVideo, and long-form audio-visual QA. Appx. §A.3 details the data construction and composition.

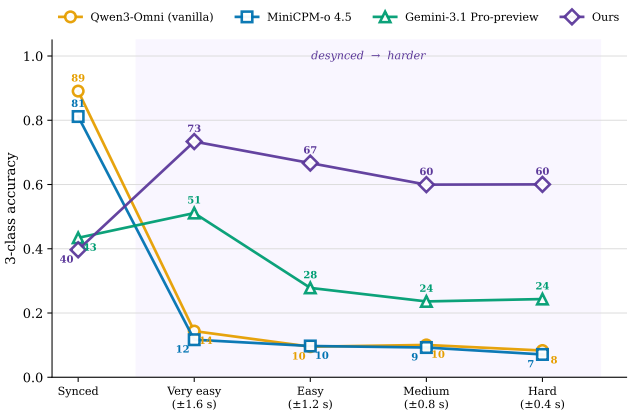
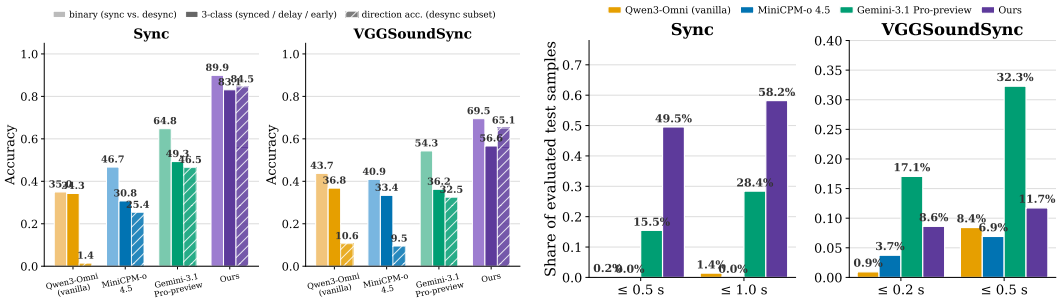


Figure 5: **Difficulty-band robustness**. Smaller offsets are harder; our model remains robust while baselines collapse under desynchronization.

Tab. 2 shows that alignment training substantially improves temporal synchronization over the vanilla Qwen3-Omni baseline. Our best 10K mixture improves **Sync** from 34.3% to 83.1% and **VGGSync** from 36.8% to 56.4%, suggesting that the model gains transferable temporal grounding rather than simply memorizing our intervention format. At the same time, it maintains or improves **V-MME**, **LVB**, and **WS**, remains competitive on **DO**, and raises the six-benchmark average accuracy from 51.3% to 63.3%. The contrast with the SFT-only mixture, which improves **Sync** but sharply hurts general benchmarks, indicates that preference alignment rather than supervised mixing is key to improving temporal grounding without incurring an alignment tax.

The recipe ablation further clarifies which data sources are responsible for this tradeoff. SFT with intervention and general video data already improves **Sync**, but substantially degrades **V-MME** and **LVB**, indicating that supervised mixing alone can over-specialize the model to intervention-style supervision. In contrast, DPO recipes recover general capability while preserving temporal gains. Self-sampled preferences provide a strong general baseline, but the best temporal results arise when targeted temporal preferences are combined with general video preference data. This



(a) Audio-visual synchronization accuracy.

(b) Localization quality under offset tolerance.

Figure 6: **Complementary synchronization results.** Left: model accuracy on binary synchronization, three-way temporal classification, and direction prediction. Right: the fraction of samples whose predicted offset is close to the ground-truth temporal displacement.

suggests that counterfactual temporal supervision supplies the grounding signal, while FineVideo and LLaVA-Video preferences regularize the model toward broad video understanding.

Fig. 5 evaluates synchronization across temporal-offset difficulty bands on **VGGSync**, using the ***Shift*** intervention from §2.1. Each band corresponds to a different offset magnitude $|\Delta|$. The high synced accuracy of vanilla Qwen3-Omni and MiniCPM-o should be read together with Fig. 4: both models strongly prefer answering "synced," making them appear accurate only when no shift is applied. Once any nonzero offset is introduced, their accuracy collapses across all bands, including large $|\Delta|$ values that should be easy to detect. Gemini-3.1-Pro follows a more expected trend, performing better on larger shifts and degrading as $|\Delta|$ becomes smaller and subtler. Our model remains stronger across all shifted bands while also reflecting the expected pattern that smaller $|\Delta|$ is harder. This suggests that temporal grounding should be judged not by synced-video accuracy alone, but by whether models show difficulty-sensitive verification under controlled audio displacement.

Fig. 6 separates temporal grounding into label-level synchronization detection and fine-grained offset localization. In Fig. 6a, our model consistently outperforms Gemini-3.1-Pro across all synchronization metrics, including binary synced/desynced classification, three-way temporal classification, and direction prediction on desynced videos. This suggests that the improvement is not limited to coarse mismatch detection, but extends to the harder problem of identifying the temporal direction of the mismatch. Fig. 6b further sharpens this distinction: most baselines rarely predict offsets close to the ground truth, whereas our model achieves the strongest localization coverage on **Sync** and remains competitive on **VGGSync**. Together, these results show that audio-visual grounding should not be measured only by whether a model flags desynchronization, but also by whether it can localize the temporal mismatch with meaningful precision.

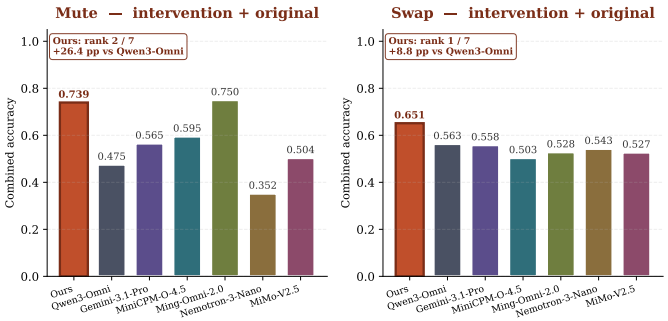


Figure 7: **Beyond temporal synchronization**. Combined Mute and Swap accuracy over original and intervened conditions.

