OMNIVAL: VALIDATED CHAIN-OF-THOUGHT DISTILLATION FOR TRUE NATIVE OMNI MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

True native omni models promise a single architecture for understanding text, images, audio, and video, yet a critical gap remains: compact open-source omni models still fall short of proprietary frontier systems, and the gap is widest where it matters most - in *grounding* each answer in the right modality and the right evidence. We trace this gap to how such models are trained. Fine-tuning and distillation transfer final outputs or representations, but never the intermediate reasoning that grounding depends on: deciding which modality holds the evidence, localizing it, and expressing it in the proper coordinate space. We introduce **OmniVAL**, a validated chain-of-thought distillation framework that makes these decisions explicit and transfers them from a frontier teacher to a compact open-source student. A teacher generates modality-routed reasoning traces; **MA-VAL**, a modality-aware rule-based validator, filters unreliable traces and turns the student's remaining errors into corrective feedback in the units of each modality, whether pixels, milliseconds, cell indices, or modality-routing fixes. Perception tools serve only as training-time scaffolding: at inference the student is a pure omni model that answers and grounds in a single autoregressive pass. Across 16 benchmarks spanning documents, GUI agency, video, audio-visual reasoning, voice interaction, ASR, and speaker diarization, OmniVAL improves Nemotron 3 Nano Omni, an open 30B model with 3B active parameters, on every benchmark, raising DocVQA grounding mAP from 71.4 to 85.5 and cutting ASR word error rate by 18%; the resulting student surpasses GPT-4o and all open-source omni baselines and substantially narrows the gap to closed-source frontier systems. We release **OmniVAL-CoT**, 110K validator-verified multimodal reasoning traces, and show that half of this validated corpus outperforms the full unfiltered teacher pool. Implementation: `https://anonymous.4open.science/r/OmniVal-216F`

1 INTRODUCTION

Multimodal systems increasingly have to work over documents, charts, graphical user interfaces, video, audio, and speech at once, and in these settings a correct answer is not enough. A reliable model must also *ground* its answer in the evidence that supports it: a bounding box in a document, a cell in a table, a click coordinate on a screen, a temporal span in a video, or a speaker turn in an audio stream. Grounding is what turns a prediction from a black-box response into something that can be inspected, audited, and acted upon, and without it even plausible answers are hard to trust in healthcare, finance, accessibility, or embodied and GUI-based agents.

True native omni models, which process text, images, audio, and video within a single architecture (NVIDIA et al., 2026; Xu et al., 2025b; Luo et al., 2025; OpenAI, 2024; Google DeepMind, 2026a), are the natural substrate for this kind of reasoning. Frontier systems such as Claude Opus 5.5 (Anthropic, 2026a), GPT-6 Sol (OpenAI, 2026), Gemini 3.1 Pro (Google DeepMind, 2026a), and Qwen3.8-Max (Qwen Team, 2026) are strong multimodal reasoners, but they are closed systems or, where weights are released (Kimi K3, Qwen3.8-Max), models of 2.4–2.8 trillion parameters, and among them only Gemini natively accepts audio. Compact open-source omni models such as Nemotron 3 Nano Omni (30B total, 3B active), Qwen3-Omni, and NExT-OMNI cover all four modalities on a single GPU, with full control over weights and no API dependency, which makes them the deployable choice. The problem is that they are weakest on the grounding tasks

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

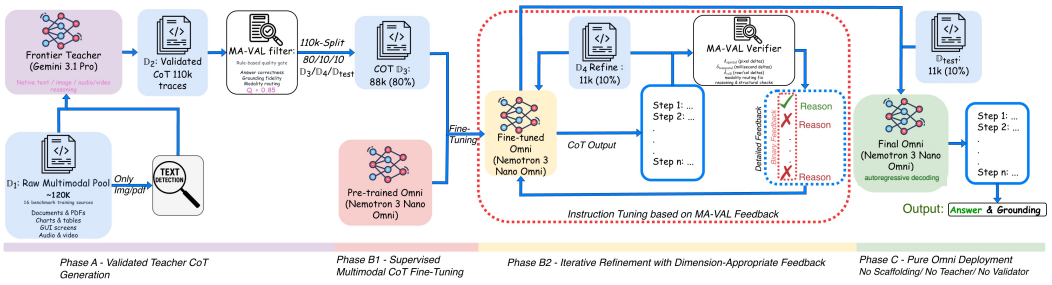

Figure 1: **Overview of the OmniVAL framework.** **Phase A** —Gemini 3.1 Pro generates modality-routed CoT traces over the multimodal pool $\mathcal{D}_1$ with training-time scaffolding; the five-module **MA-VAL** Filter keeps high-quality traces, yielding $\mathcal{D}_2$ (110K). **Phase B1** —Nemotron 3 Nano Omni is fine-tuned on $\mathcal{D}_3$. **Phase B2** —On $\mathcal{D}_4$, the MA-VAL Verifier turns student errors into pixel, millisecond, cell-index, and modality-routing corrections. **Phase C** —The student alone processes raw multimodal inputs in a single autoregressive sequence; *no teacher, validator, or perception tool is used at test time.*

themselves - long-document region localization, GUI click prediction, video moment localization, and speaker-turn attribution.

We argue that this gap is not primarily a matter of scale but of supervision. Standard fine-tuning and knowledge distillation transfer final answers, logits, or intermediate representations (Ahn et al., 2019; Gou et al., 2021); they rarely supervise the reasoning that connects an answer to its evidence, namely deciding *which modality* contains the answer, identifying *where* in that modality the evidence lies, and emitting the grounding in the coordinate space that modality uses. In an omni model this omission is costly, because the model can fail not only by localizing poorly but by reasoning over the wrong modality altogether. We call this failure mode *cross-modal mis-routing*: answering from visible screen text when the evidence is in the audio track, or grounding a document answer in the wrong region while still producing the right string. Standard metrics such as accuracy or ANLS (Yujian & Bo, 2007) reward the string and under-penalize the grounding, so apparent performance and verifiable reliability can drift apart unnoticed.

This paper introduces **OmniVAL**, a validated chain-of-thought distillation framework that teaches compact open-source omni models verifiable, modality-aware grounded reasoning. Rather than distilling final answers, OmniVAL distills *validated reasoning traces* in which the which, where, and how of grounding are written out explicitly. A frontier teacher generates modality-routed chain-of-thought traces over multimodal inputs; **MA-VAL**, a modality-aware rule-based validator, then discards unreliable traces and, once the student is trained, converts its residual grounding errors into precise corrective feedback in the natural units of each modality: pixel deltas for spatial regions, millisecond deltas for temporal spans, cell-index corrections for structured outputs, and explicit fixes for modality-routing errors. Because the student is fine-tuned on this feedback over several rounds, the corrections become durable behavior rather than prompt-time patches (Figure 1). Perception tools (text detectors, ASR, segmenters, keyframe selectors, table parsers) serve only as training-time scaffolding for trace generation, validation, and feedback; the student never sees their outputs, and at inference it is a pure omni model that reads raw inputs and answers and grounds in a single autoregressive pass. With the same 3B-active student and no change in scale, OmniVAL lifts DocVQA grounding mAP from 71.4 to 85.5, recovers 58–98% of the gap to its frontier teacher on the classical document benchmarks, and surpasses GPT-4o on all 16 benchmarks.

Our main contributions are:

- We propose **OmniVAL**, a validated chain-of-thought distillation framework that transfers grounded cross-modal reasoning from frontier omni teachers to compact, deployable open-source omni students.
- We introduce **MA-VAL**, a modality-aware rule-based dual-mode validator that filters unreliable teacher traces and returns corrective feedback in modality-specific units (pixels, milliseconds, cell indices, modality-routing fixes).

- We release **OmniVAL-CoT**, a dataset of 110K validator-verified multimodal CoT traces spanning documents, charts, GUI screens, video, audio, and speech.
- We show on 16 benchmarks across seven task families that validated reasoning supervision improves grounded omni-modal performance, most where grounding is scored: a 3B-active open student with no teacher, validator, or perception tool at inference surpasses GPT-4o and every open-source omni baseline and substantially narrows the gap to closed-source frontier systems. Ablations show that validated supervision outperforms unfiltered teacher data, that verifier feedback improves over binary filtering alone, and that cross-modal routing validation matters most where the evidence modality is in question.

2 Related Work

A *true native omni* model is a single end-to-end network jointly pretrained on text, images, audio, and video that consumes all of them directly at inference, without composing task-specific systems. Open-source examples are **NVIDIA Nemotron 3 Nano Omni** (NVIDIA et al., 2026) (30B / 3B-active hybrid MoE), **Qwen3-Omni** (Xu et al., 2025b) (30B / 3B-active Thinker–Talker MoE), and **NExT-OMNI** (Luo et al., 2025) (Discrete Flow Matching); closed-source ones are **GPT-4o** (OpenAI, 2024), **Gemini 3.1 Pro** (Google DeepMind, 2026a), and **Gemini 3.8 Flash** (Google DeepMind, 2026b). Other frontier models, Claude Opus 5.5 (Anthropic, 2026a), Claude Sonnet 5 (Anthropic, 2026b), GPT-6 Sol and GPT-6 Luna (OpenAI, 2026), Qwen3.8-Max (Qwen Team, 2026), and the open-weight Kimi K3 (max) (Moonshot AI, 2026), are strong reasoners but not omni models: none accepts audio, and only Qwen3.8-Max and Kimi K3 accept video.

Grounding has been studied extensively, but almost always one modality at a time. In documents, layout-aware models such as LayoutLM (Xu et al., 2020; 2022; Huang et al., 2022) jointly model text, image, and layout; OCR-free models (Kim et al., 2022; Lee et al., 2023; Mohammadshirazi et al., 2025) integrate recognition end-to-end; and OCR/layout-token methods such as LayTextLLM (Lu et al., 2025), DocLayLLM (Liao et al., 2025), and LayoutLLM (Luo et al., 2024) encode spatial structure explicitly. GUI grounding and agency have specialist systems such as SeeClick (Cheng et al., 2024), CogAgent (Hong et al., 2024), and OS-Atlas (Wu et al., 2024), and video, audio-visual, voice, ASR, and diarization grounding each have dedicated model families and benchmarks (ours are listed in Section 4). Each line fixes one coordinate space, a pixel box, a click point, a time span, or a speaker turn, and none asks the model to first decide which space the evidence lives in.

Chain-of-thought prompting (Wei et al., 2023) and visual CoT (Xu et al., 2025a; Zheng et al., 2023; Shao et al., 2024a) show that explicit intermediate reasoning helps multimodal performance, but they use traces as supervision without verifying that the cited evidence is in the correct modality or coordinate system. Knowledge distillation (Ahn et al., 2019; Gou et al., 2021; Ji et al., 2021) and VLM distillation (Li et al., 2024b; 2023b) transfer outputs, representations, or intermediate features, and rationale distillation (Gupta et al., 2023) transfers teacher explanations; none detects hallucinated grounding, temporal misalignment, or cross-modal mis-routing. Validation-driven learning (Madaan et al., 2023; Verma et al., 2025) shows the value of feedback and refinement, but the feedback has been discrete or textual rather than expressed in the units of the error. Closest to us, DocVAL (Neogi et al., 2025) distills validated CoT for grounded document VQA with a rule-based validator in filter and verifier modes, but is confined to single-image documents, one coordinate space, and a VLM student.

OmniVAL is the first framework, to our knowledge, to distill *validated* grounded reasoning into a true native omni student, and it differs from each line above. Against modality-specialized grounding, it trains one model to ground in spatial, temporal, structured, and speaker-level coordinate spaces and to choose among them, the decision the cross-modal reasoning validator (Module 4) supervises. Against CoT and distillation methods, it validates every teacher trace before use and expresses the student's residual errors as corrections in native units, pixel deltas, millisecond deltas, cell indices, or modality-routing fixes, rather than as a scalar reward or textual critique. Against DocVAL, it generalizes the validator with temporal, cell-level, and diarization overlap operators and matching correction units, adds routing validation, and evaluates one omni student on 16 benchmarks spanning documents, GUI agency, video, audio-visual reasoning, and speech, with inference free of teacher, validator, and perception tools.

3 METHODOLOGY

Given a multimodal input M and question q, OmniVAL trains a compact true native omni model to generate a reasoning trace c, an answer a, and a grounding output g whose form depends on the evidence modality: bounding boxes for spatial inputs, time intervals for audio/video, cell indices for tables, and speaker turns for diarization. Its design follows three principles: (i) supervision must expose the intermediate decisions behind grounding (which modality, where, and how the answer is tied to the evidence); (ii) teacher reasoning must be validated before it is used for training, since hallucinated regions, wrong timestamps, or modality-routing errors would otherwise be distilled into the student; and (iii) the student must not inherit the teacher's inference cost or any perception tool as a test-time dependency, so auxiliary tools are strictly training-time scaffolding.

3.1 OVERVIEW

Figure 1 shows the pipeline: three training phases and a deployment phase. In **Phase A**, a frontier teacher writes modality-routed chain-of-thought traces for tuples (M, q, a_{gt}, g_{gt}), where a_{gt} and g_{gt} are the ground-truth answer and grounding, and **MA-VAL**, our modality-aware validator, filters them into a verified corpus. In **Phase B1**, the student is supervised on that corpus and learns to produce (c, a, g) directly from (M, q). In **Phase B2**, the same validator runs in verifier mode, diagnosing the student's errors and turning them into corrective instruction-tuning examples. In **Phase C**, the teacher, validator, and every auxiliary tool are removed and only the fine-tuned student remains.

The design is deliberately asymmetric. During training, MA-VAL may consult OCR, ASR, segmentation masks, table structures, keyframes, and diarization cues to verify teacher and student outputs; the student never receives any of these signals as input. It learns from validated supervision, but it must ground its answers from raw multimodal inputs alone.

3.2 MA-VAL: MODALITY-AWARE VALIDATION

MA-VAL is a rule-based validator that performs quality control on grounded reasoning supervision. It runs the same checks in two modes that differ only in what they output (Figure 2): in *Filter mode* it accepts or rejects teacher traces during corpus construction, and in *Verifier mode* it analyzes student predictions and emits structured feedback consisting of error descriptions, corrected targets, and modality-specific correction vectors.

