OP-CAD: ON-POLICY CLEAN-AUDIO DISTILLATION FOR ROBUST AUDIO-VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models deployed in real-world environments encounter external noise that can interfere with their perception and understanding of multimodal inputs. We study their robustness in audio-visual understanding, focusing on question answering under environmental noise and competing speech. The challenge is to resist acoustic interference while preserving useful audio evidence. On-policy distillation provides dense teacher feedback on student-generated responses, but uniform token weighting does not explicitly prioritize positions affected by acoustic interference. We introduce OP-CAD (On-Policy Clean-Audio Distillation), a curriculum-based privileged self-distillation framework for robust audio-visual understanding. Training progresses from mild to severe environmental noise and competing speech, with selective token-level supervision at each stage. The student generates responses from corrupted audio-visual input, while a frozen teacher uses clean audio and the verified answer to supervise the same response prefixes. To allocate this supervision, OP-CAD compares teacher predictions under clean, corrupted, and visual-only contexts without revealing the answer. These matched comparisons measure sensitivity to audio removal and corruption; a bounded weighting rule emphasizes positions identified by either signal while retaining supervision throughout the response. OP-CAD outperforms the compared methods across all evaluated noise conditions. Paired analyses further show improved preservation of clean-correct answers under strong interference, with no observed aggregate clean-accuracy penalty. These results demonstrate the value of directing clean-teacher supervision toward acoustically sensitive predictions for robust audio-visual reasoning.

1 INTRODUCTION

Omni-modal large language models jointly interpret visual and acoustic information, providing a perceptual foundation for embodied agents and smartphone assistants that interact with their surroundings (Fu et al., 2024; Xu et al., 2025a;b). Yet real-world audio often contains environmental noise or background speech. Models that correctly understand a scene in quiet conditions may be misled by irrelevant sounds in noisy environments. Acoustic interference directly affects audio without directly changing the accompanying video; visual information can therefore still provide complementary cues. This motivates using the available visual context to support audio-visual understanding when audio is corrupted. We take audio-visual question answering as our starting point and study how to train omni-modal models to make correct judgments from the available audio-visual evidence under acoustic interference. Figure 1 illustrates the problem and our goal: with the video and question unchanged, interference turns the Base model's correct answer into an incorrect one, whereas OP-CAD answers correctly from the corrupted audio-visual input.

Recent work uses privileged information and on-policy self-distillation to improve audio-conditioned understanding and reasoning. CORD uses a text-conditioned teacher to guide audio reasoning (Hu et al., 2026), while EchoDistill uses a clean-audio teacher to supervise student generation under noisy audio (Lin et al., 2026). These studies demonstrate the value of richer teacher information but focus on audio-conditioned tasks. Unlike audio, visual information in real-world environments is not directly affected by these acoustic disturbances. We therefore aim to retain visual information while using privileged information from clean audio-visual input to guide learning under acoustic interference.

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.

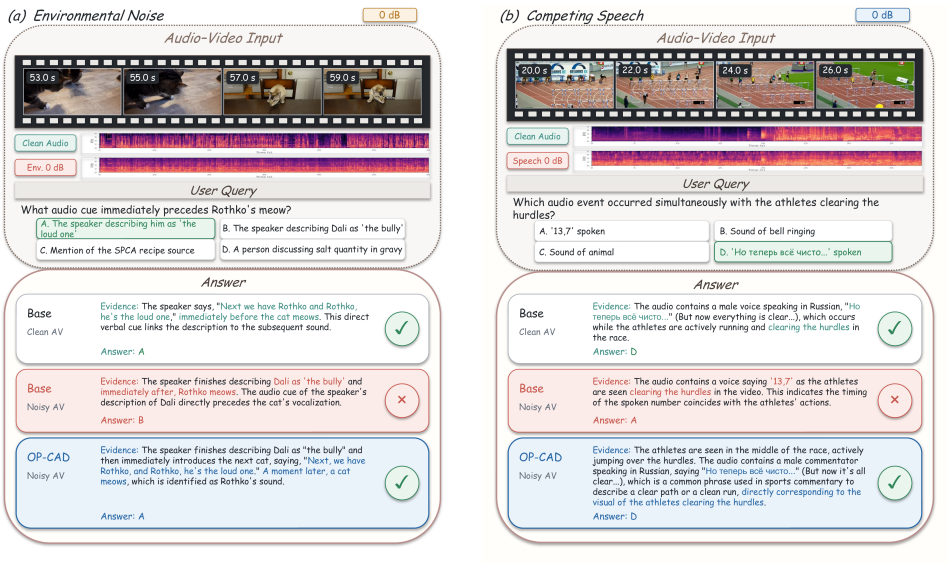


Figure 1: Prediction examples at 0 dB: (a) environmental noise; (b) competing speech.

Paired clean and corrupted audio-visual inputs make this form of supervision possible during training. The teacher receives clean audio-visual input, the student receives its corrupted counterpart, and both share the same visual information. The teacher thus has a more complete reference. OPSD provides token-level guidance along the response prefixes actually generated by the student, bringing privileged supervision onto the student's trajectories under interference (Zhao et al., 2026). However, OPSD assigns equal supervision coefficients across positions without using prediction sensitivity to audio information or acoustic interference.

We use the information differences between clean audio-visual input and restricted inputs to construct supervision targets and guide their allocation. Relative to corrupted audio-visual input, clean input provides undistorted acoustic evidence; relative to visual-only input, it provides information from the audio modality. We use both comparisons to identify sensitive positions, increasing clean-teacher supervision when either signal is strong while preserving baseline supervision throughout the response. Privileged information thus guides both what the student should learn and where stronger guidance is needed, connecting complete audio-visual observations to learning under interference.