3.4 Beyond Temporal Synchronization

Fig. 7 evaluates whether the recipe in Tab. 2 can extend beyond temporal synchronization. Starting from our best recipe, we add a small amount of Mute/Swap SFT. Re-evaluation after this additional SFT shows little change in synchronization and general-benchmark performance (Tab. 5). The resulting model ranks first on ***Swap*** and second on ***Mute***, with an average improvement of 28 percentage points over vanilla Qwen3-Omni across ***Shift***, ***Mute***, and ***Swap***.

4 Related Work

Native Omni Models and Cross-Modal Shortcuts Recent frontier multimodal models are shifting from frame-centric video-language pipelines toward native multimodal or omni-modal processing, where video, audio, images, and text are handled through a unified interface or architecture (Hurst et al., 2024; Team, 2026b; Li et al., 2025). Although such integration suggests stronger audio-visual grounding (Girdhar et al., 2023; Wu et al., 2024; Zhan et al., 2024), it does not ensure that models verify the audio stream. The shortcut behavior we observe reflects a long-standing assumption in audio-visual representation learning: natural videos provide supervision because visual and acoustic events often co-occur in synchronized and semantically aligned ways (Korbar et al., 2018; Senocak et al., 2018; Alwassel et al., 2019; Morgado et al., 2020; Chen et al., 2020). While effective for learning shared representations, these co-occurrence signals can conflate genuine grounding with statistical association (Morgado et al., 2021; Yuksekgonul et al., 2022; Thrush et al., 2022b). Models may therefore rely on visual-semantic shortcuts (Agrawal et al., 2016; Rohrbach et al., 2018): barking dogs imply barks, falling objects imply impacts, and speaking faces imply speech. Without negative cases that break these correlations (Singh et al., 2023), models can appear grounded without checking whether sound is present, synchronized, or physically consistent, producing a *Clever Hans* effect (Lapuschkin et al., 2019) in modern audio-visual models (Buch et al., 2022; Wang et al., 2024c). We address this gap using controlled audio interventions that test cross-modal verification under broken audio-visual correlations.

Preference Alignment for Video-Capable Multimodal Models Video-capable multimodal models have evolved along two related directions: video-language instruction tuning, which connects visual encoders with LLMs for video understanding (Dai et al., 2023; Lin et al., 2024), and native omni-modal modeling, which integrates video, audio, images, and text within unified interfaces or architectures (Hurst et al., 2024; Li et al., 2025; Team, 2026b). Preference-based methods such as Direct Preference Optimization (Rafailov et al., 2023) have also been adapted to video-language modeling, often using detailed captions or language-model feedback as proxies for video-grounded rewards (Zhang et al., 2025a). However, existing alignment data mainly targets helpfulness (Ouyang et al., 2022; Bai et al., 2022), visual question answering (Liu et al., 2023), instruction following (Wei et al., 2022), and safety (Zong et al., 2024; Zhu et al., 2025), with limited attention to how models use, ignore, or misattribute the audio stream. Recent work has also observed visual dominance and video-driven audio hallucination in audio-visual LLMs (Selvakumar et al., 2026; Baid et al., 2026). THUD complements these studies by separating temporal synchronization, sound existence, and cross-modal consistency through paired interventions and natural controls. In particular, Shift preserves audio content while changing its timing, testing whether semantic plausibility masks failures of temporal grounding. We further study intervention-derived preferences and their transfer to an external synchronization benchmark.

5 Conclusion

This work shows that apparent audio understanding in video-capable multimodal models can be strongly vision-driven. We identify an audio-visual *Clever Hans* effect: video-capable multimodal models can infer sounds from visual priors rather than verify audio. THUD exposes this shortcut across open and closed models through ***Shift***, ***Mute***, and ***Swap*** interventions. A two-stage alignment recipe combining intervention-derived preferences with event-level general video preferences improves audio-visual grounding while preserving broad video understanding. These findings motivate evaluating and training models under counterfactual audio-visual conditions.

AI Use Statement

We used generative AI tools for data annotation, instruction-data construction, and evaluation. Specifically, Gemini generated initial audio-visual event annotations and candidate question–answer pairs for selected FineVideo clips. GPT and Claude provided cross-model verification of visual event localization, and GPT-based text-only answering helped filter audio-dependent questions. We also used responses sampled from the SFT model as negative examples for preference construction. AI-assisted annotations and question–answer pairs were checked through cross-model agreement, human inspection of acoustic timestamps, and manual review and filtering, as described in Appx. §A.1

and Appx. §B. For evaluation, GPT-5.4 converted free-form model responses into structured predictions using the prompts documented in Appx. §G.3. We additionally used generative AI tools to assist with manuscript drafting and editing. The authors take responsibility for the final content of this work, including all AI-assisted text, annotations, and evaluation results.

Ethics Statement

This work aims to improve the reliability of multimodal models by diagnosing and reducing their reliance on visual shortcuts in audio-visual reasoning. We use video data from public or properly licensed sources for model evaluation and training under controlled audio interventions. The study does not involve human-subject experiments, crowdsourcing, or the collection of personally identifiable information. Annotation verification is performed by members of the research team. We do not perform face recognition, identity inference, biometric classification, or other person-level profiling.

Potential risks include using the interventions to construct adversarial examples or optimizing models for benchmark-specific performance rather than robust audio-visual understanding. We therefore present THUD as a diagnostic framework rather than a guarantee of deployment safety or general robustness. Models may still fail on unfamiliar sounds, noisy or edited videos, and subtle cross-modal inconsistencies. We report the limitations of the evaluated settings and encourage evaluation across diverse interventions rather than reliance on aggregate accuracy alone. If releasing data or code, we will document intended uses, licenses, and limitations, and avoid releasing sensitive or personally identifying content.