Why rule-based. A learned validator can hallucinate critiques, give unstable feedback, or misidentify the evidence modality, and in the omni setting that last error is particularly damaging, since a validator that routes wrongly will systematically teach the student to ground in the wrong modality. MA-VAL therefore relies on deterministic checks over explicit quantities (ANLS, spatial IoU, temporal IoU, cell F1, and DER), which keeps its feedback auditable and lets corrections be stated in the natural units of each modality: pixels, milliseconds, cells, or speaker turns. Full scoring formulas are given in Appendix A.3.

Validation modules. MA-VAL has five modules. **M1: Grounding Engine** dispatches to the overlap measure appropriate to the modality: spatial IoU for documents, charts, GUI screens, and images; temporal IoU for audio and video spans; cell F1 for tables; and DER for diarization. **M2: Answer Validator** scores textual correctness as $Q_{\text{ans}} = 0.7 \cdot \text{ANLS}(a_p, a_{gt}) + 0.3 \cdot \mathbf{1}[a_p \in \mathcal{T}(M)]$, where $\mathcal{T}(M)$ is the modality-specific textual evidence, such as OCR text for documents or an ASR transcript for audio. **M3: Grounding Validator** checks localization accuracy and emits correction vectors, e.g., $\delta_{\text{spatial}} = [\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2]^\top$, $\delta_{\text{temporal}} = [\Delta t_1, \Delta t_2]^\top$, or $\delta_{\text{cell}} = [\Delta r_1, \Delta c_1, \Delta r_2, \Delta c_2]^\top$. **M4: Cross-Modal Reasoning Validator** targets cross-modal mis-routing, the failure mode specific to omni models, by checking that the predicted grounding type matches the ground-truth modality, that the reasoning trace refers to the same modality as the emitted grounding, and that the prediction agrees with scaffolding evidence from the correct modality. **M5: Feedback Generator** combines the module scores into an overall quality score Q; in Filter mode it applies the threshold $Q > 0.85$, and in Verifier mode it produces a corrected target together with a priority-ordered list of fixes.

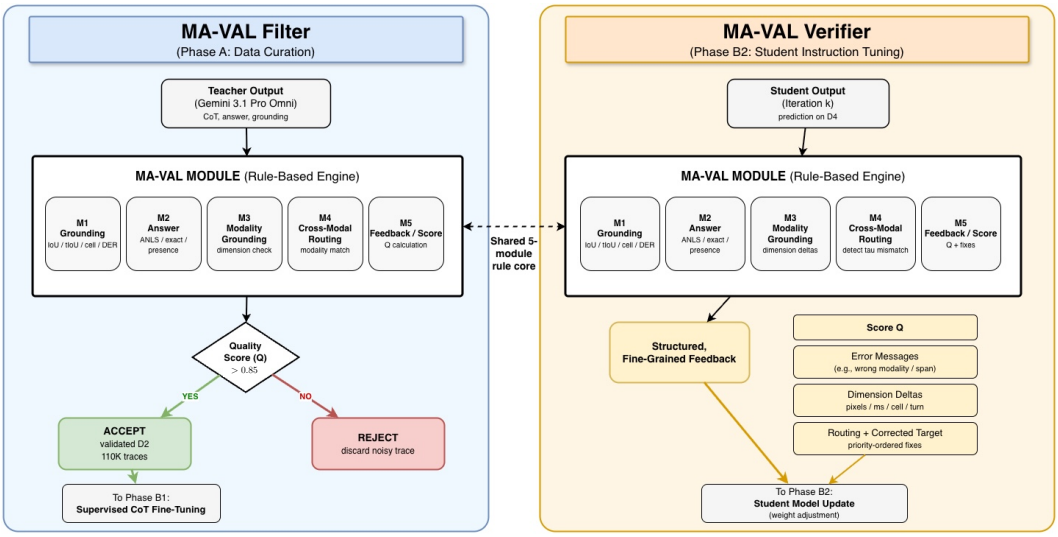

Figure 2: **MA-VAL dual-mode architecture.** MA-VAL uses the same five-module rule-based validator in two modes. In **Filter mode** (Phase A), it scores teacher-generated traces and keeps only high-quality examples for $\mathcal{D}_2$. In **Verifier mode** (Phase B2), it diagnoses student predictions and produces corrective feedback, including pixel deltas for spatial grounding, millisecond deltas for temporal grounding, cell-index corrections for structured outputs, and explicit fixes for cross-modal mis-routing.

3.3 Phase A: Validated Teacher Trace Generation

We assemble an initial pool $\mathcal{D}_1$ of approximately 120K multimodal tuples (M, q, a_{gt}, g_{gt}) spanning documents, charts, GUI screens, audio, speech, and video, drawn exclusively from the training splits of the source datasets; all evaluation examples are disjoint from $\mathcal{D}_1$ (Appendix A.1). The default teacher, Gemini 3.1 Pro (chosen for its native audio input; Section 4 and Appendix C.6 give the rationale and alternatives), is prompted to write a modality-routed chain-of-thought trace that first names the evidence modality and then explains the answer with modality-specific grounding. DB-ResNet (Liao et al., 2019), SAM 2.1 (Ravi et al., 2024), Whisper v3 (Radford et al., 2022), SigLIP (Zhai et al., 2023), pyannote (Bredin et al., 2019), and TableFormer (Nassar et al., 2022) provide training-time scaffolding whose outputs go to the teacher as auxiliary context and to MA-VAL for validation, but never to the student.

MA-VAL then scores every teacher trace in Filter mode and keeps those with $Q > 0.85$, which yields $\mathcal{D}_2$, a corpus of **110K validator-verified multimodal CoT traces** with answer and grounding supervision. We split $\mathcal{D}_2$ by modality into $\mathcal{D}_3$ for supervised training (80%, 88K), $\mathcal{D}_4$ for refinement (10%, 11K), and $\mathcal{D}_{\text{test}}$ for held-out analysis (10%, 11K). Appendix B shows example traces, and Table 9 ablates the scaffolding choices.

3.4 Phase B: Validation-Driven Student Learning

Stage B1: supervised grounded reasoning. The student is Nemotron 3 Nano Omni, a true native omni model with 30B total and 3B active parameters per token in a hybrid MoE architecture. We fully fine-tune it on $\mathcal{D}_3$: given only (M, q), it learns to generate the serialized sequence $c = (\text{CoT}, a, g)$, with g the modality-appropriate grounding target, under the standard autoregressive objective $\mathcal{L}_{\text{SFT}} = -\sum_t \log P_\theta(c_t \mid M, q, c_{<t})$. This stage teaches the output format and the reasoning pattern for modality routing, answer prediction, and grounding, without exposing the student to any scaffolding signal.

Stage B2: validator-guided iterative refinement. Imitating validated traces leaves fine-grained grounding errors, and Stage B2 converts them into corrective supervision with MA-VAL in Verifier mode. At iteration k the current student predicts $(c_i^{(k)}, a_i^{(k)}, g_i^{(k)}) = M_{\theta_{k-1}}(M_i, q_i)$ for each example

in $\mathcal{D}_4$; MA-VAL compares the prediction against the ground truth and the scaffolding evidence and produces feedback F_i describing the error and the required correction; and the student is then instruction-tuned on the feedback-augmented examples $(M_i, q_i, F_i, c_i^{gt}, a_i^{gt}, g_i^{gt})$ with the same token-level cross-entropy objective as Stage B1; no scalar reward, policy-gradient step, or inference-time validator is involved (a supervised-versus-GRPO comparison is given in Appendix D.5). Because the weights are updated after every round, the feedback produces durable changes in the model rather than temporary prompt-time behavior. We run up to $\eta_{\max} = 20$ iterations and stop when the mean per-iteration gain $\bar{\Delta}$ in composite validation score on $\mathcal{D}_4$ falls below 0.2 points over a three-iteration window (Appendix A.3); in practice, per-benchmark scores plateau after 11–15 iterations (mean 13.1 across the eight metric columns; Table 7).

3.5 Phase C: Pure Omni Deployment

After training, the teacher, MA-VAL, and all scaffolding tools are discarded. The deployed student receives only (M, q) and generates (c, a, g) in a single autoregressive sequence, so no OCR, ASR, segmenter, validator, or teacher API is needed at inference. For video, query-conditioned keyframe selection is handled by the student's own attention rather than by an external SigLIP module; Table 9 ablates the effect of SigLIP scaffolding during training.

4 Experiments

We evaluate OmniVAL on 16 benchmarks covering document understanding, GUI agency, video and audio-visual reasoning, voice interaction, ASR, and speaker diarization: DocVQA (Mathew et al., 2021), VisualMRC (Tanaka et al., 2021), FUNSD (Jaume et al., 2019), CORD (Park et al., 2019), SROIE (Huang et al., 2019), OCRBenchV2-En (Fu et al., 2025b), MMLongBench-Doc (Ma et al., 2024), and CharXiv reasoning (Wang et al., 2024) for documents; ScreenSpot-Pro (Li et al., 2025) and OSWorld (Xie et al., 2024) for GUI agency; Video-MME (Fu et al., 2025a), WorldSense (Hong et al., 2026), and DailyOmni (Zhou et al., 2026) for video and audio-visual reasoning; and VoiceBench (Chen et al., 2024), the HF Open ASR Leaderboard (Srivastav et al., 2026), and DIHARD III (Ryant et al., 2021) for speech. Each is scored with its official metric: ANLS and mAP@[0.5:0.95] for document answer grounding, the official task score for GUI, video, and audio-visual benchmarks, WER for ASR, and DER for diarization.

Model selection. Our student is Nemotron 3 Nano Omni (NVIDIA et al., 2026) and our default teacher is Gemini 3.1 Pro (Google DeepMind, 2026a); both choices are dictated by the method. Because OmniVAL fully fine-tunes the student and supervises grounding in all four modalities, the student must be an open-weight true native omni model; of the three that exist (Nemotron 3 Nano Omni, Qwen3-Omni, NExT-OMNI), Nemotron 3 Nano Omni is the strongest before any fine-tuning on every one of the 16 benchmarks (Tables 1–2) and, with a 3B-active hybrid MoE, the most deployable. Because the teacher must write reasoning traces for every example in $\mathcal{D}_1$, including its speech-only portion, it must natively accept audio as well as text, image, and video. Among current frontier models only the two Gemini models do, and Gemini 3.1 Pro is the stronger under our protocol; the earlier omni model GPT-4o also accepts audio but trails Gemini 3.1 Pro on 15 of the 16 benchmarks (all but SROIE ANLS) and yields a weaker student. We therefore use Gemini 3.1 Pro († in the tables) and not Claude Opus 5.5, although Opus is the strongest directly evaluated model in Tables 1 and 2: Opus cannot write traces for the speech-only examples, and we deliberately use one teacher for the whole corpus rather than assigning teachers per modality, since that assignment would have to decide each example's evidence modality before any trace is written, which is precisely the routing decision the traces exist to make explicit and MA-VAL's Module 4 exists to validate. The teacher ablation (Appendix C.6, Table 10) supports this choice and quantifies, as a diagnostic, how much teacher strength matters on the non-speech data.

Baselines and protocol. We compare against document-specialist VLMs, the open-source true native omni models NExT-OMNI, Qwen3-Omni, and base Nemotron 3 Nano Omni, and eight current frontier models, Claude Opus 5.5, GPT-6 Sol, Qwen3.8-Max, Kimi K3, Claude Sonnet 5, GPT-6 Luna, Gemini 3.1 Pro, and Gemini 3.8 Flash (roster and input coverage in Appendix E). Following Section 2, the tables group these by input coverage: the six frontier models without native audio input, and the closed-source true native omni models Gemini 3.1 Pro, Gemini 3.8 Flash, and the earlier

Table 1: Document text grounding. OmniVAL operates as a pure omni model (30B total, 3B active parameters) with no OCR or text detection at inference. All frontier-model scores are obtained under our evaluation protocol; parameter counts of all compared models are listed in Table 25. —: metric not produced (no mAP for models that do not output boxes). (API): scores obtained through the provider API; the released checkpoint is text-only. †: default OmniVAL teacher; the only frontier model besides Gemini 3.8 Flash with native audio input, which the teacher role requires (Section 4). **Bold**: best open-source omni; underline: best overall.

Method	DocVQA		VisualMRC		FUNSD		CORD		SROIE	
	ANLS	mAP	ANLS	mAP	ANLS	mAP	ANLS	mAP	ANLS	mAP
Document-specialist VLMs										
DocLayLLM (Llama3-8B)	78.4	—	55.0	—	84.1	—	71.3	—	84.3	—
LayoutLLM (Vicuna1.5-7B)	74.3	—	55.8	—	80.0	—	63.1	—	72.1	—
LayTextLLM (Llama2-7B)	77.2	—	41.7	—	81.0	—	82.5	—	96.1	—
DLaVA (Pixtral-12B)	85.9	46.2	52.1	—	87.6	45.5	84.4	57.9	91.4	—
Frontier models w/o native audio input										
Claude Opus 5.5	97.2	91.4	84.6	80.8	97.7	90.2	96.1	89.4	99.0	90.6
GPT-6 Sol	96.7	90.7	83.8	79.9	97.2	89.5	95.3	88.5	98.5	89.9
Qwen3.8-Max (API)	96.1	90.0	83.0	78.9	96.6	88.8	94.5	87.5	97.9	89.2
Kimi K3 (max)	95.5	89.4	82.2	78.0	96.0	88.2	93.7	86.6	97.3	88.6
Claude Sonnet 5	94.5	88.0	80.5	76.1	95.0	86.8	92.0	84.7	96.2	87.2
GPT-6 Luna	93.9	87.3	79.7	75.2	94.4	86.1	91.2	83.8	95.7	86.5
Closed-source true native omni										
Gemini 3.1 Pro†	93.4	86.6	78.9	74.2	93.9	85.4	90.4	82.8	95.2	85.8
Gemini 3.8 Flash	92.8	86.0	78.1	73.3	93.3	84.8	89.6	81.9	94.6	85.2
GPT-4o	92.5	83.1	74.6	68.9	92.5	82.4	89.2	78.7	95.5	80.4
Open-source true native omni										
NExT-OMNI	84.1	—	65.2	—	84.7	—	81.5	—	90.4	—
Qwen3-Omni	87.6	—	70.3	—	88.4	—	84.9	—	92.6	—
Nemotron 3 Nano Omni (base)	88.9	71.4	71.8	60.3	89.3	73.5	86.1	71.2	93.7	72.8
OmniVAL (Nemotron 3 Nano Omni)	**93.2**	**85.5**	**75.9**	**70.4**	**93.8**	**83.9**	**90.1**	**80.2**	**96.0**	**81.6**

GPT-4o (OpenAI, 2024). We evaluate every frontier model ourselves through its API under the same protocol, prompts, and test splits as the student, at the provider's default reasoning effort (Table 25); Qwen3.8-Max is marked (API) because its released checkpoint is text-only. Input coverage decides what can be scored: only the omni models are scored on the five audio-dependent benchmarks (n/a elsewhere), and the text+image models (Claude Opus 5.5, Claude Sonnet 5, GPT-6 Sol, GPT-6 Luna) are evaluated on Video-MME from 32 uniformly sampled frames, the benchmark's own protocol for image-only LMMs (marked §; Appendix A.1). Every OmniVAL number comes from the final student alone, with no teacher, validator, or perception tool at inference; hyperparameters and per-benchmark protocols are in Appendix A.