Building on this idea, we introduce On-Policy Clean-Audio Distillation (OP-CAD), a curriculum-based privileged self-distillation framework. The student generates responses from corrupted audio-visual input, and a frozen teacher uses clean audio-visual input and the verified answer to supervise the same prefixes. Prediction differences induced by audio removal and corruption adjust the training weights. Across examples, training follows a curriculum of progressively stronger interference; within each response, the two signals determine the supervision strength at each position. At inference, the student answers directly from audio-visual input. Experiments show that OP-CAD outperforms the compared baseline training methods under both environmental noise and competing speech, with no observed decrease in aggregate clean accuracy. Ablations further demonstrate the effectiveness of sensitivity-based supervision allocation and the mild-to-severe noise curriculum. Paired analyses show that OP-CAD better preserves clean-correct answers under strong interference than OPSD.

Our contributions are threefold:

- We propose a privileged self-distillation framework for noise-robust audio-visual QA. With shared visual context, privileged information from clean audio-visual input guides responses generated by a student under corruption.
- We design a supervision allocation mechanism based on two audio condition comparisons. Using clean audio-visual input as a common reference, audio removal and corruption identify acoustically sensitive teacher predictions and determine bounded token-level supervision weights.

- We incorporate a mild-to-severe noise curriculum into audio-visual privileged self-distillation, combining progressive interference with selective supervision within responses. Under equal training budgets, this schedule exceeds randomly mixed noise levels in both clean and corrupted conditions on both benchmarks.

2 Related Work

Audio-visual understanding and acoustic robustness. Multimodal models have progressed from image and text understanding to video and audio-visual understanding (Radford et al., 2021; Li et al., 2023; Liu et al., 2023; Lin et al., 2024; Zhang et al., 2023; Cheng et al., 2024). Omni-modal large language models further integrate vision, audio, and language within a unified generative framework (Fu et al., 2024; Li et al., 2024; Xu et al., 2025a;b). Benchmarks such as WorldSense, Daily-Omni, and OmniVideoBench advance the evaluation of joint audio-visual perception and reasoning (Hong et al., 2026; Zhou et al., 2025; Li et al., 2026). Meanwhile, AVHBench, VoiceBench, RSA-Bench, and AVTrustBench examine reliability through cross-modal conflict, acoustic degradation, and modality dependence, showing that understanding of clean input alone does not ensure reliable behavior under challenging inputs (Kim et al., 2025; Chen et al., 2026; Zhang et al., 2026; Chowdhury et al., 2025). To address unreliable inputs, Watch or Listen adapts fusion to modality reliability (Hong et al., 2023), while Focus Then Listen combines source separation and modality routing for noise-robust audio-language understanding (Yin et al., 2026). These studies advance robust perception through evaluation, input processing, and modality utilization. We focus on audio-visual QA with corrupted audio and unchanged video, investigating privileged training supervision for acoustic robustness within shared visual context.

Privileged information and on-policy distillation. Knowledge distillation guides student learning through teacher predictions (Hinton et al., 2015). Learning with privileged information allows teachers to access additional information unavailable to students during training, such as more complete modalities or cleaner inputs (Vapnik & Vashist, 2009; Garcia et al., 2018; Radevski et al., 2023; Hu et al., 2025). Multimodal distillation methods use modality-specific saliency to weight distillation losses (Jin et al., 2021) or transfer audio-visual representations to a video-only student (Chen et al., 2021). For autoregressive models, on-policy distillation supplies teacher guidance on student-generated response prefixes, covering generation states actually visited by the student (Agarwal et al., 2024). OPSD applies this principle to self-distillation, using a teacher path with privileged information, such as verified answers, to supervise the student path (Zhao et al., 2026). We adopt this training mechanism with clean audio-visual input as a privileged reference for corrupted input, and further study how input-dependent prediction changes can guide token-level supervision allocation.

Audio-conditioned distillation and supervision allocation. Recent work applies privileged distillation to audio understanding and reasoning. CORD uses a text-conditioned teacher to guide audio-conditioned generation, combining importance-weighted token-level distillation with sequence-level GRPO to bridge the audio–text reasoning gap (Hu et al., 2026). EchoDistill addresses noisy audio through clean-teacher consistency signals and audio-aware rewards, guiding more reliable student responses under interference (Lin et al., 2026). These methods explore text and clean audio privileges, respectively, with a focus on audio-conditioned tasks. In contrast, we study audio-visual QA where visual information remains available despite acoustic interference. Retaining shared visual context, OP-CAD uses clean audio-visual input as a reference to measure teacher prediction changes caused by audio removal and corruption. These changes determine token-level supervision allocation, allowing privileged information from clean inputs to provide both learning targets and supervision strengths. We compare against SFT, GRPO, and uniform OPSD using the same backbone, data, and training protocol to evaluate this allocation mechanism.

3 Method

OP-CAD uses privileged information from clean audio-visual input to guide audio-visual QA under acoustic interference. The framework has three components. First, we construct clean and corrupted audio versions and select training data through paired screening. Second, supervised warm-up initializes the student and frozen teacher, followed by a curriculum of progressively stronger acoustic

interference. Finally, the teacher supervises the student's actual response prefixes, with prediction changes caused by audio removal and corruption determining supervision strength at each position. Figure 2 presents the framework.