Reproducibility Statement

We describe the audio interventions, annotation procedure, preference construction, and alignment recipe in §2 and Appx. §A. Appx. §B provides annotation prompts, agreement thresholds, and manual verification rules. Appx. §C and Tab. 4 report the base model, training hyperparameters, hardware, and approximate runtimes. Appx. §D and Appx. §E provide preference-pair examples and the general instruction-data categories. Appx. §G documents the inference prompts, LLM-judge prompts, and average-gap calculation, while Appx. §H defines the failure-mode metrics. Together, these details describe how the training data, model configurations, and reported evaluations were constructed.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1955–1960, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1203. URL https://aclanthology.org/D16-1203/.

Humam Alwassel, Dhruv Kumar Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *ArXiv*, abs/1911.12667, 2019. URL https://api.semanticscholar.org/CorpusID:208513596.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617, 2017. URL https://api.semanticscholar.org/CorpusID:10769575.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Thomas Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova Dassarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, John Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment. *ArXiv*, abs/2112.00861, 2021. URL https://api.semanticscholar.org/CorpusID:244799619.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Thomas Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei,

Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022. URL https://api.semanticscholar.org/CorpusID:248118878.

Ami Baid, Zihui Xue, and Kristen Grauman. Don't let the video speak: Audio-contrastive preference optimization for audio-visual language models. *arXiv preprint arXiv:2604.14129*, 2026.

S. Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2907–2917, 2022. URL https://api.semanticscholar.org/CorpusID:249375461.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *ArXiv*, abs/2505.19616, 2025. URL https://api.semanticscholar.org/CorpusID:278905198.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017. URL https://api.semanticscholar.org/CorpusID:206596127.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020. URL https://api.semanticscholar.org/CorpusID:216522760.

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Audio-visual synchronization in the wild. In *BMVC*, 2021.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyu Sun, Yingjin Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuohang Lin, Hanyu Liu, Qi Gui, Qing-Yuan Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hong-Rui Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *CoRR*, 2026. URL https://api.semanticscholar.org/CorpusID:287915851.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 916–926, 2019. URL https://api.semanticscholar.org/CorpusID:208291335.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *ArXiv*, abs/2404.12390, 2024. URL https://api.semanticscholar.org/CorpusID:269214091.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020. doi: 10.1038/S42256-020-00257-Z. URL https://doi.org/10.1038/s42256-020-00257-z.

Jort F. Gemmeke, Daniel P. W. 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. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. URL https://api.semanticscholar.org/CorpusID:21519176.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023. URL https://api.semanticscholar.org/CorpusID:258564264.

Google DeepMind. Gemini 3. https://deepmind.google/models/gemini/, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14375–14385. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01363. URL https://doi.org/10.1109/CVPR52733.2024.01363.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *CoRR*, abs/2502.04326, 2025. doi: 10.48550/ARXIV.2502.04326. URL https://doi.org/10.48550/arXiv.2502.04326.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:53280782.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10, 2019. URL https://api.semanticscholar.org/CorpusID:67856367.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: chat-centric video understanding. *Science China Information Sciences*, 68, 2023a. URL https://api.semanticscholar.org/CorpusID:258588306.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22195–22206. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024b. doi: 10.1109/CVPR52733.2024.02095.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5971–5984. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.342. URL https://doi.org/10.18653/v1/2024.emnlp-main.342.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.679. URL https://doi.org/10.18653/v1/2024.acl-long.679.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12470–12481, 2020. URL https://api.semanticscholar.org/CorpusID:216553230.

Pedro Miguel Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12929–12940, 2021. URL https://api.semanticscholar.org/CorpusID:232417764.

OpenAI. Openai GPT-5 system card. *CoRR*, abs/2601.03267, 2026. doi: 10.48550/ARXIV.2601.03267. URL https://doi.org/10.48550/arXiv.2601.03267.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:4724792.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=HYEGXFnPoq.

Oskar Pfungst. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston, 1911.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2024. doi: 10.1109/CVPR52733.2024.01357.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Ramaneswaran Selvakumar, Kaousheik Jayakumar, S Sakshi, Sreyan Ghosh, Ruohan Gao, and Dinesh Manocha. Do audio-visual large language models really see and hear? *arXiv preprint arXiv:2604.02605*, 2026.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In-So Kweon. Learning to localize sound source in visual scenes. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4358–4366, 2018. URL https://api.semanticscholar.org/CorpusID:3841418.

Nikhil Singh, Chih-Wei Wu, Iroro Orife, and Mahdi M. Kalayeh. Looking similar, sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26897–26908, 2023. URL https://api.semanticscholar.org/CorpusID:258079032.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jTEKTdI3K9.

Inclusion AI Team. Ming-omni: A unified multimodal model for perception and generation. *CoRR*, abs/2506.09344, 2025a. doi: 10.48550/ARXIV.2506.09344. URL https://doi.org/10.48550/arXiv.2506.09344.

InternVL Team. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025b. doi: 10.48550/ARXIV.2504.10479. URL https://doi.org/10.48550/arXiv.2504.10479.

Nemotron 3 Nano Omni Team. Nemotron 3 nano omni: Efficient and open multimodal intelligence. 2026a. URL https://api.semanticscholar.org/CorpusID:287831524.

Qwen Team. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025c. doi: 10.48550/ARXIV.2509.17765. URL https://doi.org/10.48550/arXiv.2509.17765.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025d. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Qwen Team. Qwen3.5-omni technical report, 2026b. URL https://arxiv.org/abs/2604.15804.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5228–5238. IEEE, 2022a. doi: 10.1109/CVPR52688.2022.00517. URL https://doi.org/10.1109/CVPR52688.2022.00517.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022b. URL https://api.semanticscholar.org/CorpusID:248006414.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9568–9578. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00914. URL https://doi.org/10.1109/CVPR52733.2024.00914.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark. *CoRR*, abs/2406.08035, 2024a. doi: 10.48550/ARXIV.2406.08035. URL https://doi.org/10.48550/arXiv.2406.08035.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling video foundation models for multimodal video understanding. *ArXiv*, abs/2403.15377, 2024b. URL https://api.semanticscholar.org/CorpusID:268667436.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *ArXiv*, abs/2406.16338, 2024c. URL https://api.semanticscholar.org/CorpusID:270703034.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExt-GPT: Any-to-any multimodal LLM, 2024. URL https://openreview.net/forum?id=0A5o6dCKeK.