5 Results and Analysis

5.1 Main Results

Gains over the base student. OmniVAL improves Nemotron 3 Nano Omni on all 16 benchmarks, by $+2.3$ to $+15.9$ points on higher-is-better metrics and by 1.09 WER and 2.9 DER, and the gains are largest where localization is scored. On the classical document suite (Table 1), DocVQA rises from 88.9 to **93.2** ANLS but from 71.4 to **85.5** mAP; that the mAP gain $(+14.1)$ is more than three times the ANLS gain $(+4.3)$ is the first sign that validated CoT supervision acts mainly on localization, and the pattern repeats on VisualMRC, FUNSD, CORD, and SROIE, where mAP gains of $+8.8$ to $+10.4$ accompany ANLS gains of $+2.3$ to $+4.5$. The harder document benchmarks (Table 2) move further, OCRBenchV2-En from 65.8 to **77.5**, MMLongBench-Doc from 57.5 to **71.3**, and CharXiv reasoning from 63.6 to **77.4** ($+11.7$, $+13.8$, $+13.8$), so the transfer holds beyond forms and receipts into long-document and chart reasoning. Tasks that ground in click coordinates, video moments, and audio-visual alignment improve under the same recipe: ScreenSpot-Pro by $+15.6$, OSWorld by $+15.9$, Video-MME by $+9.1$, WorldSense by $+13.6$, and DailyOmni by $+11.9$. On speech (Table 2), VoiceBench rises from 89.4 to **91.8**, HF Open ASR WER falls from 5.95 to **4.86** and DIHARD III DER from 21.4 to **18.5** (relative reductions of 18.3% and 13.6%), so temporal

Table 2: Newer document benchmarks (OCRB-v2: OCRBenchV2-En; MMLB-Doc: MMLongBench-Doc; CharXiv reasoning), GUI agency (SS-Pro: ScreenSpot-Pro), video understanding (V-MME: Video-MME), audio-visual reasoning (WS: WorldSense; DO: DailyOmni), and speech (VB: VoiceBench; WER: HF Open ASR; DER: DIHARD III). All scores use our evaluation protocol; higher is better except WER and DER (↓). n/a: the model does not accept audio input. §: text+image model evaluated on uniformly sampled video frames (Video-MME without subtitles; no audio track), as in the official Video-MME protocol for image LMMs. (API): scores obtained through the provider API; the released checkpoint is text-only. †: default OmniVAL teacher (see Table 1). **Bold**: best open-source omni; underline: best overall.

Method	Docs & charts			GUI		Video	Video + Audio		Speech		
	OCRB-v2	MMLB-Doc	CharXiv	SS-Pro	OSWorld	V-MME	WS	DO	VB	WER ↓	DER ↓
Frontier models w/o native audio input											
Claude Opus 5.5	90.8	85.2	92.6	85.7	78.4	92.1§	n/a	n/a	n/a	n/a	n/a
GPT-6 Sol	90.2	84.5	91.8	84.8	77.9	91.4§	n/a	n/a	n/a	n/a	n/a
Qwen3.8-Max (API)	89.7	83.8	91.0	83.8	77.3	90.7	n/a	n/a	n/a	n/a	n/a
Kimi K3 (max)	89.1	83.2	90.2	82.9	76.8	90.1	n/a	n/a	n/a	n/a	n/a
Claude Sonnet 5	88.0	81.8	88.5	81.0	75.7	88.7§	n/a	n/a	n/a	n/a	n/a
GPT-6 Luna	87.5	81.1	87.7	80.1	75.1	88.0§	n/a	n/a	n/a	n/a	n/a
Closed-source true native omni											
Gemini 3.1 Pro†	87.0	80.4	86.9	79.1	74.6	87.3	80.6	88.2	94.4	4.38	15.9
Gemini 3.8 Flash	86.4	79.8	86.1	78.2	74.0	86.7	79.8	87.3	93.8	4.62	16.6
GPT-4o	76.5	68.4	73.5	65.7	56.2	79.4	65.8	80.7	91.2	5.04	18.9
Open-source true native omni											
NExT-OMNI	60.2	51.4	58.7	53.1	42.6	68.4	50.8	70.2	84.5	7.32	24.7
Qwen3-Omni	64.1	55.3	61.9	56.4	45.9	71.0	53.7	72.5	87.8	6.41	22.9
Nemotron 3 Nano Omni (base)	65.8	57.5	63.6	57.8	47.4	72.2	55.4	74.1	89.4	5.95	21.4
OmniVAL (Nemotron 3 Nano Omni)	**77.5**	**71.3**	**77.4**	**73.4**	**63.3**	**81.3**	**69.0**	**86.0**	**91.8**	**4.86**	**18.5**

grounding and speaker-turn localization improve alongside visual and document grounding. One validation-driven recipe thus improves every modality rather than overfitting to document grounding.

Position against each model group. Against the open-source true native omni models, OmniVAL beats Qwen3-Omni, NExT-OMNI, and the base model on all 16 benchmarks, and it beats the document-specialist VLMs on every column but SROIE ANLS, where LayTextLLM leads by 0.1. Against the closed-source omni group, it surpasses GPT-4o on all 16 (by 0.5–7.7 points, 0.18 WER, and 0.4 DER) and reaches the answer accuracy of Gemini 3.1 Pro on the classical documents: within 0.3 ANLS on DocVQA, FUNSD, and CORD, ahead on SROIE (96.0 vs. 95.2), and behind by 3.0 only on VisualMRC, with the remaining gap concentrated in localization (1.1–4.2 mAP). Elsewhere the gap to Gemini 3.1 Pro is 2.2–11.6 points on the newer document, GUI, video, and audio-visual benchmarks (smallest on DailyOmni, largest on WorldSense), 2.6 on VoiceBench, 0.48 WER, and 2.6 DER; measured against the base student, OmniVAL recovers 58–98% of that gap on the classical document metrics where a gap remains (it is already ahead on SROIE ANLS), 54–84% on the newer document, GUI, video, and audio-visual benchmarks, and 48–69% on speech, with 3B active parameters. Against the frontier models without audio input, Claude Opus 5.5, the strongest of them in every column it can be scored on, stays ahead by 3.0–8.7 ANLS and 5.9–10.4 mAP on the classical documents and by 10.8–15.2 points on the Table 2 benchmarks it can be evaluated on; but none of the six can be evaluated on the five audio-dependent benchmarks at all, whereas OmniVAL covers all 16 with one omni student. OmniVAL therefore narrows rather than closes the frontier gap.

5.2 Ablations and Key Findings

We ablate OmniVAL along five axes, validation strategy, cross-modal routing validation, iterative refinement, data scale, and teacher choice (full tables and per-modality breakdowns in Appendix C). **Validation is essential** (Table 5): the unfiltered 120K teacher pool gives the weakest student, MA-VAL Filter improves on it by removing low-quality traces, and Verifier-mode refinement adds a further $+1.8$ to $+12.0$ points over binary filtering while lowering WER and DER, so the supervision benefits from corrective feedback, not only from discarding bad traces. **Module 4 targets cross-modal mis-routing** (Tables 6 and 11): removing it costs only 0.3 ANLS and 1.4 mAP on DocVQA, where the evidence modality is never in question, but 8.6 and 9.0 points on WorldSense and DailyOmni, where the answer may lie in either stream, 9.7 on ScreenSpot-Pro, where a textual instruction must be routed to a screen element, and 4.8 on Video-MME; Table 11 shows why, as routing accuracy

on audio-only questions falls from 91.7 to 76.3 without it because the student defaults to the visual stream. **Iterative refinement** (Table 7) improves every metric over Stage B1 alone, least on DocVQA ANLS (+1.8) and most on ScreenSpot-Pro (+12.0) and DocVQA mAP (+6.6), so verifier feedback matters most where precise localization is scored. **Quality outweighs scale** (Table 8): 50% of the validated corpus already beats the full unfiltered 120K pool on every shared metric.

Teacher choice (Table 10). Each row is the score of the *omni student* taught by the named teacher, not the teacher's own score (those are the frontier rows of Tables 1–2); because the student natively processes audio and video, its WorldSense and DailyOmni scores say nothing about whether its teacher accepts audio. Teachers are grouped by input coverage, since only a teacher with native audio input can write traces for the speech-only portion of $\mathcal{D}_1$ and is therefore eligible for the full corpus. Among these omni teachers, Gemini 3.1 Pro (†) gives the best student on every column, ahead of Gemini 3.8 Flash, Gemini 2.5 Flash (Comanici et al., 2025), and GPT-4o by 0.7, 5.9, and 6.6 DocVQA mAP. On the nine non-speech benchmarks, student quality tracks the teacher's directly evaluated strength: the Claude Opus 5.5-taught student reaches 90.3 DocVQA mAP against 85.5 for the default, the older teachers Claude Sonnet 4.5 (Anthropic, 2025), GPT-5 (OpenAI, 2025), Gemini 2.5 Flash, and GPT-4o yield 78.9–83.5, and removing teacher CoT altogether costs 14.1 mAP on DocVQA and 7.9–14.2 points elsewhere; the open-weight teachers Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025) and Llama 4-400B-A17B (Meta AI, 2025) stay within 2.9–3.6 mAP of the default, so OmniVAL does not depend on proprietary teacher APIs. An Opus-taught student cannot, however, be trained on the full OmniVAL corpus, because Claude Opus 5.5 cannot write traces for the speech-only examples: the 90.3 row is an upper bound on the non-speech portion, not an available configuration, which is why the speech benchmarks are excluded from Table 10 and why its second group does not show that a teacher without audio input would be better for the full corpus. Nor is splitting the corpus between an omni teacher and a stronger text+image teacher a route to those 4.8 points: the split must decide, before any trace is written, which modality each example's evidence lives in, trivial for pure documents and pure speech but the very routing decision that the audio-visual, video, and GUI examples are in the corpus to teach, and it would train the student on two reasoning distributions that differ systematically by modality, confounding the single-recipe cross-modal transfer of Tables 1–2. The headroom therefore measures how much OmniVAL stands to gain from stronger omni teachers, and the second group of Table 10 shows that gain is real (Appendix C.6). Together the ablations support the core claim: grounded omni reasoning improves when routing, localization, and correction are explicitly validated rather than merely distilled. Appendix D probes the claim from further directions: transfer to six general image, video, and audio benchmarks and two cross-modal hallucination benchmarks; a factorial separation of CoT, grounding, validation, and refinement; MA-VAL under noisy or swapped scaffolding tools; the cost of scaffolded teacher generation and of selective refinement; inference-time reasoning cost; causal modality interventions; and transfer of the pipeline to Qwen3-Omni.

6 Conclusion

The gap between compact open-source omni models and frontier systems on grounded tasks is usually read as a gap in scale, and that gap is real: 3B active parameters against closed or trillion-parameter systems. Much of it, we have shown, is a gap in supervision: standard fine-tuning and distillation never teach which modality holds the evidence, where it lies, or in which coordinate system to express it. **OmniVAL** does: a frontier teacher writes modality-routed reasoning traces, **MA-VAL** validates every trace before it is learned from and turns residual student errors into pixel, millisecond, cell-index, and modality-routing corrections, and the student stays a pure omni model with no teacher, validator, or perception tool at inference. On 16 benchmarks across seven task families it improves Nemotron 3 Nano Omni on every one, surpasses every open-source omni baseline and GPT-4o, and substantially narrows, without closing, the frontier gap, most where localization is scored: DocVQA grounding mAP rises from 71.4 to 85.5, ScreenSpot-Pro by 15.6 points, and ASR word error rate falls by 18.3%. Measured against the base student, it recovers 58–98% of the gap to its own frontier teacher on the classical document benchmarks, and the same recipe transfers to a second student and to general image, video, and audio benchmarks that no scaffolding tool touches (Appendix D). Two lessons carry beyond this paper: half of a validated corpus beats the full unfiltered teacher pool, so trace quality outweighs volume; and validating cross-modal routing lifts audio-only routing accuracy from 76.3 to 91.7, so modality choice must be supervised, not assumed. We release **OmniVAL-CoT**, 110K validator-verified traces; limitations are discussed in Appendix F.

AI USE STATEMENT

Frontier generative AI models are part of the methodology of this work: they are used as teacher models to generate chain-of-thought traces for OmniVAL-CoT (Section 3.3) and are also evaluated as comparison baselines (Section 4). These uses are integral to the experiments and are described explicitly in the paper. Separately, we used generative AI tools in a limited capacity to assist with LaTeX formatting, and organization of some tables. All AI-assisted material was carefully reviewed and verified by at least two authors before inclusion in the manuscript. Generative AI tools were not used to conduct experiments, get experimental measurements or results, or independently draw scientific conclusions from the experiments. They were also not used to generate the substantive scientific content of the manuscript. The authors independently verified the experimental results, numerical claims, methodological descriptions, citations, and conclusions presented in this work. We take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work does not involve human-subject experiments or newly collected personal data. OmniVAL-CoT is derived from established benchmark datasets and model-generated traces, and each source dataset is used in accordance with its applicable license, terms of use, or access agreement (e.g., DIHARD III via LDC); see Appendix I. Grounding over documents, audio, video, and speaker turns could be misused for surveillance or privacy-invasive analysis, while plausible but incorrect evidence localization may create false confidence in model outputs. We discuss these risks and their broader implications in Appendix G, and describe our release safeguards, intended-use guidance, and restrictions on privacy-invasive and sensitive applications in Appendix H.