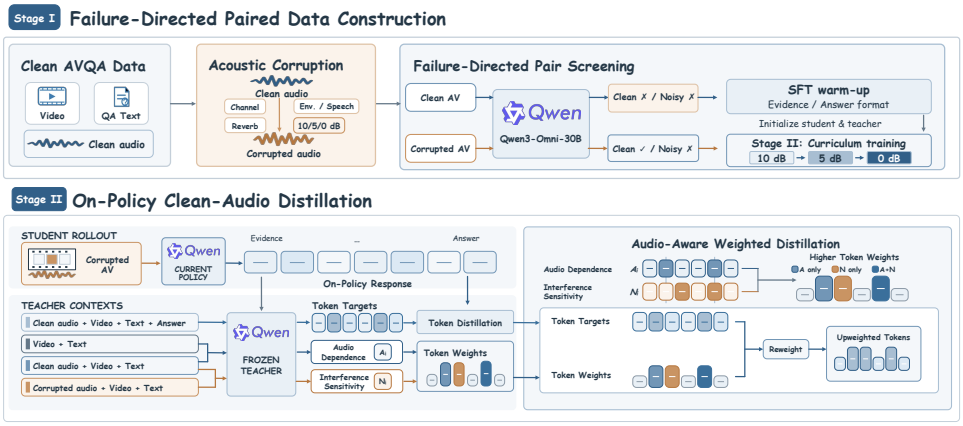


Figure 2: Overview of the OP-CAD training framework.

3.1 Paired Audio-Visual Data Construction

To learn robustness to acoustic interference, we construct paired clean and corrupted audio-visual examples. Candidate questions come from WorldSense (Hong et al., 2026), OmniBench (Li et al., 2025), and the AVQA subset of OmniInstruct-v1 (m-a-p, 2024). For each example, we modify only the audio data by applying audio degradation.

Specifically, we apply a measured room impulse response, simulate an 8-kHz μ-law channel, and mix environmental noise or competing speech into the clean audio to obtain corrupted audio at SNRs of 10, 5, and 0 dB. Environmental noise comes from MUSAN, DNS Challenge, and NOISEX-92; competing speech comes from VoxCeleb2 (Snyder et al., 2015; Reddy et al., 2020; Varga & Steeneken, 1993; Chung et al., 2018). Training and evaluation use disjoint interference recordings and room impulse responses, with disjoint speakers for competing speech.

We use the base model M_0 to answer the same question under clean and corrupted inputs, partitioning data by the paired predictions:

- Clean-correct, corrupted-incorrect examples enter the robustness training set $\mathcal{D}_{\text{rob}}$, focusing training on examples affected by acoustic interference.
- Examples answered incorrectly under both conditions enter the supervised warm-up set $\mathcal{D}_{\text{warm}}$ to learn structured evidence and answer outputs.
- Examples answered correctly under corruption are excluded from both sets, regardless of correctness on clean input.

3.2 Supervised Warm-Up and Noise Curriculum

We first fine-tune the base model on $\mathcal{D}_{\text{warm}}$ to learn the following response structure:

```
Evidence:  <audio-visual evidence supporting the answer>
Answer:  <one valid option letter>
```

Each supervised response pairs evidence generated by Gemini 3 Flash (Google DeepMind, 2025) with the dataset's verified answer. All generated warm-up evidence underwent human review to check whether the audio-visual content supported the explanation. The resulting model M_{warm} initializes both the student S_{θ_0} and teacher T. During subsequent robustness training, the student is updated while the teacher remains frozen. The base model M_0 used for data screening and the teacher T used for supervision thus serve different roles.

Robustness training follows a $10 \to 5 \to 0$ dB curriculum of increasing acoustic interference (Bengio et al., 2009). We partition $\mathcal{D}_{\text{rob}}$ into three SNR stages, each containing environmental noise and competing speech examples at its designated level and trained for one epoch. The student generates new responses with its current policy; the teacher remains fixed. The curriculum organizes difficulty across examples, while token weighting organizes supervision within each response, determining the training order and the strength of detailed teacher guidance, respectively.

3.3 On-Policy Clean-Teacher Supervision

For each training example, the student receives video v, corrupted audio a^n, question q, and options $\mathcal{C}$, denoted by $x^n = (v, a^n, q, \mathcal{C})$. Replacing the audio with its clean version a^c gives x^c; removing audio gives the visual-only input x^v. Adding the verified answer y to x^c produces the privileged teacher input $x^{c,+}$.

These inputs serve distinct purposes: $x^{c,+}$ constructs the distillation target, whereas x^c, x^n, and x^v determine supervision allocation. The latter three exclude the verified answer and share the video, question, and options, so allocation signals reflect prediction differences caused by audio conditions.

The student samples a response $\hat{r} \sim S_\theta(\cdot \mid x^n)$. Under each input condition, the teacher then predicts at the same prefixes of this response. Comparisons thus correspond to generation states actually visited by the student, without differences in prefixes from independently generated responses.

Let $\mathcal{I}_x$ index the generated positions of example x. At position i, the temperature-scaled next-token distributions of the privileged teacher and student are $p^{c,+}_{T,i} = T_\tau(\cdot \mid x^{c,+}, \hat{r}_{<i})$ and $p^n_{S,i} = S_{\theta,\tau}(\cdot \mid x^n, \hat{r}_{<i})$. Following the reverse-KL form of OPSD, the token-level distillation loss is

$$D_i = \mathrm{KL}\left(p^n_{S,i} \,\middle\|\, p^{c,+}_{T,i}\right). \tag{1}$$

The teacher uses clean audio-visual input and the verified answer to provide the learning target, while the student learns to align with it under corrupted input. The verified answer is used only for the privileged target, not for the weight computation below.

3.4 Supervision Allocation through Audio Condition Comparisons

Uniform OPSD assigns the same weight to every D_i. OP-CAD instead uses teacher prediction sensitivity to changes in audio conditions to identify positions requiring stronger supervision. Figure 3 contrasts the two approaches.