Xiaomi MiMo Team. Xiaomi mimo-v2.5: A leap in agency and multimodality. https://mimo.xiaomi.com/mimo-v2-5/, April 2026. Accessed: 2026-05-04.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Y. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *ArXiv*, abs/2210.01936, 2022. URL https://api.semanticscholar.org/CorpusID:252734947.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9637–9662, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 694–717, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.30. URL https://aclanthology.org/2025.naacl-long.30/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. URL https://arxiv.org/abs/2410.02713.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025, 2025b. URL https://openreview.net/forum?id=EElFGvt39K.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *CoRR*, abs/2505.17862, 2025. doi: 10.48550/ARXIV.2505.17862. URL https://doi.org/10.48550/arXiv.2505.17862.

Boyu Zhu, Xiaofei Wen, Wenjie Jacky Mo, Tinghui Zhu, Yanan Xie, Peng Qi, and Muhao Chen. Omniguard: Unified omni-modal guardrails with deliberate reasoning. *ArXiv*, abs/2512.02306, 2025. URL https://api.semanticscholar.org/CorpusID:283457894.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy M. Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *ArXiv*, abs/2402.02207, 2024. URL https://api.semanticscholar.org/CorpusID:267413047.

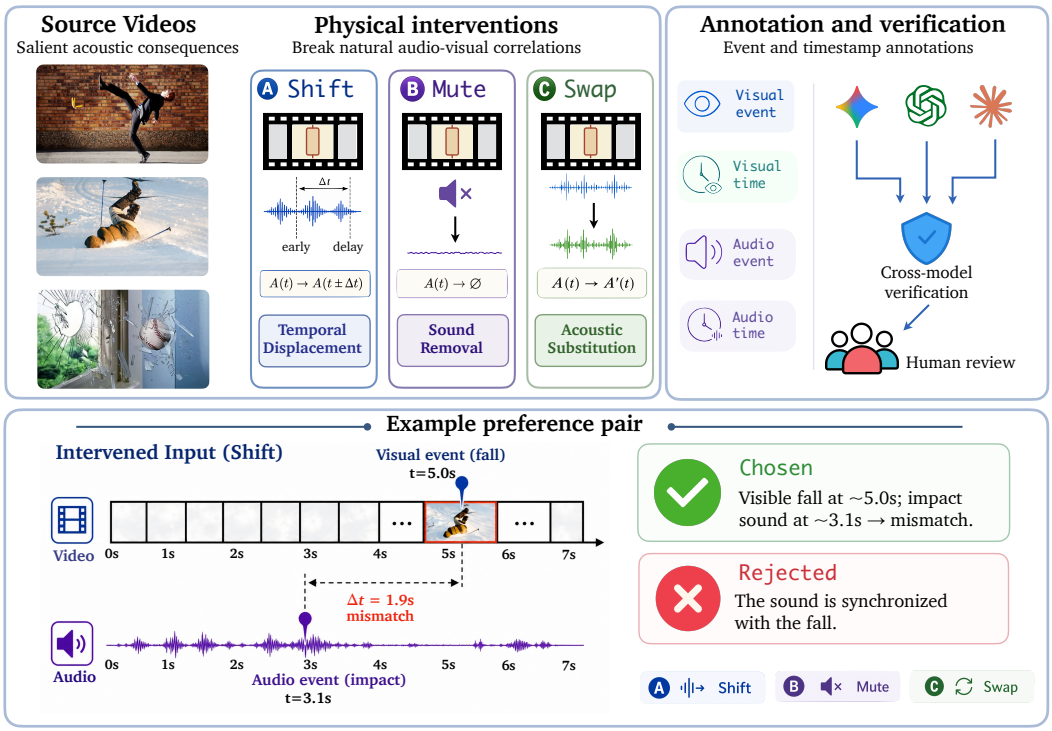


Figure 8: Pipeline for intervention **data construction**. We create ***Shift***, ***Mute***, and ***Swap*** variants from source videos with salient acoustic events, annotate visual/audio events and timestamps via cross-model verification with human review, and construct chosen–rejected preference pairs for training. The bottom panel shows a representative ***Shift*** example.

A Schematic Overviews of Data Construction and Alignment

A.1 Data Construction Pipeline

Construction pipeline. Fig. 8 summarizes the construction of our intervention-derived preference data. Gemini generates initial event-time annotations. Visual timestamps are verified with GPT and Claude through frame-unit analysis, while acoustic timestamps are checked through human inspection. After agreement filtering and manual review, we apply ***Shift***, ***Mute***, and ***Swap*** to the validated source videos. The resulting preference pairs contrast chosen responses grounded in the observed audio-visual relation with rejected responses that are visually plausible but inconsistent with the audio evidence.

Data splits. We partition the data by source video, keeping each original clip and all its intervention variants in the same split. Audio donors for Swap are sampled only within the corresponding split.

A.2 Intervention Summary

Tab. 3 summarizes the three interventions in THUD. Each intervention keeps the visual stream fixed while perturbing the audio track to target one grounding dimension: ***Shift*** probes temporal synchronization, ***Mute*** probes sound existence, and ***Swap*** probes source consistency. These controlled cases test whether models verify the observed audio or simply infer plausible sounds from visual priors.

A.3 Preference Data Sources

This section describes the preference data sources used in our alignment recipe study. Each preference example is represented as a pair of responses, where the chosen response provides the desired behavior and the rejected response provides a shortcut-prone or incorrect alternative.

Original synchronization preferences (OP). Original synchronization preferences are constructed from the annotated audio-visual event tuples introduced in §2.2. For each video, we represent the aligned visual and acoustic event as

$$z_i = (e_i^v, t_i^v, e_i^a, t_i^a), \tag{9}$$

where e_i^v and t_i^v denote the visual event and its timestamp, while e_i^a and t_i^a denote the corresponding acoustic event and timestamp. The chosen response is the annotated answer derived from the original aligned event. The rejected response is produced by perturbing one or more components of z_i, such as the visual event, visual timestamp, acoustic event, or acoustic timestamp, creating a plausible but incorrect synchronization explanation.