REPRODUCIBILITY STATEMENT

Appendix A provides the benchmark evaluation protocols, including train/evaluation split disjointness (Appendix A.1), the complete hyperparameter and compute configuration (Table 4), the full MA-VAL scoring formulas and convergence criteria (Appendix A.3), and pseudocode for the complete OmniVAL training and refinement pipeline (Algorithm 1). Appendix B provides representative teacher, student, and verifier examples, and Appendix J documents the composition, construction, splits, and intended use of the OmniVAL-CoT release. Code, preprocessing scripts, and reproduction instructions are available in the anonymous repository at `https://anonymous.4open.science/r/OmniVal-216F`.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer, 2019. URL `https://arxiv.org/abs/1904.05835`.

Anthropic. Introducing Claude Sonnet 4.5, September 2025. URL `https://www.anthropic.com/news/claude-sonnet-4-5`.

Anthropic. System card: Claude opus 5.5, September 2026a. URL `https://www.anthropic.com/claude-opus-5-5-system-card`.

Anthropic. Introducing Claude Sonnet 5, June 2026b. URL `https://www.anthropic.com/news/claude-sonnet-5`.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL `https://arxiv.org/abs/2006.11477`.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren,

Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization, 2019. URL https://arxiv.org/abs/1911.01255.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants, 2024. URL https://arxiv.org/abs/2410.17196.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024. URL https://arxiv.org/abs/2401.10935.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. URL https://arxiv.org/abs/2507.05595.

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, 2025a. URL https://arxiv.org/abs/2405.21075.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2025b. URL https://arxiv.org/abs/2501.00321.

Google DeepMind. Gemini 3.1 Pro model card, February 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.8 Flash model card, September 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, March 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01453-z. URL http://dx.doi.org/10.1007/s11263-021-01453-z.

Avani Gupta, Saurabh Saini, and P J Narayanan. Concept distillation: Leveraging human-centered explanations for model improvement, 2023. URL https://arxiv.org/abs/2311.15303.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2026. URL https://arxiv.org/abs/2502.04326.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents, 2024. URL https://arxiv.org/abs/2312.08914.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025. URL https://arxiv.org/abs/2501.13826.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022. URL https://arxiv.org/abs/2204.08387.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019. doi: 10.1109/icdar.2019.00244. URL http://dx.doi.org/10.1109/ICDAR.2019.00244.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents, 2019. URL https://arxiv.org/abs/1905.13538.

Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7945–7952, 2021.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, pp. 498–517. Springer, 2022.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks, 2020. URL https://arxiv.org/abs/2012.14952.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18893–18912. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lee23g.html.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio, 2024. URL https://arxiv.org/abs/2410.12787.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use, 2025. URL https://arxiv.org/abs/2504.07981.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024a. URL https://arxiv.org/abs/2311.17005.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102, 2023a.

Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability, 2023b. URL https://arxiv.org/abs/2307.03135.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models, 2024b. URL https://arxiv.org/abs/2403.02781.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization, 2019. URL https://arxiv.org/abs/1911.08947.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding, 2025. URL https://arxiv.org/abs/2408.15045.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, Ernie Chang, Saksham Suri, Hongyu Xu, Qi Qian, Wei Wen, Balakrishnan Varadarajan, Zhuang Liu, Hu Xu, Florian Bordes, Raghuraman Krishnamoorthi, Bernard Ghanem, Vikas Chandra, and Yunyang Xiong. Videoauto-r1: Video auto reasoning via thinking once, answering twice, 2026. URL https://arxiv.org/abs/2601.05175.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, Hao Liu, and Can Huang. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding, 2025. URL https://arxiv.org/abs/2407.01976.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding, 2024. URL https://arxiv.org/abs/2404.05225.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching, 2025. URL https://arxiv.org/abs/2510.13721.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations, 2024. URL https://arxiv.org/abs/2407.01523.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix, 2025. URL https://arxiv.org/abs/2505.13032.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021. URL https://arxiv.org/abs/2007.00398.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Ahmad Mohammadshirazi, Pinaki Prasad Guha Neogi, Ser-Nam Lim, and Rajiv Ramnath. Dlava: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness, 2025. URL https://arxiv.org/abs/2412.00151.

Moonshot AI. Kimi K3: Open frontier intelligence, July 2026. URL https://github.com/MoonshotAI/Kimi-K3.

Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. Tableformer: Table structure understanding with transformers, 2022. URL https://arxiv.org/abs/2203.01017.

Pinaki Prasad Guha Neogi, Ahmad Mohammadshirazi, Ser-Nam Lim, and Rajiv Ramnath. DocVAL: Validated chain-of-thought distillation for grounded document VQA, 2025. URL https://arxiv.org/abs/2511.22521.

NVIDIA, :, Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, Mike Ranzinger, Greg Heinrich, Guo Chen, Lukas Voegtle, Philipp Fischer, Timo Roman, Karan Sapra, Collin McCarthy, Shaokun Zhang, Fuxiao Liu, Hanrong Ye, Yi Dong, Mingjie Liu, Yifan Peng, Piotr Zelasko, Zhehuai Chen, Nithin Rao Koluguri, Nune Tadevosyan, Lilit Grigoryan, Ehsan Hosseini Asl, Pritam Biswas, Leili Tavabi, Yuanhang Su, Zhiding Yu, Peter Jin, Alexandre Milesi, Netanel Haber, Yao Xu, Sarah Amiraslani, Nabin Mulepati, Eric Tramel, Jaehun Jung, Ximing Lu, Brandon Cui, Jin Xu, Zhiqi Li, Shihao Wang, Yuanguo Kuang, Shaokun Zhang, Huck Yang, Boyi Li, Hongxu Yin, Song Han, Bilal Kartal, Pavlo Molchanov, Adi Renduchintala, Charles Wang, David Mosallanezhad, Soumye Singhal, Luis Vega, Katherine Cheung, Sreyan Ghosh, Yian Zhang, Alexander Bukharin, Venkat Srinivasan, Johnny Greco, Andre Manoel, Maarten Van Segbroeck, Suseella Panguliri, Rohit Watve, Divyanshu Kakwani, Shubham Pachori, Jeffrey Glick, Radha Sri-Tharan, Aileen Zaman, Khanh Nguyen, Shi Chen, Jiaheng Fang, Qing Miao, Wenfei Zhou, Yu Wang, Zaid Pervaiz Bhat, Varun Praveen, Arihant Jain, Ramanathan Arunachalam, Tomasz Kornuta, Ashton Sharabiani, Amy Shen, Wei Huang, Yi-Fu Wu, Ali Roshan Ghias, Huiying Li, Brian Yu, Nima Tajbakhsh, Chen Cui, Wenwen Gao, Li Ding, Terry Kong, Manoj Kilaru, Anahita Bhiwandiwalla, Marek Wawrzos, Daniel Korzekwa, Pablo Ribalta, Grzegorz Chlebus, Besmira Nushi, Ewa Dobrowolska, Maciej Jakub Mikulski, Kunal Dhawan, Steve Huang, Jagadeesh Balam, Yongqiang Wang, Nikolay Karpov, Valentin Mendelev, George Zelenfroynd, Meline Mkrtchyan, Qing Miao, Omri Almog, Bhavesh Pawar, Rameshwar Shivbhakta, Sudeep Sabnis, Ashrton Sharabiani, Negar Habibi, Geethapriya Venkataramani, Pamela Peng, Prerit Rodney, Serge Panev, Richard Mazzarese, Nicky Liu, Michael Fukuyama, Andrii Skliar, Roger Waleffe, Duncan Riach, Yunheng Zou, Jian Hu, Hao Zhang, Binfeng Xu, Yuhao Yang, Zuhair Ahmed, Alexandre Milesi, Carlo del Mundo, Chad Voegele, Zhiyu Cheng, Nave Assaf, Andrii Skliar, Daniel Afrimi, Natan Bagrov, Ran Zilberstein, Ofri Masad, Eugene Khvedchenia, Natan Bagrov, Borys Tymchenko, Tomer Asida, Daniel Afrimi, Parth Mannan, Victor Cui, Michael Evans, Katherine Luna, Jie Lou, Pinky Xu, Guyue Huang, Negar Habibi, Michael Boone, Pradeep Thalasta, Adeola Adesoba, Dina Yared, Christopher Parisien, Leon Derczynski, Shaona Ghosh, Wes Feely, Micah Schaffer, Radha Sri-Tharan, Jeffrey Glick, Barnaby Simkin, George Zelenfroynd, Tomasz Grzegorzek, Rishabh Garg, Aastha Jhunjhunwala, Sergei Kolchenko, Farzan Memarian, Haran Kumar, Shiv Kumar, Isabel Hulseman, Anjali Shah, Kari Briski, Padmavathy Subramanian, Joey Conway, Udi Karpas, Jane Polak Scowcroft, Annie Surla, Shilpa Ammireddy, Ellie Evans, Jesse Oliver, Tom Balough, Chia-Chih Chen, Sandip Bhaskar, Alejandra Rico, Bardiya Sadeghi, Seph Mard, Katherine Cheung, Meredith Price, Laya Sleiman, Saori Kaji, Wesley Helmholz, Wendy Quan, Michael Lightstone, Jonathan Cohen, Jian Zhang, Oleksii Kuchaiev, Boris Ginsburg, Jan Kautz, Eileen Long, Mohammad Shoeybi, Mostofa Patwary, Oluwatobi Olabiyi, Andrew Tao, Bryan Catanzaro, and Udi Karpas. Nemotron 3 nano omni: Efficient and open multimodal intelligence, 2026. URL https://arxiv.org/abs/2604.24954.

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

OpenAI. Introducing GPT-5, August 2025. URL https://openai.com/index/introducing-gpt-5/.

OpenAI. Introducing GPT-6 Sol and Luna, September 2026. URL https://openai.com/index/introducing-gpt-6-sol-and-luna/.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. {CORD}: A consolidated receipt dataset for post-{ocr} parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. URL https://openreview.net/forum?id=SJl3z659UH.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. The third dihard diarization challenge, 2021. URL https://arxiv.org/abs/2012.01477.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. *International Conference on Learning Representations (ICLR)*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024a. URL https://arxiv.org/abs/2403.16999.

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*, 2024b.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Rao Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation, 2026. URL https://arxiv.org/abs/2510.06961.

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, 2025. URL https://arxiv.org/abs/2410.18325.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images, 2021. URL https://arxiv.org/abs/2101.11272.

Pulkit Verma, Ngoc La, Anthony Favier, Swaroop Mishra, and Julie A. Shah. Teaching llms to plan: Logical chain-of-thought instruction tuning for symbolic planning, 2025. URL https://arxiv.org/abs/2509.13351.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024. URL https://arxiv.org/abs/2406.18521.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. Os-atlas: A foundation action model for generalist gui agents, 2024. URL https://arxiv.org/abs/2410.23218.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. URL https://arxiv.org/abs/2404.07972.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2025a. URL https://arxiv.org/abs/2411.10440.

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, 2025b. URL https://arxiv.org/abs/2509.17765.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding, 2022. URL https://arxiv.org/abs/2012.14740.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20, pp. 1192–1200. ACM, August 2020. doi: 10.1145/3394486.3403172. URL http://dx.doi.org/10.1145/3394486.3403172.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Li Yujian and Liu Bo. A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models, 2023. URL https://arxiv.org/abs/2310.16436.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2026. URL https://arxiv.org/abs/2505.17862.

A Experimental Configuration

A.1 Evaluation Protocol

For fair comparison, we follow each benchmark's standard test-set protocol. Where official test-set grounding annotations are leaderboard-hidden (e.g., DocVQA), we use validation splits for grounding metrics while reporting ANLS on the official test split, following DLaVA (Mohammadshirazi et al., 2025). **Train/evaluation disjointness.** All benchmark examples used to construct $\mathcal{D}_1$, and hence OmniVAL-CoT, $\mathcal{D}_3$, and $\mathcal{D}_4$, are drawn exclusively from the designated training splits of the source datasets; benchmarks that provide no training split (e.g., OSWorld) contribute no examples to $\mathcal{D}_1$ and are used for evaluation only. Every example on which we report a score, including the DocVQA validation examples used for grounding mAP, is disjoint from OmniVAL-CoT and was never seen by the teacher, the validator, or the student during training. Table 3 reports the asset-level audit behind this statement: examples were grouped by underlying source asset (document, screenshot, video, clip, recording, or session), not merely by question or annotation, and checked for exact identifier or hash matches and for modality-appropriate near duplicates (perceptual image hashes, video keyframe hashes, and audio fingerprints). No exact or near-duplicate asset overlap was found on any benchmark.

The same audit over the 10,100 evaluation assets of the general-purpose benchmarks in Appendix D (4,200 images, 3,100 videos, 2,800 recordings) also found zero overlap.

Table 3: Asset-level leakage audit between the OmniVAL-CoT training splits ($\mathcal{D}_3$, $\mathcal{D}_4$) and the official evaluation splits. Representative benchmarks shown; the complete audit covers all 16.

Benchmark	Evaluation split	Grouping unit	Exact overlap	Near-duplicate overlap
DocVQA	Official test	Document ID	0	0
ScreenSpot-Pro	Official test	Screenshot ID	0	0
Video-MME	Official test	Video ID	0	0
WorldSense	Official test	Clip ID	0	0
HF Open ASR	Official eval	Recording ID	0	0
DIHARD III	Official eval	Session ID	0	0

Video input for frontier baselines. Video-MME is evaluated in its "without subtitles" setting. Models with native video input (the student, Qwen3-Omni, NExT-OMNI, Gemini 3.1 Pro, Gemini 3.8 Flash, Qwen3.8-Max, Kimi K3) receive the video file directly. Text+image models (Claude Opus 5.5, Claude Sonnet 5, GPT-6 Sol, GPT-6 Luna) receive 32 uniformly sampled frames as an ordered image sequence with the same question prompt; this follows the official Video-MME protocol for image-only LMMs and is marked § in the tables. No model receives the audio track on Video-MME. Benchmarks that require audio input (WorldSense, DailyOmni, VoiceBench, HF Open ASR, DIHARD III) are scored only for models that natively accept audio; text+image models are marked n/a rather than evaluated through a transcript. Speaker diarization on DIHARD III follows the official Track 1 "oracle SAD" protocol with no forgiveness collar and overlapped speech scored, evaluated by the standard `dscore` toolkit. HF Open ASR uses the official aggregated WER across the 8-corpus leaderboard configuration. Per-benchmark preprocessing scripts are released alongside our code.