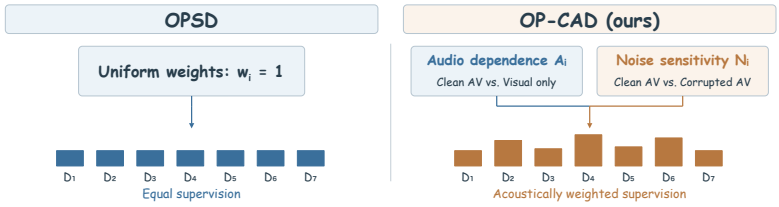


Figure 3: Token-level supervision allocation in OPSD and OP-CAD.

For $u \in \{c, n, v\}$, let $p^u_{T,i} = T_\tau(\cdot \mid x^u, \hat{r}_{<i})$ denote the teacher prediction under the corresponding input. Using clean audio-visual input as a common reference, we construct two signals:

$$A_i = \mathrm{JS}(p^c_{T,i}, p^v_{T,i}), \qquad N_i = \mathrm{JS}(p^c_{T,i}, p^n_{T,i}), \tag{2}$$

where JS denotes Jensen–Shannon divergence. Audio removal sensitivity A_i measures how much the teacher prediction changes when audio is removed while the video and prefix remain fixed. Corruption sensitivity N_i measures the change when clean audio is replaced by corrupted audio. The former captures the prediction difference associated with audio beyond the visual context; the latter captures the difference caused by acoustic interference.

Because the two divergences may have different scales, we normalize each by its maximum within the current response to obtain $\widetilde{A}_i$ and $\widetilde{N}_i$. If a signal's response maximum is at most ϵ, its normalized values are set to zero; otherwise, each value is divided by that maximum. The normalized signals

lie in $[0, 1]$ and represent relative sensitivity within a response. We then take their maximum and construct the supervision weight:

$$s_i = \max(\widetilde{A}_i, \widetilde{N}_i), \qquad w_i = 1 + \lambda \,\mathrm{stopgrad}(s_i), \quad \lambda \geq 0. \tag{3}$$

The baseline weight of 1 preserves supervision at every generated position, while the increment is bounded by λ, giving $w_i \in [1, 1 + \lambda]$. Positions sensitive to both interventions do not exceed this bound through signal addition. The $\mathrm{stopgrad}$ operation treats the computed allocation score as constant during backpropagation.

For a mini-batch $\mathcal{B}$, let $D_{x,i}$ and $w_{x,i}$ denote the distillation loss and supervision weight at position i of example x. The final objective is

$$\mathcal{L}_{\text{OP-CAD}}(\theta) = \frac{\sum_{x \in \mathcal{B}} \sum_{i \in \mathcal{I}_x} w_{x,i} D_{x,i}}{\sum_{x \in \mathcal{B}} \sum_{i \in \mathcal{I}_x} w_{x,i}}. \tag{4}$$

Each example contributes one student response, and all teacher conditions score its same prefixes. Sampled prefixes, teacher distributions, and supervision weights are detached; gradients flow only through the student distribution under corrupted input. Normalizing by the total weight preserves relative allocation while keeping the loss scale comparable to uniform OPSD. Setting $\lambda = 0$ recovers the uniform OPSD objective.

4 Experiments

4.1 Experimental Design

We evaluate all 1,197 questions in Daily-Omni and 1,000 questions in OmniVideoBench under clean audio and environmental noise or competing speech at SNRs of 10, 5, and 0 dB. We compare the original Qwen3-Omni-Instruct model (Base), the SFT warm-up checkpoint, and continued training with SFT, GRPO, OPSD, or OP-CAD. Base and SFT warm-up measure performance before and after supervised initialization. SFT, GRPO, OPSD, and OP-CAD compare robustness training strategies with shared warm-up initialization, data, and noise severity curriculum. SFT uses fixed evidence and answer targets; GRPO (Shao et al., 2024) samples eight responses per prompt with correctness and format rewards. OPSD (Zhao et al., 2026) uses the same frozen teacher and privileged target as OP-CAD but assigns uniform token weights, providing the direct control for supervision allocation.

Training uses rank-64 LoRA (Hu et al., 2022) on the language backbone and multimodal projectors, with frozen audio and visual encoders, a learning rate of 10^{-5}, and an effective batch size of 32. OPSD and OP-CAD sample one response of up to 128 new tokens per example, using the complete vocabulary with $\tau = 1$ and OP-CAD weight strength $\lambda = 0.5$. We report deterministic multiple-choice accuracy and the *harm rate*, the fraction of each model's clean-correct answers that become incorrect under corruption.

Table 1: Full test accuracy (%). Env./Speech: three-SNR averages. Bold: best result. Superscripts: changes from SFT warm-up in percentage points, using displayed scores.

	Daily-Omni			OmniVideoBench		
Method	Clean	Env.	Speech	Clean	Env.	Speech
Base	67.75	61.38	57.84	37.40	36.70	34.80
SFT warm-up	68.76	61.74	59.82	38.10	36.80	35.93
+ SFT	68.42 $^{\downarrow 0.34}$	61.88 $^{\uparrow 0.14}$	59.12 $^{\downarrow 0.70}$	39.00 $^{\uparrow 0.90}$	36.23 $^{\downarrow 0.57}$	36.20 $^{\uparrow 0.27}$
+ GRPO	68.34 $^{\downarrow 0.42}$	62.91 $^{\uparrow 1.17}$	59.93 $^{\uparrow 0.11}$	41.50 $^{\uparrow 3.40}$	37.87 $^{\uparrow 1.07}$	36.87 $^{\uparrow 0.94}$
+ OPSD	69.34 $^{\uparrow 0.58}$	63.49 $^{\uparrow 1.75}$	60.62 $^{\uparrow 0.80}$	41.40 $^{\uparrow 3.30}$	37.40 $^{\uparrow 0.60}$	35.13 $^{\downarrow 0.80}$
+ OP-CAD	**70.09** $^{\uparrow 1.33}$	**66.97** $^{\uparrow 5.23}$	**63.55** $^{\uparrow 3.73}$	**41.60** $^{\uparrow 3.50}$	**40.23** $^{\uparrow 3.43}$	**38.93** $^{\uparrow 3.00}$

4.2 Robustness Across Acoustic Conditions

Table 1 separates the effect of supervised initialization from subsequent robustness training. Warm-up improves accuracy on clean and corrupted inputs over Base on both benchmarks, but continued

training does not yield uniform gains across strategies and conditions. OP-CAD achieves the highest aggregate accuracy among the listed methods in clean audio and both interference families.