SFT-policy negatives (SP). Self-sampled negatives are generated from the SFT model itself. Given the same video-question input, we use the reference annotation as the chosen response and treat the SFT model's incorrect or shortcut-prone output as the rejected response. This data source encourages the model to correct its own post-SFT failure modes.

Counterfactual temporal preferences (CTP). Counterfactual temporal preferences are constructed by pairing original and shifted videos. For an original video, the chosen response corresponds to the original synchronized condition, while the response describing the shifted condition is used as the rejected response. For a shifted video, this assignment is reversed: the shifted-condition answer is chosen, and the original synchronized answer is rejected. This forces the model to distinguish true temporal alignment from visually plausible but temporally inconsistent audio.

FineVideo descriptive preferences (FV-D). FV-D is derived from the FineVideo data described in Appx. §E. We use the description, localization, and attribution tasks, which encourage the model to produce faithful video descriptions, localize relevant events, and attribute answers to appropriate visual or audio evidence.

FineVideo audio-visual QA preferences (FV-AVQA). FV-AVQA corresponds to the audio-dependent QA subset in Appx. §E. These examples ask questions that require audio evidence. Candidate questions and answers are generated by Gemini and then manually filtered. We further retain examples where GPT-based text-only answering fails, since GPT does not receive the audio stream. This filtering emphasizes cases where the answer cannot be reliably inferred from visual or textual priors alone.

FineVideo audio-visual QA long-form preferences (FV-AVQA-L). FV-AVQA-L is a long-form version of FV-AVQA. Instead of only selecting an answer option, the chosen response includes both the answer and an explanation grounded in the audio-visual evidence. This data source encourages the model to justify audio-dependent answers rather than relying on short-form guesses.

LLaVA-Video multiple-choice QA (LV-MCQA). LV-MCQA is a multiple-choice video QA dataset titled LLaVA-Video-178K (Zhang et al., 2024). We include it as a general video preference source to regularize the model toward broad video understanding and reduce over-specialization to intervention-style examples.

Final preference mixture. The final DPO training set contains 10,383 preference pairs: 5,112 THUD pairs covering Shift, Mute, and Swap (49.2%); 3,955 FineVideo pairs (38.1%); and 1,316 long-form audio-visual QA pairs (12.7%). These proportions describe the composition of the preference dataset.

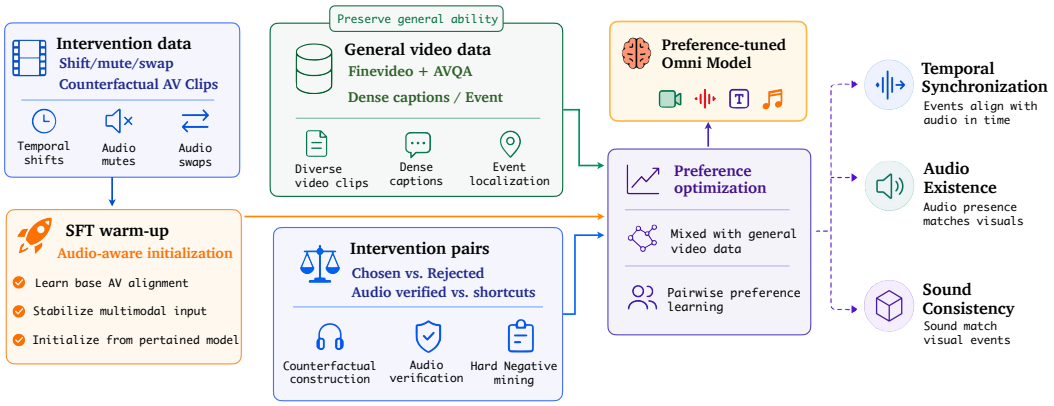


Figure 9: Two-stage intervention-driven **alignment pipeline**. Counterfactual intervention data is first used for SFT warm-up, and intervention preference pairs are then mixed with general video data during preference optimization. This design encourages audio-verified responses while preserving general video understanding.

A.4 Alignment Pipeline

Fig. 9 illustrates our two-stage post-training pipeline designed to detect ***Shift***, ***Mute***, and ***Swap*** failures while preserving general video understanding. The training process integrates our targeted intervention dataset and re-annotated general video instructions derived from FineVideo (Farré et al., 2024). In Stage 1, an SFT warm-up on intervention data establishes basic audio-aware patterns. Stage 2 applies DPO using a mixture of intervention preference pairs and general video data. The preference pairs teach the model to reject visually plausible shortcuts, while the general data acts as a regularizer to preserve broad multimodal capabilities.

Table 3: Summary of our three physical interventions. Each intervention breaks a different natural audio-visual correlation and poses a diagnostic question about audio-visual grounding.

Intervention	Operation	Broken correlation	Diagnostic question
Shift	$a_{1:T} \to a_{1:T}^{+\Delta}$	temporal synchrony	Is the sound synchronized?
Mute	$a_{1:T} \to \varnothing$	sound existence	Is any sound present?
Swap	$a_{1:T} \to a'_{1:T}$	source consistency	Does the sound match the event?

B Annotation and Verification Details

Video-to-frame-unit conversion. For visual verification with GPT and Claude, we convert each video into N temporally ordered frame units. Given a video of duration T seconds, we split it into non-overlapping windows

$$u_j = [s_j, e_j], \qquad j = 1, \ldots, N,$$

where s_j and e_j denote the start and end time of the j-th unit. From each unit, we sample representative frames and present them in temporal order, together with the timestamp range of the unit. The verifier is asked to select the unit that contains the target visual event and optionally refine the timestamp within that unit. This frame-unit format allows models without direct video ingestion to perform temporal localization over visual evidence.

Gemini annotation prompt. We use Gemini to produce the initial audio-visual event annotation. The prompt is designed to avoid generic captioning and instead force event-level localization:

> You are given a video with audio. Identify the most salient visible event that has a corresponding acoustic consequence. Return a JSON object with the following fields:
> `visual_event`: a short description of the visible event.
> `visual_time`: the timestamp in seconds when the visible event occurs.
> `audio_event`: a short description of the corresponding sound.
> `audio_time`: the timestamp in seconds when the sound occurs.
> `confidence`: high / medium / low.
> If the visual event or audio event cannot be localized reliably, return `uncertain`.