A.2 Complete Hyperparameters

Key Configurations. Phase A: Teacher (Gemini 3.1 Pro, temperature 0.7) generates 120K modality-routed CoT traces; modality-appropriate scaffolding (DB-ResNet, SAM 2.1, Whisper v3, SigLIP, pyannote, TableFormer) provides teacher context and validation inputs (never student inputs). Stage B1: Student (Nemotron 3 Nano Omni, 30B total / 3B active MoE) receives only (M, q) and produces (CoT, a, g); learning rate 1.5×10^{-4}, effective batch 128, 3 epochs over $\mathcal{D}_3$ (88K). Stage B2: 2 epochs per iteration over $\mathcal{D}_4$ (11K); per-benchmark scores plateau after 11–15 iterations (mean 13.1), and the loop stops when $\bar{\Delta} < 0.2$ over 3 iterations. Phase C: pure omni inference without teacher, validator, or auxiliary perception modules (single-GPU deployment). We fully fine-tune all student parameters with AdamW.

Table 4 provides the complete configuration across all phases.

A.3 MA-VAL Scoring Formulas

Complete mathematical formulations for MA-VAL. All formulas use scaffolding outputs $\mathcal{S}$ for validation; scaffolding is never provided to the student. Here $\tau \in \{\text{spatial}, \text{temporal}, \text{cell}, \text{diarization}\}$ is the grounding type (subscript p: predicted, gt: ground truth), τ_c is the modality referenced in the reasoning trace, and r is the scaffolding evidence unit a grounding points to (e.g., an OCR region, ASR segment, shot, or table cell).

Component quality scores:

$$Q_{\text{ans}} = 0.7 \cdot \mathrm{ANLS}(a_p, a_{gt}) + 0.3 \cdot \mathbf{1}[a_p \in \mathcal{T}(M)] \tag{1}$$

$$Q_{\text{ground}} = 0.8 \cdot \mathrm{Overlap}_{\tau}(g_p, g_{gt}) + 0.2 \cdot \mathbf{1}[r_p = r_{gt}] \tag{2}$$

$$Q_{\text{modal}} = 0.5 \cdot \mathbf{1}[\tau_p = \tau_{gt}] + 0.3 \cdot \mathbf{1}[\tau_c = \tau_p] + 0.2 \cdot \mathbf{1}[g_p \text{ consistent with } \mathcal{S}] \tag{3}$$

$$Q_{\text{reason}} = \tfrac{1}{3}\left(S_{\text{struct}} + S_{\text{coord}} + S_{\text{align}}\right) \tag{4}$$

where S_{struct} scores whether the trace follows the routing–localization–answer structure, S_{coord} whether coordinates stated in the trace match the emitted grounding, and S_{align} whether the evidence cited in the trace matches the scaffolding content.

Modality-specific overlap operators:

$$\text{Overlap}_{\text{spatial}}(g_p, g_{gt}) = \text{IoU}_{\text{2D}}(g_p, g_{gt}) \tag{5}$$

$$\text{Overlap}_{\text{temporal}}(g_p, g_{gt}) = \text{tIoU}(g_p, g_{gt}) \tag{6}$$

$$\text{Overlap}_{\text{cell}}(g_p, g_{gt}) = \text{F1}_{\text{cell}}(g_p, g_{gt}) \tag{7}$$

$$\text{Overlap}_{\text{diarization}}(g_p, g_{gt}) = 1 - \min\big(\text{DER}(g_p, g_{gt}),\, 1\big) \tag{8}$$

All four operators take values in $[0, 1]$. In the diarization operator, DER is expressed as a fraction (e.g., 0.185), not as the percentage used in the result tables (e.g., 18.5); the clipping at 1 keeps the overlap non-negative in the rare case that DER exceeds 100%.

Aggregate quality score:

$$Q = \alpha_{\text{ans}} Q_{\text{ans}} + \alpha_{\text{ground}} Q_{\text{ground}} + \alpha_{\text{modal}} Q_{\text{modal}} + \alpha_{\text{reason}} Q_{\text{reason}} \tag{9}$$

with default weights $(\alpha_{\text{ans}}, \alpha_{\text{ground}}, \alpha_{\text{modal}}, \alpha_{\text{reason}}) = (0.35, 0.35, 0.15, 0.15)$.

Convergence detection. Let Score_j be the composite validation score on $\mathcal{D}_4$ (in points) after B2 iteration j, and $\Delta_j = \text{Score}_j - \text{Score}_{j-1}$. Then

$$\bar{\Delta}^{(k)} = \frac{1}{w} \sum_{j=k-w+1}^{k} \Delta_j \tag{10}$$

$$\text{Converged at } k \iff \bar{\Delta}^{(k)} < 0.2 \text{ and } \max_{k-w<j\leq k} |\Delta_j| < 0.4 \tag{11}$$

with $w = 3$ iterations.

A.4 Complete Algorithm

Algorithm 1 provides complete pseudocode with three key design principles: (1) scaffolding is used only by the teacher and validator (line 4), (2) the student receives only (M, q) without scaffolding inputs (line 14), and (3) MA-VAL uses scaffolding for feedback while the student learns to ground through pure multimodal reasoning (lines 23–25).

Table 4: Complete hyperparameter configuration for OmniVAL.

Parameter	**Phase A**	**Stage B1**	**Stage B2**	**Phase C**
Models				
Teacher	Gemini 3.1 Pro	—	—	—
Student	—	Nemotron 3 Nano Omni (30B total, 3B active MoE)		
Optimization				
Learning rate	N/A	1.5×10^{-4}	1×10^{-5}	N/A
Effective batch size	N/A	128	64	1
Sequence length	N/A	4096	4096	4096
Epochs	N/A	3	2 / iteration	N/A
Optimizer	N/A	AdamW	AdamW	N/A
Weight decay	N/A	0.01	0.01	N/A
Warmup steps	N/A	500	200	N/A
MA-VAL configuration				
Mode	Filter (binary)	N/A	Verifier (detailed)	N/A
$Q_{\min}$ threshold	0.85	N/A	0.85	N/A
Weights $\boldsymbol{\alpha}$ (ans, ground, modal, reason)	(.35, .35, .15, .15)	—	(.35, .35, .15, .15)	—
Throughput	48 ex/s (CPU)	—	11 ex/s (CPU)	—
Retention	91.7% (110K/120K)	—	—	—
Iterative refinement (Stage B2 only)				
Max iterations $\eta_{\max}$	—	—	20	—
Convergence window w	—	—	3	—
Threshold on $\bar{\Delta}$ (points)	—	—	0.2	—
Per-benchmark plateau (iterations)	—	—	11–15 (mean 13.1)	—
Data splits				
$\mathcal{D}_3$ (training, 80%)	—	88,000	—	—
$\mathcal{D}_4$ (refinement, 10%)	—	—	11,000	—
$\mathcal{D}_{\text{test}}$ (test, 10%)	—	—	—	11,000
Compute				
GPUs	CPU (MA-VAL)	8×H100 80GB	8×H100 80GB	1×A100 40GB
Memory (peak)	96 GB RAM	78 GB	78 GB	28 GB
Wall time	45 CPU-min	42 h	28 h	—

Algorithm 1 OmniVAL: Validated Multimodal CoT Distillation for Nemotron 3 Nano Omni

Require: $\mathcal{D}_1 = \{(M_i, q_i, a_i^{gt}, g_i^{gt})\}_{i=1}^{N}$; teacher T; scaffolding tools $\{\mathcal{S}^{(m)}\}$; student M_{θ_0}
Ensure: Fine-tuned student M_{θ^*}

$\mathcal{D}_{\text{filt}} \leftarrow \varnothing$
▷ Phase A: Teacher Generation + MA-VAL Filter
for each $(M_i, q_i, a_i^{gt}, g_i^{gt}) \in \mathcal{D}_1$ **do**
  $\mathcal{S}_i \leftarrow$ modality-appropriate scaffolding over M_i ▷ teacher/validator only
  $(\mathrm{CoT}_i^T, a_i^T, g_i^T) \leftarrow T(M_i, \mathcal{S}_i, q_i, a_i^{gt}, g_i^{gt})$
  if MA-VAL-FILTER$(M_i, \mathcal{S}_i, \mathrm{CoT}_i^T, a_i^T, g_i^T, a_i^{gt}, g_i^{gt}) > 0.85$ **then**
    $\mathcal{D}_{\text{filt}} \leftarrow \mathcal{D}_{\text{filt}} \cup \{(M_i, q_i, \mathrm{CoT}_i^T, a_i^T, g_i^T)\}$
  end if
end for
Split $\mathcal{D}_{\text{filt}} \to \mathcal{D}_3$ (88K), $\mathcal{D}_4$ (11K), $\mathcal{D}_{\text{test}}$ (11K), stratified by modality
▷ Stage B1: Supervised Multimodal CoT Learning
for epoch $e = 1$ to 3 **do**
  for each $(M_i, q_i, \mathrm{CoT}_i^{gt}, a_i^{gt}, g_i^{gt}) \in \mathcal{D}_3$ **do**
    Critical: student receives only (M_i, q_i), no scaffolding
    $\mathcal{L} \leftarrow -\sum_t \log P_\theta(c_t \mid M_i, q_i, c_{<t})$ where $c = [\mathrm{CoT}, a, g]$
    Update θ via AdamW
  end for
end for
▷ Stage B2: Iterative Refinement with MA-VAL Verifier
for iteration $k = 1$ to 20 **do**
  $\mathcal{D}_{\text{corr}}^{(k)} \leftarrow \varnothing$
  for each $(M_i, q_i, c_i^{gt}, a_i^{gt}, g_i^{gt}) \in \mathcal{D}_4$ **do**
    $(c_i^{(k)}, a_i^{(k)}, g_i^{(k)}) \leftarrow M_{\theta_{k-1}}(M_i, q_i)$ ▷ no scaffolding to student
    $F_i \leftarrow$ MA-VAL-VERIFIER$(M_i, \mathcal{S}_i, c_i^{(k)}, a_i^{(k)}, g_i^{(k)}, a_i^{gt}, g_i^{gt})$
    $\mathcal{D}_{\text{corr}}^{(k)} \leftarrow \mathcal{D}_{\text{corr}}^{(k)} \cup \{(M_i, q_i, F_i, c_i^{gt}, a_i^{gt}, g_i^{gt})\}$
  end for
  Fine-tune θ on $\mathcal{D}_{\text{corr}}^{(k)}$ for 2 epochs
  Evaluate composite score on $\mathcal{D}_4$
  if Converged **then**
    break
  end if
end for
return M_{θ^*} ▷ pure omni-modal model

B SAMPLE OUTPUTS AND PER-MODALITY EXAMPLES

B.1 TEACHER, STUDENT, AND MA-VAL INTERACTION

Teacher Generation (Phase A) – Document Modality:

Input: Receipt image, Q: "State the total."

Scaffolding (for teacher / MA-VAL): 15 detected text regions, including Region #2 $[510, 800, 570, 830] \rightarrow$ "\$45.99", Region #14 $[445, 795, 505, 825] \rightarrow$ "TOTAL:"

Teacher Output: Modality-routing: "the answer is in the document." Step-by-step spatial reasoning $\rightarrow$ Answer: \$45.99, BBox: $[510, 800, 570, 830]$

MA-VAL Filter: $Q = 1.0 > 0.85 \rightarrow$ ACCEPT

Student Prediction (Iteration 3, Stage B2) – Audio-Visual Modality:

Input (Pure Omni): Cooking video clip, Q: "State the time at which the chef adds salt." **(no scaffolding signals)**

Output: Modality-routing: "the answer is in the audio track of the video." Steps 1–4 (identifies pouring sound but mis-localizes) $\rightarrow$ Answer: "00:40", Span: $[39.2, 41.5]$

MA-VAL Verifier Feedback:

Validation: Answer: ANLS$= 0.80$ ($Q_{\text{ans}} = 0.56$), Span: tIoU$= 0.00$ ($Q_{\text{ground}} = 0$), Modality: correct ($Q_{\text{modal}} = 1.0$), Reasoning: 0.74, Overall: $Q = 0.46 < 0.85 \rightarrow$ INVALID

Detailed Feedback:

- **Answer Error:** Got "00:40" (off by 5 seconds). Ground truth: "00:45".
- **Span Error:** Predicted $[39.2, 41.5]$, ground truth $[44.6, 46.3]$ (no overlap). Move the start **5.4 s later** and the end 4.8 s later, narrowing the span by 0.6 s.
- **Reasoning Issue:** Step 3 mistook "pouring water" for "adding salt". Listen for the distinctive granular sound of salt being sprinkled, not pouring liquid.
- **Priority Fixes:** (1) Distinguish liquid pour from granular sprinkle; (2) re-localize span $\sim$5 s later; (3) verify against verbal cue ("a pinch of salt").

B.2 PER-MODALITY COT EXAMPLES

Receipt extraction (SROIE). "The question asks for the total. The answer must be a numerical value labeled 'TOTAL'. Looking at the receipt, I find the line 'TOTAL: \$45.99' near the bottom. The bounding box for the value is $[510, 800, 570, 830]$. **Answer:** \$45.99 **Box:** $[510, 800, 570, 830]$."

Chart reasoning (CharXiv). "The question asks which method scores highest on the y-axis at x=0.7. Looking at the line chart, the orange curve is at the top at x=0.7, with value approximately 0.82. The legend maps orange to 'Method C'. **Answer:** Method C **Box:** $[315, 145, 412, 188]$ (legend entry)."

GUI grounding (ScreenSpot-Pro). "The user says 'click the save button'. I see a toolbar at the top with icons. The save icon (floppy disk) is the third from the left at approximately $[148, 32, 178, 62]$. **Answer:** click **Box:** $[148, 32, 178, 62]$."

Video understanding (Video-MME). "The question asks what the speaker holds up at the climax. Scanning the video, the climax is around t=2 : 14. At t=2 : 14.3, the speaker holds up a red book. **Answer:** a red book **Span:** $[133.8, 136.1]$."