OPSD provides the direct comparison for supervision allocation because it shares OP-CAD's teacher target and training protocol. OP-CAD improves environmental noise and competing speech averages by 3.48 and 2.92 percentage points on Daily-Omni and by 2.83 and 3.80 points on OmniVideoBench. It also exceeds OPSD at every tested SNR. These results show that weighting supervision by acoustic sensitivity improves robustness over uniform distillation with the same teacher target.

Table 2: Daily-Omni accuracy (%) by question type and duration. Superscripts: changes from Base within each condition in percentage points, using displayed scores. Bold: best per column.

Cond.	Method	*Daily-Omni* Question Type						Video Duration		Avg.
		AV Align	Comp.	Ctx. Und.	Evt. Seq.	Infer.	Reas.	30s	60s	
Clean	Base	58.82	77.86	60.62	60.13	81.82	81.14	68.01	67.45	67.75
	+ SFT	**61.34**$^{\uparrow 2.52}$	77.10$^{\downarrow 0.76}$	64.77$^{\uparrow 4.15}$	57.84$^{\downarrow 2.29}$	83.77$^{\uparrow 1.95}$	80.57$^{\downarrow 0.57}$	68.62$^{\uparrow 0.61}$	**68.18**$^{\uparrow 0.73}$	68.42$^{\uparrow 0.67}$
	+ GRPO	59.24$^{\uparrow 0.42}$	79.39$^{\uparrow 1.53}$	63.73$^{\uparrow 3.11}$	59.80$^{\downarrow 0.33}$	82.47$^{\uparrow 0.65}$	80.00$^{\downarrow 1.14}$	70.48$^{\uparrow 2.47}$	65.82$^{\downarrow 1.63}$	68.34$^{\uparrow 0.59}$
	+ OPSD	56.72$^{\downarrow 2.10}$	77.86$^{\leftrightarrow 0.00}$	**68.91**$^{\uparrow 8.29}$	**60.78**$^{\uparrow 0.65}$	84.42$^{\uparrow 2.60}$	**82.29**$^{\uparrow 1.15}$	71.41$^{\uparrow 3.40}$	66.91$^{\downarrow 0.54}$	69.34$^{\uparrow 1.59}$
	+ OP-CAD	60.50$^{\uparrow 1.68}$	**80.15**$^{\uparrow 2.29}$	**68.91**$^{\uparrow 8.29}$	59.80$^{\downarrow 0.33}$	**85.06**$^{\uparrow 3.24}$	81.71$^{\uparrow 0.57}$	**72.18**$^{\uparrow 4.17}$	67.64$^{\uparrow 0.19}$	**70.09**$^{\uparrow 2.34}$
Env.	Base	49.30	70.99	59.93	54.68	75.97	71.05	62.34	60.24	61.38
	+ SFT	54.06$^{\uparrow 4.76}$	69.72$^{\downarrow 1.27}$	59.93$^{\leftrightarrow 0.00}$	53.81$^{\downarrow 0.87}$	75.54$^{\downarrow 0.43}$	70.86$^{\downarrow 0.19}$	62.44$^{\uparrow 0.10}$	61.21$^{\uparrow 0.97}$	61.88$^{\uparrow 0.50}$
	+ GRPO	50.70$^{\uparrow 1.40}$	74.81$^{\uparrow 3.82}$	60.62$^{\uparrow 0.69}$	55.12$^{\uparrow 0.44}$	76.62$^{\uparrow 0.65}$	74.67$^{\uparrow 3.62}$	64.45$^{\uparrow 2.11}$	61.09$^{\uparrow 0.85}$	62.91$^{\uparrow 1.53}$
	+ OPSD	51.54$^{\uparrow 2.24}$	73.03$^{\uparrow 2.04}$	60.79$^{\uparrow 0.86}$	55.66$^{\uparrow 0.98}$	79.44$^{\uparrow 3.47}$	75.24$^{\uparrow 4.19}$	65.79$^{\uparrow 3.45}$	60.79$^{\uparrow 0.55}$	63.49$^{\uparrow 2.11}$
	+ OP-CAD	**54.76**$^{\uparrow 5.46}$	**77.86**$^{\uparrow 6.87}$	**65.28**$^{\uparrow 5.35}$	**57.84**$^{\uparrow 3.16}$	**83.12**$^{\uparrow 7.15}$	**79.05**$^{\uparrow 8.00}$	**68.62**$^{\uparrow 6.28}$	**65.03**$^{\uparrow 4.79}$	**66.97**$^{\uparrow 5.59}$
Speech	Base	47.34	66.92	58.03	50.22	71.43	66.48	58.22	57.39	57.84
	+ SFT	48.74$^{\uparrow 1.40}$	67.18$^{\uparrow 0.26}$	56.82$^{\downarrow 1.21}$	53.49$^{\uparrow 3.27}$	74.03$^{\uparrow 2.60}$	66.48$^{\leftrightarrow 0.00}$	59.61$^{\uparrow 1.39}$	58.55$^{\uparrow 1.16}$	59.12$^{\uparrow 1.28}$
	+ GRPO	46.50$^{\downarrow 0.84}$	72.01$^{\uparrow 5.09}$	57.86$^{\downarrow 0.17}$	52.83$^{\uparrow 2.61}$	74.24$^{\uparrow 2.81}$	71.24$^{\uparrow 4.76}$	61.15$^{\uparrow 2.93}$	58.48$^{\uparrow 1.09}$	59.93$^{\uparrow 2.09}$
	+ OPSD	47.48$^{\uparrow 0.14}$	74.05$^{\uparrow 7.13}$	56.82$^{\downarrow 1.21}$	53.49$^{\uparrow 3.27}$	75.11$^{\uparrow 3.68}$	72.38$^{\uparrow 5.90}$	63.01$^{\uparrow 4.79}$	57.82$^{\uparrow 0.43}$	60.62$^{\uparrow 2.78}$
	+ OP-CAD	**50.42**$^{\uparrow 3.08}$	**75.06**$^{\uparrow 8.14}$	**63.04**$^{\uparrow 5.01}$	**55.77**$^{\uparrow 5.55}$	**77.49**$^{\uparrow 6.06}$	**74.67**$^{\uparrow 8.19}$	**65.33**$^{\uparrow 7.11}$	**61.45**$^{\uparrow 4.06}$	**63.55**$^{\uparrow 5.71}$