Frame-unit visual verification prompt. For GPT and Claude, we provide the ordered frame units and the candidate visual event proposed by Gemini:

> You are given temporally ordered frame units from a video. Each unit contains representative frames and a timestamp range. Target visual event: `[visual_event]`.
> Select the frame unit where this event occurs.
> Return: `unit_id`, `timestamp_range`, and a one-sentence justification. If the event is not visible or cannot be localized, return `uncertain`.

Agreement and filtering rules. We retain a sample only if it satisfies the following conditions:

1. **Visual agreement:** Gemini, GPT, and Claude localize the visual event within $\epsilon_v = 0.8$ seconds, or select overlapping frame units.
2. **Audio verification:** the acoustic event is audible and its timestamp can be verified by human inspection within $\epsilon_a = 0.5$ seconds of the Gemini prediction.
3. **Event clarity:** the visual event has a clear onset or peak moment, such as impact, fall, collision, breakage, or contact.
4. **Acoustic salience:** the corresponding sound is not dominated by unrelated background music, speech, or noise.
5. **Intervention validity:** after applying ***Shift***, ***Mute***, or ***Swap***, the correct answer remains unambiguous.

Manual review protocol. Samples failing automatic agreement are manually reviewed. We use the following decision rules. If the disagreement is due to a small boundary ambiguity, we correct the timestamp to the clearest event onset. If the visual event is partially occluded, spread over a long interval, or lacks a well-defined moment, we discard the sample. If the sound is too weak, masked by background noise, or not clearly tied to the visual event, we discard the sample. For ***Swap***, we additionally discard cases where the substituted audio is trivially unrelated or too similar to the original audio; retained swaps must be acoustically plausible but inconsistent with the visible event.

C Experimental Configuration

This section provides additional implementation details for our supervised fine-tuning (SFT), preference optimization, and evaluation experiments. All training experiments are conducted on 8*NVIDIA H200 GPUs, while evaluation experiments are conducted on either 8*NVIDIA H200 or 8*NVIDIA H100 GPUs. A single SFT run takes approximately 6 hours, and DPO training on 10K examples takes approximately 20 hours. For evaluation, the average inference time across the six datasets is approximately 5 hours per dataset.

Base model. We use Qwen3-Omni-30B-A3B-Instruct as the base omni-modal model. Video inputs are processed with audio enabled by setting `use_audio_in_video=true`.

Training configuration. We summarize the training configurations for supervised fine-tuning and preference optimization in Tab. 4. Both stages are launched with `torchrun` on a single node with 8 GPUs, using DeepSpeed ZeRO-3 for memory-efficient distributed training.

Table 4: Training configurations for the supervised fine-tuning and DPO stages.

Configuration	SFT	DPO
Initialization	Qwen3-Omni-30B-A3B-Instruct	SFT checkpoint
Fine-tuning type	Full-parameter tuning	LoRA
Epochs	3	1
Learning rate	2×10^{-6}	1×10^{-6}
Scheduler / warmup	Cosine / 0.03	Cosine / 0.03
Weight decay / grad norm	0.01 / 1.0	0.0 / 1.0
Precision	bf16	bf16
Cutoff length	131,072	131,072
Video max pixels	501,760	250,880
Audio in video	Enabled	Enabled
Batch size	1 per GPU; accum. 4; effective 32	1 per GPU; accum. 8; effective 64
Memory optimization	DeepSpeed ZeRO-3	DeepSpeed ZeRO-3
Preference loss	–	Sigmoid DPO, $\beta = 0.1$
LoRA setting	–	rank 32; alpha 64; dropout 0.05
Workers	16 preprocessing; 8 dataloader	16 preprocessing; 8 dataloader
Distributed setup	8 GPUs, single node	8 GPUs, single node
Hardware	H200 GPUs	H200 GPUs

Verification after additional Mute/Swap SFT. Table 5 compares temporal synchronization and general understanding before and after the additional Mute/Swap SFT.

Table 5: Accuracy (%) before and after additional Mute/Swap SFT. Benchmark abbreviations follow Table 2.

Checkpoint	Sync	VGGSync	V-MME	LVB	WS	DO	Avg.
Before extra SFT	83.1	56.4	70.1	52.1	50.3	67.9	63.3
After extra SFT	*82.8*	*57.0*	*68.9*	*51.7*	*50.5*	*68.2*	*63.2*

D Preference Pair Examples

Examples of Preference Pair Construction

Shift	**Chosen:** The visible fall occurs at $\sim$ 5.0s, while the impact sound is heard at $\sim$ 3.1s, indicating a synchronization mismatch. **Rejected:** The sound is synchronized with the fall.
Mute	**Chosen:** The audio track is silent throughout the clip; no music, speech, ambient noise, or sound effects are detected. **Rejected:** The animated music-video scene is described as containing female vocals, a chaotic electronic rock beat, rain, thunder, TV static, and glass-shattering effects.
Swap	**Chosen:** The visuals show an optics diffraction demonstration, but the audio describes how to use a centrifuge, indicating an audio-source mismatch. **Rejected:** The narrator explains the diffraction setup as the hand shines a smartphone light through the pen tip.

E FINEVIDEO-DERIVED GENERAL INSTRUCTION DATA

Description
Describe visible events and audible cues.

Localization
Locate visual/audio events in time.

Attribution
Infer the source or material of a cue.

Audio-dependent QA
Answer questions that require audio evidence.

F QUALITATIVE GPT-5.5 OUTPUTS (VISUAL-ONLY INPUT)

We provide representative raw outputs of GPT-5.5 across the three tasks of our test data—***Mute***, ***Shift***, and ***Swap***—when the model is given visual frames only. Since the model does not have access to the audio track, these examples illustrate how it tends to respond when forced to reason about audio without being able to hear it.

Task: *Mute* — What does the audio sound like?

```
video:  Bats & Balls Fail Compilation 47.mp4
```

GPT-5.5 raw output:
"The audio is likely from an indoor gym: basketball dribbles and sneaker squeaks as someone runs toward the hoop, followed by the thud/clank of a dunk attempt. Then there's a sudden loud crash as the backboard shatters, with glass raining down onto the floor. You can hear startled reactions—shouting, laughter, and movement—as people react and run over afterward."