Audio (HF Open ASR). "Transcribing the audio segment $[0, 5.6]$ s: 'Welcome to the conference on multimodal learning'. **Answer:** Welcome to the conference on multimodal learning."

Speaker diarization (DIHARD III). "The audio contains two speakers. Speaker A speaks $[0.0, 4.2]$ and $[8.6, 12.1]$. Speaker B speaks $[4.2, 8.6]$. **Diarization:** `[(A, 0.0, 4.2), (B, 4.2, 8.6), (A, 8.6, 12.1)]`."

C Ablation Studies Experiments

C.1 Validation Strategy

Table 5: Validation strategy ablation. "No validation" trains Stage B1 on the full unfiltered 120K teacher pool; "Filter only" trains Stage B1 on the MA-VAL-filtered corpus with no Stage B2 refinement (identical to the Stage B1-only row of Table 7); "Filter + Verifier" is the full OmniVAL.

Validation Mode	DocVQA ANLS	DocVQA mAP	OCRBenchV2 Score	ScreenSpot-Pro / OSWorld SS-Pro	ScreenSpot-Pro / OSWorld OSWorld	Video-MME Score	HF Open ASR WER ↓	DIHARD III DER ↓
No validation (120K unfiltered)	89.7	73.1	67.4	58.9	48.9	73.1	5.62	20.4
MA-VAL Filter only (binary)	91.4	78.9	71.2	61.4	55.5	75.2	5.18	19.4
MA-VAL Filter + Verifier (ours)	**93.2**	**85.5**	**77.5**	**73.4**	**63.3**	**81.3**	**4.86**	**18.5**

C.2 Cross-Modal Reasoning Validator (Module 4)

Table 6: Cross-modal reasoning validator (Module 4) ablation. Impact is smallest on the single-modality document benchmark (DocVQA) and largest on benchmarks that require routing between input streams (WorldSense, DailyOmni) or from a textual instruction to a screen element (ScreenSpot-Pro).

MA-VAL Configuration	DocVQA ANLS	DocVQA mAP	ScreenSpot-Pro	Video-MME	WorldSense	DailyOmni
MA-VAL w/o Module 4 (no cross-modal)	92.9	84.1	63.7	76.5	60.4	77.0
MA-VAL full (ours)	**93.2**	**85.5**	**73.4**	**81.3**	**69.0**	**86.0**

C.3 Training Strategy

Table 7: Training strategy ablation. Stage B1: supervised learning on validated CoT only. B2 iterations add refinement on MA-VAL Verifier feedback.

Training Config	DocVQA ANLS	DocVQA mAP	OCRBenchV2 Score	ScreenSpot-Pro Score	Video-MME Score	DailyOmni Score	HF ASR WER ↓	DIHARD III DER ↓
Base Nemotron 3 Nano Omni (no FT)	88.9	71.4	65.8	57.8	72.2	74.1	5.95	21.4
Stage B1 only (supervised CoT)	91.4	78.9	71.2	61.4	75.2	77.6	5.18	19.4
B1 + B2 (5 iterations)	92.0	82.5	73.0	66.0	77.8	80.5	5.00	18.9
B1 + B2 (10 iterations)	92.7	84.2	75.5	69.5	79.5	83.0	4.92	18.7
B1 + B2 (converged, ours)	**93.2**	**85.5**	**77.5**	**73.4**	**81.3**	**86.0**	**4.86**	**18.5**
(iterations)	*(13)*	*(13)*	*(14)*	*(15)*	*(12)*	*(13)*	*(11)*	*(14)*

C.4 Data Scale

Table 8: Data scale ablation. All rows use full Stage B2 on $\mathcal{D}_4$ (11K).

Data Scale	Size	DocVQA ANLS	DocVQA mAP	OCRBenchV2	ScreenSpot-Pro	Video-MME	DailyOmni
25%	22K	90.5	75.5	68.0	59.8	74.0	75.8
50%	44K	91.8	80.3	71.4	63.5	76.5	78.5
75%	66K	92.6	83.5	74.5	68.5	79.0	82.2
100%	88K	**93.2**	**85.5**	**77.5**	**73.4**	**81.3**	**86.0**
Unvalidated 120K (reference)	120K	89.7	73.1	67.4	58.9	73.1	75.0

C.5 Scaffolding Tool Analysis

Table 9 analyzes the role of training-time scaffolding tools. Removing all scaffolding from Phase A (so MA-VAL validates using only ground-truth annotations and the teacher's self-consistency)

causes substantial degradation across all modalities. For documents, removing DB-ResNet text detection costs 1.8 ANLS and 8.6 mAP on DocVQA; for audio, Whisper v3 with forced alignment is essential; for diarization, pyannote provides scaffolding speaker segmentation that lets MA-VAL verify Module 3's structured output. The student, as a reminder, never sees any scaffolding signals at inference.

Table 9: Scaffolding tool ablation. Scaffolding is used only in Phase A and Stage B2 validation, never at student inference.

Scaffolding Configuration	**DocVQA** ANLS	mAP	**ScreenSpot-Pro** Score	**Video-MME** Score	**HF Open ASR** WER ↓	**DIHARD III** DER ↓
Full scaffolding (ours)	**93.2**	**85.5**	**73.4**	**81.3**	**4.86**	**18.5**
w/o DB-ResNet	91.4	76.9	73.4	81.3	4.86	18.5
w/o SAM 2.1	93.2	85.5	65.0	77.5	4.86	18.5
w/o Whisper v3	93.2	85.5	73.4	78.5	5.46	19.7
w/o pyannote (diarization)	93.2	85.5	73.4	79.5	4.86	19.9
w/o SigLIP (keyframe)	93.2	85.5	73.4	76.0	4.86	18.5
No scaffolding at all	89.6	73.4	58.4	72.7	5.74	20.8

C.6 Teacher Model Analysis (Detailed)

Table 10 compares teachers on nine grounding-heavy, non-speech benchmarks. The student and all other training settings are fixed; only the Phase A teacher changes. Every number in the table is a score of the Nemotron 3 Nano Omni *student* trained with the named teacher; it is not a score of the teacher itself. The teachers' own directly evaluated scores are the frontier rows of Tables 1–2, and the two should not be conflated: in particular, the WorldSense and DailyOmni columns are populated for every teacher because the *student* natively accepts audio and video, not because the teacher does. The table groups teachers by input coverage. For the teachers in the second group, which have no native audio input, the audio content of WorldSense- and DailyOmni-style training examples reaches the teacher only through the training-time scaffolding in $\mathcal{S}_i$ (Whisper v3 transcripts and pyannote diarization segments); this adapted teacher setting is distinct from the raw-input baseline evaluation, in which those same models are marked n/a on audio benchmarks. The Gemini 3.1 Pro row (†) is the main OmniVAL model reported in Tables 1 and 2.

Within the omni group, the only teachers eligible for the full corpus, Gemini 3.1 Pro gives the best student on every column, ahead of Gemini 3.8 Flash by 0.7 DocVQA mAP and of the earlier omni models Gemini 2.5 Flash and GPT-4o by 5.9 and 6.6; this is the comparison that determines the default teacher. Across both groups, student quality increases with the teacher's directly evaluated strength: the eight current frontier teachers rank in the same order as they do as directly evaluated baselines in Tables 1–2, and the Claude Opus 5.5-taught student is strongest on the nine non-speech benchmarks (90.3 DocVQA mAP, 4.8 above the default), followed by the GPT-6 Sol-taught student (89.6); the older teachers Claude Sonnet 4.5, GPT-5, Gemini 2.5 Flash, and GPT-4o lie 2.0–6.6 DocVQA mAP below the default, and the two open-weight teachers, Qwen3-VL-235B-A22B-Thinking and Llama 4-400B-A17B, are 2.9 and 3.6 mAP below it. An Opus-taught student cannot be trained on the full OmniVAL corpus, because Claude Opus 5.5 cannot generate traces for the speech-only examples (voice interaction, ASR, and diarization), which have no visual evidence at all; a teacher without audio input can add nothing to such examples beyond the transcript and segmentation it is given. The 90.3 row is therefore an upper bound on the non-speech portion of the corpus, not an available configuration, and the same holds for every row in the second group. Gemini 3.1 Pro remains the default teacher for this reason of input coverage, not of ranking. Because teachers without audio input cannot be meaningfully compared on speech data, the speech benchmarks (VoiceBench, HF Open ASR, DIHARD III) are excluded from Table 10, and the ranking in the second group does not establish that any of those models is a better teacher for the full omni corpus.

Why one omni teacher, and not Claude Opus 5.5 or a per-modality teacher assignment. Since Claude Opus 5.5 is the strongest directly evaluated frontier model in Tables 1 and 2 and gives the strongest student on the non-speech benchmarks in Table 10, we state explicitly why OmniVAL uses Gemini 3.1 Pro as its single teacher, and why we do not split the corpus between an omni teacher for speech and a stronger text+image teacher for everything else.

(i) Eligibility. The speech-only portion of $\mathcal{D}_1$ (voice interaction, ASR, diarization) contains no visual evidence. The only inputs a teacher without audio could receive for these examples are the Whisper transcript and the pyannote segments, which are themselves the answers; any "reasoning trace" it wrote would paraphrase the label rather than reason over evidence, and MA-VAL would have nothing to validate it against. Claude Opus 5.5 is therefore not a possible teacher for OmniVAL as defined, independently of how strong it is elsewhere. The only real alternative is a split corpus, which we reject for the two reasons that follow.

(ii) Teacher assignment presupposes the routing decision. To send an example to the right teacher, a per-modality assignment must first decide which modality its evidence lives in. For pure-document and pure-speech examples this is trivial. For the audio-visual, video, and GUI examples in $\mathcal{D}_1$, which are exactly the examples on which cross-modal mis-routing occurs (Table 11) and which Module 4 exists to supervise, it is the answer to the routing question that the trace is supposed to make explicit. Assigning these examples to a teacher without audio input would mean that the routing step of every such trace is dictated by the scaffolding pipeline, by which stream happened to be transcribed and handed over, instead of being reasoned over the raw inputs; the student would then imitate routing decisions that no model actually made. A single omni teacher that perceives every stream is the only configuration in which each routing statement in $\mathcal{D}_2$ is a decision the teacher made from the evidence.

(iii) One teacher, one trace distribution. Stage B1 fine-tunes the student on a single serialized format (routing, reasoning, answer, grounding), and MA-VAL's prompts, threshold, and correction vocabulary are calibrated once against one teacher's output. Two teachers would introduce differences in trace style, length, and confidence that are perfectly correlated with modality, so the student could learn that document reasoning and audio reasoning are different registers. That is the opposite of the property Section 5 demonstrates, one recipe improving all 16 benchmarks, and it would make the cross-modal transfer claim unfalsifiable within our own experiments, because any per-modality difference could be attributed to the teacher split rather than to the method.

(iv) What the second group of Table 10 is for. Those rows are a controlled diagnostic, obtained by giving audio to the teacher through scaffolding, of how much teacher strength matters when everything else is fixed. They establish that student quality scales with teacher strength, that OmniVAL will improve as omni teachers improve, and that open-weight teachers are viable. They are not candidate configurations, and we do not propose per-modality assignment as future work. The direct route to a stronger OmniVAL student is a stronger omni teacher.

Table 10: Teacher comparison on nine grounding-heavy, non-speech benchmarks. Every cell is a score of the Nemotron 3 Nano Omni *student* trained with the named Phase A teacher (all other settings fixed), not a score of the teacher itself; the WorldSense and DailyOmni columns are populated for all teachers because the student natively accepts audio and video. Teachers are grouped by input coverage. Only the omni teachers (native audio input) can generate traces for the speech-only portion of $\mathcal{D}_1$ and are therefore eligible teachers for the full OmniVAL corpus; teachers without native audio input see audio only through ASR/diarization scaffolding, so their rows are a comparison on the non-speech portion of the corpus only and are not available configurations for the full method. Speech benchmarks are excluded for the same reason. Rows within each group are sorted by student DocVQA mAP. Teacher reasoning-effort settings follow Table 25. †: default OmniVAL teacher; this row equals the main OmniVAL results in Tables 1 and 2. (API): teacher accessed through the provider API; the released checkpoint is text-only. **Bold**: best within each group. Abbreviations as in Table 2.

Teacher Model	DocVQA mAP	OCRB-v2	MMLB-Doc	CharXiv	SS-Pro	OSWorld	V-MME	WorldSense	DailyOmni
Omni teachers (native audio input; eligible for the full corpus)									
Gemini 3.1 Pro†	**85.5**	**77.5**	**71.3**	**77.4**	**73.4**	**63.3**	**81.3**	**69.0**	**86.0**
Gemini 3.8 Flash	84.8	76.7	70.3	76.2	72.7	62.5	80.3	67.8	85.3
Gemini 2.5 Flash	79.6	70.3	63.4	68.1	60.7	50.9	74.5	59.8	77.0
GPT-4o	78.9	69.7	62.9	67.5	60.1	50.3	74.0	59.3	76.5
Teachers without native audio input (non-speech portion only; audio reaches them through scaffolding)									
Claude Opus 5.5	**90.3**	**83.5**	**78.4**	**85.6**	**78.2**	**69.3**	**88.4**	**77.2**	**90.8**
GPT-6 Sol	89.6	82.7	77.4	84.4	77.5	68.5	87.4	76.0	90.1
Qwen3.8-Max (API)	88.9	81.8	76.4	83.2	76.8	67.6	86.4	74.9	89.4
Kimi K3 (max)	88.2	81.0	75.4	82.1	76.1	66.8	85.4	73.7	88.7
Claude Sonnet 5	86.9	79.2	73.3	79.7	74.8	65.0	83.3	71.3	87.4
GPT-6 Luna	86.2	78.4	72.3	78.5	74.1	64.2	82.3	70.2	86.7
Claude Sonnet 4.5	83.5	73.4	67.0	71.6	63.6	53.9	77.0	62.4	79.7
GPT-5	83.1	73.0	66.5	71.1	63.2	53.4	76.6	62.0	79.3
Qwen3-VL-235B-A22B-Thinking	82.6	72.5	66.0	70.6	62.8	53.0	76.2	61.5	78.9
Llama 4-400B-A17B	81.9	71.9	65.4	70.0	62.3	52.4	75.7	61.0	78.3
No teacher									
w/o teacher CoT (direct FT)	71.4	67.6	60.1	65.0	59.4	49.1	73.4	57.6	75.0

C.7 Per-Modality Cross-Modal Routing Analysis

To further analyze the contribution of MA-VAL's Module 4 (cross-modal reasoning validator), we report per-modality routing accuracy on a stratified subset of WorldSense and DailyOmni examples that explicitly require cross-modal grounding (audio-only, visual-only, or requiring both streams).