Tables 2 and 3 examine whether aggregate improvements are concentrated in a particular task or duration. OP-CAD exceeds OPSD in every corrupted-condition category and duration column. The observed gains extend across multiple question types, audio types, and video lengths.

Table 3: OmniVideoBench accuracy (%) by audio type and duration. Superscripts: changes from Base within each condition in percentage points, using displayed scores. Bold: best per column.

Cond.	Method	*OmniVideoBench* Audio Type			Video Duration				Avg.
		Music	Sound	Speech	(0,1] min	(1,5] min	(5,10] min	(10,30] min	
Clean	Base	30.77	32.65	39.11	49.12	35.29	36.40	33.33	37.40
	+ SFT	32.97$^{\uparrow 2.20}$	37.41$^{\uparrow 4.76}$	40.03$^{\uparrow 0.92}$	47.95$^{\downarrow 1.17}$	39.41$^{\uparrow 4.12}$	35.96$^{\downarrow 0.44}$	35.25$^{\uparrow 1.92}$	39.00$^{\uparrow 1.60}$
	+ GRPO	**36.26**$^{\uparrow 5.49}$	36.73$^{\uparrow 4.08}$	**43.04**$^{\uparrow 3.93}$	49.12$^{\leftrightarrow 0.00}$	40.59$^{\uparrow 5.30}$	39.47$^{\uparrow 3.07}$	**39.46**$^{\uparrow 6.13}$	41.50$^{\uparrow 4.10}$
	+ OPSD	28.57$^{\downarrow 2.20}$	**40.82**$^{\uparrow 8.17}$	**43.04**$^{\uparrow 3.93}$	48.54$^{\downarrow 0.58}$	42.06$^{\uparrow 6.77}$	**40.79**$^{\uparrow 4.39}$	36.40$^{\uparrow 3.07}$	41.40$^{\uparrow 4.00}$
	+ OP-CAD	32.97$^{\uparrow 2.20}$	39.46$^{\uparrow 6.81}$	**43.04**$^{\uparrow 3.93}$	**49.71**$^{\uparrow 0.59}$	**43.53**$^{\uparrow 8.24}$	38.60$^{\uparrow 2.20}$	36.40$^{\uparrow 3.07}$	**41.60**$^{\uparrow 4.20}$
Env.	Base	25.64	**38.10**	37.75	**44.83**	35.49	33.92	35.38	36.70
	+ SFT	31.87$^{\uparrow 6.23}$	37.64$^{\downarrow 0.46}$	36.48$^{\downarrow 1.27}$	44.64$^{\downarrow 0.19}$	35.39$^{\downarrow 0.10}$	32.31$^{\downarrow 1.61}$	35.25$^{\downarrow 0.13}$	36.23$^{\downarrow 0.47}$
	+ GRPO	30.40$^{\uparrow 4.76}$	35.60$^{\downarrow 2.50}$	39.20$^{\uparrow 1.45}$	41.13$^{\downarrow 3.70}$	37.84$^{\uparrow 2.35}$	38.60$^{\uparrow 4.68}$	35.12$^{\downarrow 0.26}$	37.87$^{\uparrow 1.17}$
	+ OPSD	27.84$^{\uparrow 2.20}$	37.87$^{\downarrow 0.23}$	38.45$^{\uparrow 0.70}$	43.47$^{\downarrow 1.36}$	37.25$^{\uparrow 1.76}$	38.74$^{\uparrow 4.82}$	32.44$^{\downarrow 2.94}$	37.40$^{\uparrow 0.70}$
	+ OP-CAD	**33.33**$^{\uparrow 7.69}$	**38.10**$^{\leftrightarrow 0.00}$	**41.47**$^{\uparrow 3.72}$	44.44$^{\downarrow 0.39}$	**39.22**$^{\uparrow 3.73}$	**40.94**$^{\uparrow 7.02}$	**38.19**$^{\uparrow 2.81}$	**40.23**$^{\uparrow 3.53}$
Speech	Base	25.27	36.73	35.56	40.94	33.43	33.04	34.10	34.80
	+ SFT	31.87$^{\uparrow 6.60}$	38.10$^{\uparrow 1.37}$	36.35$^{\uparrow 0.79}$	42.69$^{\uparrow 1.75}$	35.00$^{\uparrow 1.57}$	34.65$^{\uparrow 1.61}$	34.87$^{\uparrow 0.77}$	36.20$^{\uparrow 1.40}$
	+ GRPO	31.50$^{\uparrow 6.23}$	36.05$^{\downarrow 0.68}$	37.66$^{\uparrow 2.10}$	39.57$^{\downarrow 1.37}$	37.65$^{\uparrow 4.22}$	37.43$^{\uparrow 4.39}$	33.59$^{\downarrow 0.51}$	36.87$^{\uparrow 2.07}$
	+ OPSD	29.30$^{\uparrow 4.03}$	36.51$^{\downarrow 0.22}$	35.56$^{\leftrightarrow 0.00}$	42.69$^{\uparrow 1.75}$	34.71$^{\uparrow 1.28}$	36.11$^{\uparrow 3.07}$	29.89$^{\downarrow 4.21}$	35.13$^{\uparrow 0.33}$
	+ OP-CAD	**32.97**$^{\uparrow 7.70}$	**41.95**$^{\uparrow 5.22}$	**39.06**$^{\uparrow 3.50}$	**45.03**$^{\uparrow 4.09}$	**38.73**$^{\uparrow 5.30}$	**38.01**$^{\uparrow 4.97}$	**36.02**$^{\uparrow 1.92}$	**38.93**$^{\uparrow 4.13}$