Task: *Shift* — Is the audio in sync with the video?

```
video:  Are You Serious!  - Throwback Thursday Sept 2017
11_delay_2.39s.mp4
```

GPT-5.5 raw output:
"The audio and video appear to be synchronized.
The clearest sync points are the physical impact moments: the man's jump/trip over the boards, the body hitting the ground, and the subsequent movement/laughter reaction. The impact sound lines up with the visual fall closely enough that I do not perceive a meaningful lead or lag. Any offset, if present, seems very small—likely under about 0.1–0.2 seconds—and not enough to call the tracks out of sync."

Task: *Swap* — Does the audio match the video content?

```
video:  Bats & Balls Fail Compilation 47.mp4
audio swapped from:  FailArmy After Dark - That's Going To Cost You
Ep.10 (103.mp4)
```

GPT-5.5 raw output:
"yes"

G EVALUATION PROMPTS

This section documents the exact prompts used to elicit model responses on each of the three tasks, together with the GPT-based judge prompts used to parse those free-form responses into a structured prediction. For MUTE and SWAP, the numbers reported in the main text correspond to the *neutral* prompt setting, in which the model is asked an open-ended description question rather than being directly cued about the hypothesis under test. For SHIFT, a single structured prompt is used, which simultaneously asks the model to make an aligned/misaligned decision and estimate the offset.

G.1 Average Gap Calculation

We summarize shortcut reliance by the average accuracy drop from each non-intervened control to its paired counterfactual condition:

$$\Delta_{\text{shortcut}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \left(\text{Acc}_{\text{Orig},d} - \text{Acc}_{\text{Interv},d} \right), \quad \mathcal{D} = \{\text{Sync}, \text{Exist.}, \text{Consist.}\}. \tag{10}$$

Larger values indicate a larger performance collapse under counterfactual interventions. We report the gap only when all three dimensions are available. For free-form outputs, GPT-5.4 is used as an LLM judge to adjudicate predicted labels.

G.2 Inference Prompts

Mute — Inference Prompt

"Describe the audio you hear in this video."

Swap — Inference Prompt

"Describe what you see in the video and what you hear in the audio."

Shift — Inference Prompt

"Watch this video and listen to its audio carefully. Determine whether the audio and video tracks are synchronized. If they are not synchronized, identify the direction of the offset (audio delayed or audio early relative to video) and estimate the offset in seconds. Explain your reasoning."

G.3 GPT Judge Prompts

Because the inference prompts above elicit free-form text, we use GPT-5.4 as a structured-output extractor to obtain task-level predictions. The judge receives the system prompt below and the model's raw response as the user message, and returns a single JSON object. For the Mute parser, ambiguous responses are assigned to `synced` by default, as specified in the prompt. This fallback is a parsing convention and does not, by itself, establish that the response explicitly describes audible content. Response-content categories and scoring rules for the failure-mode analysis are detailed in Appx. §H.

Mute — Judge System Prompt

You are evaluating a model's response about a video whose audio has been secretly replaced with silence. The user will give you the model's response. Classify the model's assessment into exactly ONE of four categories and return ONLY valid JSON (no markdown, no explanation):

```
{"prediction": "synced"|"delay"|"early"|"muted",
"explanation": "<one sentence>"}
```

Rules:

- `synced`: *The model claims audio and video are synchronized (it is hallucinating audio).*
- `delay`: *The model claims audio is delayed relative to video (hallucinating audio + claiming offset).*
- `early`: *The model claims audio is early relative to video (hallucinating audio + claiming offset).*
- `muted`: *The model states there is no audio, or that the audio is silent, muted, or absent.*
- *If the model describes specific sounds (e.g. "thud", "music", "splash") but does not explicitly say the video is silent, classify based on its sync/desync conclusion, not as* `muted`.
- *If the model is ambiguous or gives no clear answer, default to* `synced`.

SWAP — Judge System Prompt

You are a structured-output extractor. Given a model's response describing a video's visual and audio content, identify the audio–visual relationship asserted by the response. The input may contain either original or altered audio; do not assume either condition. Classify the response into exactly ONE of the five categories below. Return ONLY valid JSON, with no text outside the JSON object:

`{"prediction": "<label>", "explanation": "<one sentence>"}`

Rules:

- *"mismatched": the response states, explicitly or through an unambiguous equivalent expression, that the audio does not correspond to the visible content or event. The exact word "mismatched" is not required.*
- *"synced": the response clearly affirms that the audio corresponds to the visible content or event, without asserting a temporal offset.*
- *"delay": the response treats the audio and visual event as semantically corresponding, but states that the audio occurs after the visual event.*
- *"early": the response treats the audio and visual event as semantically corresponding, but states that the audio occurs before the visual event.*
- *"uncertain": the response provides no clear relationship judgment, contains unresolved contradictory judgments, or states that the relationship cannot be determined.*
- *Separate descriptions of visual and audio content do not, by themselves, establish either a match or a mismatch. Do not infer a relationship judgment merely because the response mentions different visual and acoustic events.*
- *General uncertainty, confusion, or lack of detail does not establish a mismatch. Likewise, failure to explicitly report a mismatch does not establish a match.*
- *If the response clearly asserts a semantic mismatch, choose "mismatched" even if it also mentions a temporal offset. Otherwise, use "uncertain" whenever none of the other categories is supported by the response.*

SHIFT — Judge System Prompt

"You are a structured-output extractor. The user will give you a model's free-text response about audio-video synchronization. Extract the following fields and return ONLY valid JSON (no markdown, no explanation):

`{"synced": <bool>, "direction": "none"|"delay"|"early", "offset_sec": <float>, "t_v": <float or null>, "t_a": <float or null>, "explanation": "<one sentence>"}`

Rules:

- *synced: true if the model says audio and video are synchronized, false otherwise.*
- *direction: "delay" means audio comes AFTER the visual event; "early" means audio comes BEFORE the visual event; "none" if synced is true.*
- *offset_sec: estimated time gap in seconds. 0.0 if synced.*
- *t_v: the timestamp (in seconds) the model attributes to the VISUAL event. null if not mentioned.*
- *t_a: the timestamp (in seconds) the model attributes to the AUDIO event. null if not mentioned.*
- *If you cannot determine a field, use the default (true / "none" / 0.0 / null / "").*"

H FAILURE-MODE DEFINITIONS

This appendix defines the eight failure modes reported in Fig. 3. All rates lie in $[0, 1]$; higher values indicate more frequent failures.