Table 11: Cross-modal routing accuracy with and without Module 4 on stratified cross-modal subsets. Without Module 4, the student frequently mis-routes to the visual stream when the answer is audio-only.

Configuration	Audio-only Q	Visual-only Q	AV-joint Q
MA-VAL w/o Module 4	76.3	89.1	73.6
MA-VAL full (ours)	**91.7**	**93.4**	**88.2**

The largest gap (15.4 points on audio-only questions) indicates that without Module 4, the student defaults to grounding in the visual stream, ignoring the audio track—exactly the cross-modal mis-routing failure mode Module 4 is designed to address.

D Additional Analyses

This appendix collects controlled analyses that probe the claims of the main paper from directions the main ablations do not cover: transfer to general-purpose and hallucination benchmarks, disentanglement of the supervision signal, robustness of MA-VAL to noisy or swapped scaffolding tools, the cost of trace generation and refinement, inference-time reasoning cost, causal reliance on the correct modality, and transfer of the pipeline to a second student.

Protocol. Rows labeled *OmniVAL* are the final model of Tables 1–2, and rows labeled *No scaffolding* coincide with the corresponding row of Table 9. The second-student transfer uses a task-balanced

44K-example subset of $\mathcal{D}_2$ with five Stage B2 iterations, and the supervised-versus-RL pilot uses 5K examples, to keep the number of controlled reruns tractable; these are marked as such, and their absolute values are lower than full-corpus training by construction. Every table is internally controlled: all rows within a table share the same data, training budget, optimizer, and inference policy.

D.1 Transfer to General-Purpose Image, Video, and Audio Benchmarks

The 16 main benchmarks are grounding-oriented, and several align with the type of evidence the training-time scaffolding tools produce (OCR boxes, temporal spans, speaker turns). To test whether the gains transfer to settings where no such structure is available, we evaluated the unchanged checkpoints on six general-purpose benchmarks (Table 12): MMBench (Liu et al., 2024) and MMMU (Yue et al., 2024) for image understanding, MVBench (Li et al., 2024a) and Video-MMMU (Hu et al., 2025) for video, and MMAU (Sakshi et al., 2025) and MMAR (Ma et al., 2025) for audio reasoning. None of these benchmarks was used to construct $\mathcal{D}_1$, and the asset audit of Appendix A.1 found zero overlap with OmniVAL-CoT.

Table 12: General-purpose benchmarks not used in training. Accuracy (%); higher is better. Macro Avg. is the unweighted mean over the six benchmarks.

Model	MMBench	MMMU	MVBench	Video-MMMU	MMAU	MMAR	Macro Avg.
Base Nemotron 3 Nano Omni	68.4	42.7	58.2	39.8	61.5	53.2	53.97
Unvalidated CoT	70.1	44.8	62.5	42.3	65.8	57.6	57.18
MA-VAL Filter only	71.2	45.9	65.1	44.0	68.2	60.1	59.08
OmniVAL	**72.6**	**47.3**	**67.8**	**46.2**	**70.9**	**63.4**	**61.37**

OmniVAL improves the macro-average by $+7.40$ points over the base model, $+4.19$ over unvalidated CoT, and $+2.29$ over Filter-only, with consistent gains across image VQA, pure video understanding, and pure audio reasoning ($+9.6$ MVBench, $+9.4$ MMAU, and $+10.2$ MMAR over the base). The ordering base $<$ unvalidated $<$ filtered $<$ full is the same as on the grounding benchmarks, so the validated supervision transfers to unstructured, tool-unseen tasks rather than only to tasks that resemble the scaffolders.

D.2 Cross-Modal Hallucination Benchmarks

Because OmniVAL's central claim concerns grounding in the correct modality, we also evaluated on two benchmarks designed to expose cross-modal hallucination (Table 13): the Curse of Multi-Modalities benchmark (CMM) (Leng et al., 2024), which reports perception accuracy (PA) and hallucination resistance (HR) overall and on visually- and audio-dominant subsets, and AVHBench (Sung-Bin et al., 2025), which tests audio-to-visual and visual-to-audio hallucination.

Table 13: Cross-modal hallucination benchmarks. All columns are accuracies (%); higher is better. Vis.-Dom. and Aud.-Dom. HR are CMM hallucination resistance on the visually- and audio-dominant subsets.

Model	CMM PA	CMM HR	Vis.-Dom. HR	Aud.-Dom. HR	AVH A→V	AVH V→A
Base Nemotron 3 Nano Omni	67.2	58.4	62.1	54.8	61.3	59.7
Unvalidated CoT	72.8	66.5	69.8	63.2	67.1	65.4
OmniVAL w/o Module 4	76.1	71.3	74.2	68.5	72.0	70.8
OmniVAL	**79.4**	**76.8**	**78.9**	**74.7**	**76.2**	**75.5**

Relative to the base model, OmniVAL improves CMM hallucination resistance by 18.4 points, audio-dominant HR by 19.9, and AVHBench by 14.9 and 15.8 points in the two directions. Removing Module 4 costs 5.5 CMM-HR, 6.2 audio-dominant HR, and 4.2/4.7 AVHBench points, which ties the improvement to modality-route validation specifically rather than to fine-tuning in general.

D.3 Disentangling CoT, Grounding Supervision, Validation, and Refinement

The main ablations vary the validation strategy but not the composition of the supervision itself. To separate the contribution of natural-language reasoning traces from that of grounding targets, and both from validation, we ran a factorial ablation with the same examples and the same Stage B1 training budget (Table 14); only the last row adds Stage B2.

Table 14: Factorial supervision ablation (same examples and Stage B1 budget; the last row adds Stage B2).

Supervision	DocVQA ANLS	DocVQA mAP	OCRBenchV2-En	ScreenSpot-Pro	Video-MME	DailyOmni
Answer only	90.5	74.0	69.0	58.0	73.0	75.0
Answer + grounding	91.0	77.5	70.0	60.5	74.5	76.5
Answer + CoT	91.5	75.5	70.5	58.5	74.0	75.8
Unvalidated CoT + grounding	89.8	72.8	67.8	57.5	72.8	74.8
Filtered CoT + grounding	91.4	78.9	71.2	61.4	75.2	77.6
OmniVAL (+ Stage B2)	**93.2**	**85.5**	**77.5**	**73.4**	**81.3**	**86.0**

Each ingredient contributes separately, and one of them can hurt. Grounding targets add $+3.5$ DocVQA mAP over answer-only training and reasoning text alone adds $+1.5$. Combining *unvalidated* teacher traces with their grounding targets, however, falls to 72.8 mAP, below answer-only training: the reasoning text is harmless (answer+CoT is 75.5), but the hallucinated boxes and spans in unfiltered traces act as wrong grounding labels and undo the benefit of grounding supervision. Filtering the traces with MA-VAL recovers this and more (78.9, $+6.1$ over unvalidated), and Stage B2 refinement adds a further $+6.6$ mAP with gains on every column, the largest on ScreenSpot-Pro ($+12.0$) and DailyOmni ($+8.4$). The improvements therefore do not come from natural-language CoT alone: grounding supervision, trace validation, and iterative correction each make a distinct contribution, validation is what makes teacher grounding usable at all, and CoT is complementary to grounding rather than a substitute for it. This also bears on whether explicit reasoning is always beneficial (Liu et al., 2026): reasoning text by itself moves DocVQA mAP by only $+1.5$; the large localization gains come from validated grounding and corrective refinement.

D.4 Does Verifier Feedback Help Beyond Additional Training?

Stage B2 performs additional supervised updates, so its gains could in principle come from more optimization rather than from the content of the feedback. We matched the Stage B2 examples and the number of update steps exactly and varied only what the feedback contained (Table 15).

Table 15: Stage B2 feedback-content control (identical examples and update steps).

Stage B2 supervision	DocVQA mAP	ScreenSpot-Pro	DailyOmni	DER ↓
Corrected target only	83.2	65.0	80.0	19.0
Generic "try again" feedback	83.8	67.2	81.8	18.8
Modality-specific deltas	84.5	69.5	83.7	18.6
Full MA-VAL feedback (OmniVAL)	**85.5**	**73.4**	**86.0**	**18.5**

With identical examples and update counts, the content of the feedback moves the result monotonically: corrected targets alone reach 83.2 DocVQA mAP and 65.0 ScreenSpot-Pro, generic feedback 83.8 and 67.2, modality-specific pixel, time, cell, and routing corrections 84.5 and 69.5, and the full MA-VAL feedback (error diagnosis, corrected target, and prioritized fixes) 85.5 and 73.4, best on all four metrics. The spread is largest where localization is scored (8.4 points on ScreenSpot-Pro, 6.0 on DailyOmni) and small on DER (0.5). Stage B2 gains therefore arise from the structured content of the verifier feedback, not from extra optimization steps.

D.5 Refinement Schedule, Selective Refinement, and Supervised versus RL Refinement

The adaptive-convergence schedule of Section 3.4 refines every failed example on every iteration until the composite score plateaus. To measure how much of its benefit is recoverable at lower cost, we evaluated single-pass and selective schedules from the same Stage B1 checkpoint (Table 16). "Lowest 25%" refines only the quarter of $\mathcal{D}_4$ with the lowest MA-VAL quality score; "selective" re-scores after each pass and refines only examples still below threshold.

Table 16: Refinement-schedule ablation from the common Stage B1 checkpoint of Table 7. The Stage B1 and five-pass rows equal the corresponding rows of Table 7; the adaptive row is the final model.

Schedule	Refined fraction	Iterations	Student generations	B2 hours	DocVQA mAP	ScreenSpot-Pro	Video-MME	DailyOmni
Stage B1 only	0%	0	0	0	78.9	61.4	75.2	77.6
Single pass, all failures	all failed	1	11K	4.5	80.8	64.0	76.5	79.2
Single pass, lowest 25%	25%	1	2.75K	1.1	80.2	63.2	76.1	78.8
Selective refinement	dynamic	3	5.2K	2.1	81.8	65.2	77.2	80.2
Full five-pass refinement	all failed	5	55K	22.5	82.5	66.0	77.8	80.5
Adaptive convergence (OmniVAL)	dynamic	11–15	147K	28.0	**85.5**	**73.4**	**81.3**	**86.0**

The table traces a cost–accuracy curve. Measured as the fraction of the Stage B1-to-converged gain recovered, a single pass over the lowest-scoring 25% (1.1 hours, 4% of the full cost) recovers 20% on DocVQA mAP and 15% on ScreenSpot-Pro; a single pass over all failures (4.5 hours) recovers 29% and 22%; three selective passes (2.1 hours, 7.5% of the cost) recover 44% and 32%; five full passes (22.5 hours) recover 55% and 38%; and adaptive convergence (28.0 hours) recovers the rest. Two points follow. Selective refinement is the efficient operating point when compute is constrained: it matches or exceeds five full passes on no metric but recovers roughly half of the attainable gain at less than a tenth of the cost, because MA-VAL's scores identify the examples that still contain recoverable error. And the later iterations are not wasted: the benchmarks that score precise localization (ScreenSpot-Pro, DailyOmni) gain the most from the iterations after the fifth, which is consistent with Table 7 and with the finding in Appendix D.4 that fine-grained corrections matter most where localization is scored. Adaptive convergence is the procedure used for the main results; selective refinement is the recommended configuration under a compute budget, and neither adds any inference-time cost.

Supervised refinement versus RL. Stage B2 is supervised: the input, the student's response, and the structured feedback are serialized into an example whose target is the corrected response, and the student is updated with token-level cross-entropy. A natural alternative is to use the validator score as a reward. In a reduced-scale pilot in which Stage B1 and each refinement method are trained on the same 5K examples with three refinement iterations per method, we compared supervised refinement with GRPO (Shao et al., 2024b) on the MA-VAL score and with the two in sequence (Table 17).

Table 17: Supervised versus RL refinement in a reduced-scale pilot. Stage B1 and every refinement method are trained on the same 5K examples, with three refinement iterations per method; the first row is the shared 5K-trained Stage B1 starting checkpoint. Absolute values are therefore below the full-corpus Stage B1 of Table 7, and only within-table differences are meaningful.

Stage B2 method	Signal	H100-hours	DocVQA mAP	Video-MME	CMM HR
Stage B1 only (5K)	—	0	76.3	74.5	58.4
SFT	structured feedback	**28**	**81.2**	**76.4**	**68.1**
GRPO	scalar reward	40	78.5	75.1	63.2
SFT $\rightarrow$ GRPO	combined	40	**81.8**	**76.8**	**68.9**

Supervised refinement exceeds GRPO alone by 2.7 DocVQA mAP, 1.3 Video-MME, and 4.9 CMM-HR while using less compute, which is consistent with MA-VAL providing dense, localized corrections that token-level supervision preserves and that a scalar reward compresses away. GRPO after SFT adds a further 0.4–0.8 points, so RL is complementary rather than a replacement; a full-scale RL study is left to future work.

D.6 Scaffolded versus Blind Teacher Generation

Scaffolding outputs are supplied to the teacher during trace generation (Section 3.3, Algorithm 1), not only to MA-VAL afterwards. To quantify what this buys, we compared scaffolded one-shot generation with blind generation and with blind generation plus resampling on the same stratified source set (Table 18). Acceptance is the fraction of traces passing the MA-VAL filter; calls and tokens are teacher usage per accepted trace.

Table 18: Teacher-generation efficiency. Routing accuracy and grounding quality are MA-VAL scores of the accepted traces.

Generation pipeline	Scaffolding	Max samples	First-pass accept.	Final accept.	Calls/accepted ↓	Tokens/accepted ↓	Routing acc.	Grounding
Blind generation	no	1	54.2	54.2	1.85	4,120	68.7	62.4
Blind + resampling	no	3	54.2	81.6	2.01	4,480	72.3	68.1
OmniVAL generation	yes	1	**91.7**	91.7	**1.09**	**2,560**	94.2	91.8
Scaffolded + one correction	yes	2	91.7	**94.2**	1.14	2,750	**95.1**	**93.2**

Against blind three-sample resampling, scaffolded one-shot generation raises final acceptance by 10.1 points, cuts teacher calls per accepted trace by 45.8% and teacher tokens by 42.9%, and improves routing accuracy by 21.9 and grounding quality by 23.7 points. The pipeline therefore does not rely on repeated sampling to find acceptable traces; scaffolding improves the trace before validation, and validation then removes what remains.