4.3 Component Analysis

Supervision signals. We compare audio removal sensitivity alone, corruption sensitivity alone, and their combination while fixing initialization, curriculum, and teacher targets (Table 4). Both individual signals improve all four corrupted-condition averages over uniform OPSD. The combined rule performs best in three aggregates; on OmniVideoBench competing speech, corruption sensitivity alone scores 39.00% versus 38.93% for OP-CAD. Both signals therefore provide useful allocation criteria, and their combination improves most evaluated noise-family averages.

Table 4: Accuracy (%) with single-signal and combined allocation. Bold marks the best result.

	Daily-Omni			OmniVideoBench		
Variant	Clean	Env.	Speech	Clean	Env.	Speech
OPSD	69.34	63.49	60.62	41.40	37.40	35.13
Audio-rem. proxy	69.34	64.47	61.71	40.70	39.63	38.50
Corr.-sens. proxy	68.76	65.41	62.13	**43.80**	39.73	**39.00**
OP-CAD	**70.09**	**66.97**	**63.55**	41.60	**40.23**	38.93

Noise curriculum. We compare OP-CAD's 10, 5, then 0 dB curriculum with randomly mixed SNRs under matched total training steps and data budgets (Table 5). Curriculum learning performs better than random mixing. From the displayed scores, environmental noise and competing speech gains are 1.81 and 0.88 percentage points on Daily-Omni, and 1.96 and 1.80 points on OmniVideoBench. With the training budget held fixed, these gains demonstrate the effectiveness of the mild-to-severe curriculum for improving acoustic robustness.

Table 5: Curriculum order ablation: accuracy (%) under matched training budgets.

Training schedule	Daily-Omni			OmniVideoBench		
	Clean	Env.	Speech	Clean	Env.	Speech
Randomly mixed	67.67	65.16	62.67	39.80	38.27	37.13
Curriculum learning	**70.09**	**66.97**	**63.55**	**41.60**	**40.23**	**38.93**

4.4 Prediction Stability under Interference

Aggregate noisy accuracy combines retained clean successes and newly correct answers. Figure 4 separates these outcomes using two measures at 0 dB. Harm rate measures the fraction of each model's own clean-correct answers that become incorrect; clean-correct retention measures noisy accuracy on the fixed set of questions that Base answers correctly under clean audio. The former assesses each model's vulnerability to interference, while the latter evaluates all models on the same questions that Base answers correctly under clean audio.

OP-CAD has lower harm rates than OPSD in all four conditions and the lowest rate among the compared methods in three. The largest reduction occurs on OmniVideoBench competing speech, from 39.86% to 33.65%, with a paired difference of -6.20 points. SFT has a slightly lower harm rate of 32.05% in this condition.

OP-CAD achieves the highest retention on Base's clean-correct questions in three conditions. Under OmniVideoBench environmental noise, Base retains 71.12% accuracy versus 66.04% for OP-CAD, indicating that improved full test accuracy does not entail better preservation of every baseline success.

The overall degradation is also smaller: under competing speech at 0 dB on OmniVideoBench, the net loss of correct answers from clean to noisy input is 34 for OP-CAD versus 71 for OPSD. Together with the clean results in Table 1, this shows improved resistance to interference without an observed decrease in aggregate clean accuracy.



Figure 4: Prediction stability at 0 dB: harm rate (left) and clean-correct retention (right).

4.5 Training Example of Supervision Allocation

Figure 5 shows how the two signals distribute supervision on a recorded training rollout under competing speech at 5 dB. Audio removal sensitivity is strongest around "voice" and "sound," whereas corruption sensitivity peaks around "observational" and "flowing." The signals therefore emphasize different positions in this example. Their combination retains both sets of peaks while preserving baseline supervision elsewhere, illustrating how the weighting rule operates along a student's response under shared visual context.

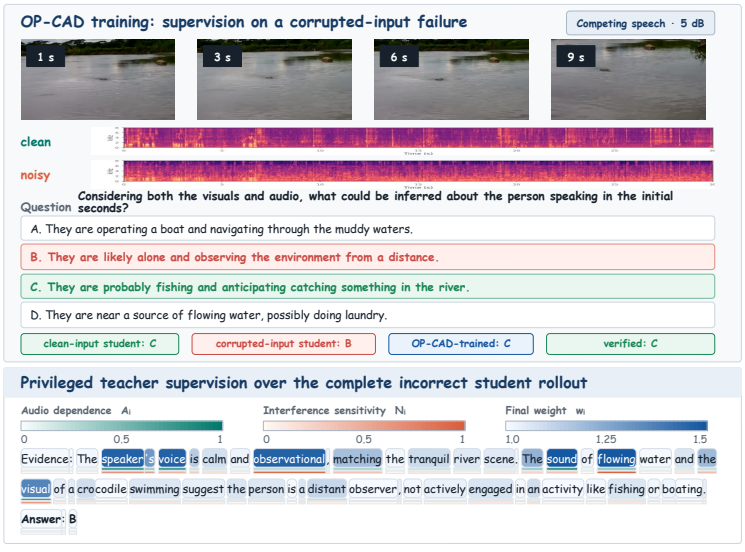


Figure 5: Token-level supervision allocation on an OP-CAD training rollout.