For audio-existence evaluation, we distinguish explicit silence reports (`muted`), descriptions of audible content (`audio_described`), and visual-only responses (`visual_only`). These content categories supplement the task-level parsing in Appx. §G.3. Correct responses report silence on silenced clips and audible content on original clips. Visual-only responses are incorrect in both conditions and contribute to Audio Dodge rather than Mute Hallucination. All evaluated responses remain in the denominators of these audio-existence metrics.

A. Audio Hallucination. Errors in which the model invents or accepts audio content that is incompatible with the input.

Mute Hallucination. On videos whose audio track has been replaced with silence, the prompt is *"Describe the audio you hear in this video." Mute Hallucination* is the fraction of all responses categorized as `audio_described`: those describing concrete audio content, such as speech, music, ambient noise, or impacts, as being present in the clip. Visual-only responses are excluded from the hallucination count but remain in the denominator.

Swap False-Match. On videos whose audio track has been replaced with the soundtrack of an unrelated video, we use the neutral prompt *"Describe what you see in the video and what you hear in the audio." Swap False-Match* is the fraction of all responses that incorrectly affirm semantic correspondence between the substituted audio and the visible content. Under the parser in Appx. §G.3, these responses receive the label `synced`, `delay`, or `early`, each of which requires an affirmative correspondence judgment. A response that separately describes the visual and audio content without judging their relationship is labeled `uncertain`, rather than counted as a false match. Such responses remain in the denominator and receive no credit for correct relationship classification. Audio-description correctness is evaluated separately from relationship classification.

B. Audio Denial. Symmetric errors on naturally paired (un-intervened) videos.

False Silence. On original videos with audible content, the same neutral Mute prompt is used. *False Silence* is the fraction of all responses categorized as `muted`. Visual-only responses are not counted as false silence, although they are incorrect for audio-existence accuracy.

Swap False-Mismatch. On videos with their original audio, we use the same neutral prompt and the same response parser. *Swap False-Mismatch* is the fraction of all responses that incorrectly assert an audio–visual semantic mismatch and therefore receive the label `mismatched`. Temporal-offset claims alone do not count as semantic mismatches. Responses with no clear relationship judgment are labeled `uncertain`; they remain in the denominator, are not counted as false mismatches, and receive no credit for correct relationship classification.

C. Question Avoidance.

Audio Dodge. Fraction of responses to the neutral Mute prompt categorized as `visual_only`: the model describes the visual scene without describing audible content or explicitly reporting silence. We compute this rate separately for silenced videos and original controls, and report their arithmetic mean:

$$D_{\text{audio}} = \frac{1}{2}\left(\frac{N^{\text{Mute}}_{\text{visual_only}}}{N^{\text{Mute}}} + \frac{N^{\text{Orig}}_{\text{visual_only}}}{N^{\text{Orig}}}\right).$$

Here, N^{Mute} and N^{Orig} are the total numbers of evaluated responses in the two conditions.

D. Temporal Failures (sync task). The sync benchmark contains synced originals together with two intervention variants per video: an audio-delayed copy (DELAY) and an audio-advanced copy (EARLY), both with offsets of approximately ± 2 s. Each model's free-form response is parsed by the judge into a boolean `pred_synced` and a categorical `pred_direction` $\in \{\text{DELAY}, \text{EARLY}, \text{NONE}\}$.

Offset Blindness. Fraction of desync samples (delay or early) for which the model judges the clip to be synced. This isolates pure failure to perceive temporal misalignment.

Direction Confusion. Among the desync samples on which the model correctly judges the clip to be non-synced, the fraction for which it picks the *wrong* direction (calls a delay an early or vice versa). This isolates direction-of-offset perception from offset detection itself.

False Sync Alarm. Fraction of synced original samples for which the model claims the clip is desynced. Symmetric counterpart to Offset Blindness: false alarms on naturally aligned audio.

For Gemini-3.1-pro, per-row sync predictions were not retained, so its Offset Blindness, Direction Confusion, and False Sync Alarm are derived from the aggregate `sync_desync_accuracy`, `direction_accuracy_on_desync`, and per-category accuracies in the saved `metrics.json`, which are mathematically equivalent to the per-row counts.

I LIMITATIONS.

Our training recipe is currently evaluated on a limited set of base models, so its effectiveness across broader omni-modal model families remains to be further studied. In addition, our recipe experiments primarily validate the effect of applying DPO after SFT for improving temporal synchronization. We have not yet conducted a complete training study for the Mute and Swap settings, which probe audio existence and cross-modal consistency. Extending the recipe to these intervention types is an important direction for future work.

J NEW ASSETS

We introduce intervention-based evaluation assets for probing audio-visual grounding under three controlled settings: Mute, Swap, and Shift. These assets are intended for diagnostic evaluation, testing whether multimodal models verify the audio stream rather than relying on visual-semantic shortcuts.

For each selected video, we construct intervention variants by modifying only the audio stream. Mute removes the audio track to test audio-existence verification. Swap replaces the original audio with mismatched audio to test cross-modal consistency. Shift temporally displaces the audio to test synchronization and offset reasoning. Each example is paired with its intervention type, evaluation condition, and target label.

The assets are verified by members of the research team to ensure that the intervention is valid and the label is unambiguous. We reject examples with unclear visual events, corrupted or inaudible audio, failed interventions, or ambiguous labels. The assets are used only for model evaluation and alignment research, and are not intended as a guarantee of real-world robustness. Their limitations include restricted intervention coverage, possible residual annotation noise, and limited coverage of all real-world audio-visual failure modes.