5 Conclusion

OP-CAD weights clean-teacher supervision by the teacher's sensitivity to audio removal and corruption. On two audio-visual benchmarks, it improves accuracy on noisy input over uniform OPSD under matched training conditions, with no observed decrease in aggregate clean accuracy. Single-signal ablations show that the two comparisons are useful but that their combination is not best in every condition. The method changes supervision during training; inference uses the student alone, without a clean audio reference, verified answer, or additional teacher passes. The current evidence is limited to one backbone, one run per method, and controlled acoustic corruptions.

AI USE STATEMENT

OpenAI Codex assisted with manuscript editing, figure layout, reference formatting, methodological feedback, interpretation of results, table recomputation, and implementation and testing of the accompanying code. Gemini 3 Flash generated the offline evidence used in supervised training, as described in Section 3.1 and Appendix B.

ETHICS STATEMENT

This work studies acoustic robustness using existing audio-visual datasets and synthetic interventions. Third-party media remain subject to their source licenses; redistribution should be limited to permitted material, with identifiers and reconstruction instructions provided otherwise. Better robustness does not establish reliability for safety-critical deployment. Applications involving recorded speech and video should account for privacy, consent, and potential differences in performance across speakers and environments.

REPRODUCIBILITY STATEMENT

The objective and data construction are defined in Section 3. Implementation details, data composition, and corruption parameters appear in Appendices A and B. Appendix C provides per-condition results, statistical procedures, and diagnostics; Appendix D specifies prompts, configurations, and data-access limitations. An anonymous source-code package accompanies the paper.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *Proceedings of the International Conference on Learning Representations*, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380.

Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7016–7025, 2021.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026. doi: 10.1162/tacl.a.628.

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

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. AVTrustBench: Assessing and enhancing reliability and robustness in audio-visual LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1590–1601, 2025.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Proceedings of Interspeech*, 2018.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. VITA: Towards open-source interactive omni multimodal LLM. arXiv preprint arXiv:2408.05211, 2024.

Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision*, pp. 103–118, 2018.

Google DeepMind. Gemini 3 Flash model card. Model card, December 17, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *Proceedings of the International Conference on Learning Representations*, 2026.

Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18783–18794, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022.

Jing Hu, Danxiang Zhu, Xianlong Luo, Dan Zhang, Shuwei He, Yishu Lei, Haitao Zheng, Shikun Feng, Jingzhou He, Yu Sun, Hua Wu, and Haifeng Wang. CORD: Bridging the audio-text reasoning gap via weighted on-policy cross-modal distillation. arXiv preprint arXiv:2601.16547, 2026.

Rui Hu, Delai Qiu, Shuyu Wei, Jiaming Zhang, Yining Wang, Shengping Liu, and Jitao Sang. Investigating and enhancing vision-audio capability in omnimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7452–7463, 2025. doi: 10.18653/v1/2025.findings-acl.389.

Woojeong Jin, Maziar Sanjabi, Shaoliang Nie, Liang Tan, Xiang Ren, and Hamed Firooz. MSD: Saliency-aware knowledge distillation for multimodal understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3557–3569, 2021. doi: 10.18653/v1/2021.findings-emnlp.302.

Sung-Bin Kim, Hyun-Bin Oh, Jung-Mok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *Proceedings of the International Conference on Learning Representations*, 2025.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Jiafu Tang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. OmniVideoBench: Towards audio-visual understanding evaluation for omni MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, 2023.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Baichuan-Omni technical report. arXiv preprint arXiv:2410.08565, 2024.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024. doi: 10.18653/v1/2024.emnlp-main.342.

Liang Lin, Chunxi Luo, Kaiwen Luo, Jie Zhang, Jin Wang, Yuanhe Zhang, Cai Yuchen, Qiankun Li, Gongli Xi, Zhenhong Zhou, Kun Wang, and Junhao Dong. EchoDistill: Alignment noisy-to-clean self-distillation for robust audio LLMs. arXiv preprint arXiv:2605.23954, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

m-a-p. OmniInstruct-v1. Dataset, 2024.

Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5213–5224, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Chandan K. A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In *Proceedings of Interspeech*, pp. 2492–2496, 2020. doi: 10.21437/Interspeech.2020-3038.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484, 2015.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993. doi: 10.1016/0167-6393(93)90095-3.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. arXiv preprint arXiv:2509.17765, 2025b.

Han Yin, Yang Xiao, Younghoo Kwon, Ting Dang, and Jung-Woo Choi. Focus then listen: An empirical study of plug-and-play audio enhancer for noise-robust large audio language models. arXiv preprint arXiv:2603.04862, 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023. doi: 10.18653/v1/2023.emnlp-demo.49.

Yibo Zhang, Kaiwen Luo, Liang Lin, Shilinlu Yan, Jin Wang, Yaoqi Guo, Yitian Chen, Yalan Qin, Zhenhong Zhou, Kun Wang, and Li Sun. RSA-bench: Benchmarking audio large models in real-world acoustic scenarios. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18350–18374, 2026. doi: 10.18653/v1/2026.findings-acl.913.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. arXiv preprint arXiv:2601.18734, 2026.

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