D.7 Robustness to Noisy Scaffolding Tools

Table 9 removes scaffolding tools outright. Real tools are imperfect rather than absent, so we also corrupted each tool's output gradually while leaving the other channels intact: OCR token and box noise, ASR substitutions and timestamp shifts, mask erosion and dilation, and diarization boundary and speaker-label noise. Mild, moderate, and severe correspond to matched corruption levels such as 5/10/20% text noise, 2/5/10% box jitter, and 100/250/500 ms temporal jitter (Table 19).

Table 19: Graceful degradation under scaffolding noise. Trace retention is the MA-VAL acceptance rate on $\mathcal{D}_1$.

Noise level	Trace retention	DocVQA mAP	Video-MME	WER ↓	DER ↓
Clean (OmniVAL)	91.7%	85.5	81.3	4.86	18.5
Mild	89.0%	84.8	80.6	4.90	18.7
Moderate	82.0%	83.2	78.8	5.05	19.2
Severe	70.0%	80.5	76.0	5.35	20.0
No scaffolding	—	73.4	72.7	5.74	20.8

As tools become noisier, retention falls because more corrupted traces are rejected (91.7% to 70.0%), while downstream performance declines much more slowly: mild corruption costs 0.7 DocVQA mAP and 0.7 Video-MME, moderate 2.3 and 2.5, and even severe corruption (5.0 and 5.3) leaves every metric better than training without scaffolding. MA-VAL thus acts as a training-time shield against imperfect perception tools rather than a channel that propagates their errors into the student.

Alternative scaffolding tools. We also replaced each default tool while holding the rest of the pipeline fixed. For OCR, DB-ResNet yields 85.5 DocVQA mAP, TrOCR (Li et al., 2023a) 83.1, and PaddleOCR (Cui et al., 2025) 84.9; for ASR, Whisper v3 yields 4.86 WER, Whisper v2 4.98, and wav2vec 2.0 (Baevski et al., 2020) 5.32; for diarization, pyannote yields 18.5 DER, VBx (Landini et al., 2020) 19.3, and a ResNet34 x-vector system 20.1. Stronger standalone tools give higher retention and better students, but no single tool is essential: MA-VAL consumes normalized evidence from interchangeable scaffolders.

D.8 INFERENCE-TIME REASONING COST

Longer outputs can improve accuracy on their own, so we measured what the trained student actually emits at inference. Held-out questions were split into perception/localization and reasoning-intensive subsets, and every model was run under three inference policies on the same checkpoint: *direct* (answer only), *always-CoT* (explicit reasoning), and *adaptive*, which selects between them according to whether the query requires simple perception or multi-step cross-modal reasoning. No parameters differ between policies. Token counts are the student's own outputs; teacher traces are absent at inference (Table 20).

Table 20: Accuracy (%) and output length under three inference policies on the same held-out set. Avg./P90 are output tokens per response.

Model	Policy	Perception	Reasoning	Grounding	Overall	Avg./P90 tokens
Base Nemotron 3 Nano Omni	direct	73.8	65.2	67.5	69.8	142 / 387
NExT-OMNI	direct	68.2	58.4	60.1	63.2	138 / 372
NExT-OMNI	always-CoT	66.8	64.5	66.2	66.1	398 / 1,012
NExT-OMNI	adaptive	69.1	63.8	65.7	66.2	244 / 658
Qwen3-Omni	direct	72.4	64.8	68.2	68.8	145 / 389
Qwen3-Omni	always-CoT	70.8	72.3	74.0	72.8	408 / 1,038
Qwen3-Omni	adaptive	73.2	71.2	73.5	72.6	252 / 675
OmniVAL	direct	78.5	71.0	76.3	74.9	148 / 395
OmniVAL	always-CoT	77.2	**79.8**	**82.1**	**79.2**	412 / 1,047
OmniVAL	adaptive	**79.1**	78.5	81.4	79.0	256 / 687

Three observations follow. First, direct answering beats always-on CoT for simple perception for every model, while explicit CoT improves reasoning by 8.8 points and grounding by 5.8 over OmniVAL-direct; reasoning should be invoked selectively, and the adaptive policy retains 79.0 of the 79.2 always-CoT overall score while emitting 37.9% fewer tokens on average and 34.4% fewer at P90. Second, output length does not explain OmniVAL's advantage over other omni models: at nearly identical token counts, OmniVAL-direct exceeds Qwen3-Omni-direct by 6.1 points and NExT-OMNI-direct by 11.7, and the gaps remain 6.4/13.1 under always-CoT and 6.4/12.8 under adaptive inference. Third, under identical output caps OmniVAL leads at every budget (Table 21); OmniVAL-adaptive at 128 tokens (73.4) already exceeds Qwen3-Omni at 512 tokens (72.4).

Table 21: Overall accuracy (%) under identical output-token caps.

Model / policy	128 tokens	256 tokens	512 tokens
NExT-OMNI + CoT	58.2	63.8	65.9
Qwen3-Omni + CoT	64.1	69.4	72.4
OmniVAL + CoT	71.8	75.2	78.8
OmniVAL + adaptive	**73.4**	**76.8**	**79.0**

D.9 CAUSAL MODALITY INTERVENTIONS

To test whether the trained student actually relies on the correct evidence rather than merely producing better-formatted answers, we intervened on the inputs of a fixed evaluation subset: removing the answer-relevant modality (accuracy should drop), removing the irrelevant modality (accuracy should not drop), and replacing irrelevant evidence with mismatched distractors (the answer should not flip).

Table 22: Causal modality interventions on a fixed evaluation subset. Relevant drop: accuracy loss when the answer-relevant modality is removed (higher means stronger reliance on the right evidence). Irrelevant drop and distractor flip: lower is better. Hallucination rate: lower is better. Grounding: accuracy of the emitted grounding.

Model	Relevant drop ↑	Irrelevant drop ↓	Distractor flip ↓	Hallucination ↓	Grounding ↑
Base Nemotron 3 Nano Omni	24.7	12.3	28.6	34.2	61.8
Unvalidated CoT	29.1	10.8	22.1	26.7	68.4
OmniVAL w/o Module 4	32.4	9.5	18.3	22.1	72.0
OmniVAL	**36.8**	**6.2**	**12.7**	**16.5**	**78.6**

Compared with the base model, OmniVAL depends more on the genuinely relevant evidence ($+12.1$ relevant-modality drop), less on irrelevant evidence (-6.1), and is far less likely to flip under distractors (-15.9). Table 23 decomposes the errors on the same subset by type: answer errors fall from 28.3% to 14.2%, grounding errors from 38.2% to 21.4%, routing errors from 21.4% to 6.8%, and hallucinations from 34.2% to 16.5%, with the largest relative reduction (68%) on routing errors, the failure mode Module 4 targets. The final model has therefore changed not only in answer accuracy but in its causal reliance on the correct modality and grounding target, and removing Module 4 accounts for a substantial part of that change.

Table 23: Error decomposition on the fixed evaluation subset of Table 22 (% of examples; lower is better). The hallucination column coincides with the hallucination-rate column of Table 22.

Model	Answer errors	Grounding errors	Routing errors	Hallucinations
Base Nemotron 3 Nano Omni	28.3	38.2	21.4	34.2
OmniVAL	**14.2**	**21.4**	**6.8**	**16.5**
Relative reduction	50%	44%	68%	52%

D.10 Transfer to a Second Student: Qwen3-Omni

OmniVAL makes no assumption specific to Nemotron 3 Nano Omni. We applied the identical pipeline to Qwen3-Omni (the open-weight Qwen3-Omni-30B-A3B checkpoint) on the 44K subset with five Stage B2 iterations (Table 24).

Table 24: OmniVAL applied to Qwen3-Omni (44K subset, five Stage B2 iterations). The base model is prompted to emit boxes for this experiment so that DocVQA mAP can be measured; Table 1 reports — for it because its default output contains no boxes.

Model	DocVQA mAP	ScreenSpot-Pro	Video-MME	DailyOmni	WER ↓	DER ↓
Qwen3-Omni (base)	74.2	56.4	71.0	72.5	6.41	22.9
+ Unvalidated CoT	78.5	59.1	73.8	75.2	5.92	21.6
+ OmniVAL	**81.9**	**61.7**	**76.1**	**77.8**	**5.44**	**20.3**

OmniVAL improves Qwen3-Omni on every representative benchmark over both the base model and the unvalidated-CoT variant, with the same ordering as for Nemotron 3 Nano Omni. The framework is therefore not specific to one student family; the main paper uses Nemotron 3 Nano Omni because it is the strongest of the open-weight true native omni models before fine-tuning (Section 4).

E Frontier Baseline Roster and Input Coverage

Table 25 summarizes the frontier models we compare against, together with the input modalities each accepts according to its official documentation and the reasoning-effort setting used in our evaluation. The row order, which is used in all result tables, is the models' ranking under our evaluation protocol (Tables 1–2); it is monotone across every benchmark on which a model can be evaluated. Among the frontier models only the two Gemini models (3.1 Pro and 3.8 Flash) accept

audio, so the others cannot be evaluated on VoiceBench, HF Open ASR, DIHARD III, WorldSense, or DailyOmni. The result tables therefore group the frontier models by input coverage, following Section 2: the six models without native audio input form one group, and the closed-source true native omni models form another, comprising the two Gemini models and the earlier GPT-4o, which is evaluated on all 16 benchmarks; our student, Nemotron 3 Nano Omni, accepts all four modalities. The released Qwen3.8-Max weights (Qwen3.8-2.4T-A95B) are text-only, so its multimodal results use the Qwen3.8-Max API; these results do not evaluate the released checkpoint. Tables 1–2 report all results on the 16 benchmarks for every model in this roster and for the student before and after OmniVAL; the six models without audio input are marked n/a on the WorldSense, DailyOmni, and speech columns of Table 2.

Table 25: Comparison roster: frontier models, student, and open-source omni baselines (September 2026). Rows are ordered by the models' scores under our evaluation protocol (Tables 1–2) and grouped by input coverage (frontier models without native audio input; closed-source true native omni models); the same order and grouping are used in all result tables. Parameter counts (total / active per token) are those published by each provider; closed-source providers do not disclose them. †: default OmniVAL teacher. Effort settings are those used when we evaluate each model through its API.

Model	Provider	Weights	Params (total / active)	Input modalities	Reasoning effort	Ref.
Claude Opus 5.5	Anthropic	closed	undisclosed	text, image	default (medium)	Anthropic (2026a)
GPT-6 Sol	OpenAI	closed	undisclosed	text, image	default (medium)	OpenAI (2026)
Qwen3.8-Max (API)	Alibaba	API*	2.4T / 95B	text, image, video	default	Qwen Team (2026)
Kimi K3 (max)	Moonshot AI	open	2.8T / 104B	text, image, video	max	Moonshot AI (2026)
Claude Sonnet 5	Anthropic	closed	undisclosed	text, image	default (high)	Anthropic (2026b)
GPT-6 Luna	OpenAI	closed	undisclosed	text, image	default (medium)	OpenAI (2026)
Gemini 3.1 Pro†	Google	closed	undisclosed	text, image, audio, video	default (high)	Google DeepMind (2026a)
Gemini 3.8 Flash	Google	closed	undisclosed	text, image, audio, video	default (medium)	Google DeepMind (2026b)
GPT-4o	OpenAI	closed	undisclosed	text, image, audio, video	default	OpenAI (2024)
Nemotron 3 Nano Omni (student)	NVIDIA	open	30B / 3B	text, image, audio, video	default	NVIDIA et al. (2026)
Qwen3-Omni (30B-A3B)	Alibaba	open	30B / 3B	text, image, audio, video	default	Xu et al. (2025b)
NExT-OMNI	CAS / NUS	open	7B††	text, image, audio, video	default	Luo et al. (2025)

*The released Qwen3.8-Max checkpoint (Qwen3.8-2.4T-A95B) is text-only; all multimodal results in this paper are obtained through the provider API and do not evaluate the released weights. ††Language backbone initialized from Qwen2.5-7B; the paper does not report a separate total.

F LIMITATIONS

OmniVAL narrows but does not eliminate the frontier-model gap, and its teacher must be an omni model with native audio input, so it cannot yet draw on the strongest text+image reasoners; its student will improve as omni teachers do. Training-time supervision depends on scaffolding quality. MA-VAL checks answers, grounding, modality consistency, and coordinate alignment, but it cannot establish that a reasoning trace is faithful. Our benchmarks mostly assume a single grounding target, and multi-evidence grounding remains open. MA-VAL's threshold and weights are hand-set and we have not studied their sensitivity; Appendix D shows that supervision degrades gracefully rather than not at all as scaffolding tools become noisy; and our leakage audit is asset-level and cannot exclude topical overlap between benchmark sources.

G BROADER IMPACT

OmniVAL can have positive societal impact by improving the verifiability and auditability of omni-modal model outputs. By requiring answers to be grounded in modality-specific evidence, such as document regions, GUI coordinates, video moments, audio spans, table cells, or speaker turns, the framework can help users inspect model predictions and identify errors more easily in applications such as accessibility, document analysis, human-computer interaction, and assistive multimodal agents. However, the same grounding capabilities may also introduce risks if deployed in sensitive domains without safeguards. More accurate localization over documents, audio, video, or speaker turns could be misused for surveillance, privacy-invasive analysis, or automated decision support without sufficient human oversight. Incorrect grounding may also create false confidence when a model provides a plausible but wrong evidence location. These risks motivate careful deployment

practices, including privacy-preserving data handling, human review in high-stakes settings, clear uncertainty communication, and restrictions on uses involving surveillance or sensitive personal data.

H Responsible Release and Safeguards

OmniVAL-CoT and the accompanying implementation are released for research use. The dataset is constructed from established benchmark sources and validator-generated annotations; we do not release private user data, scraped personal data, or non-consensual recordings. Data access and use should follow the licenses and terms of the underlying benchmark datasets. To reduce misuse risk, the release will include documentation describing intended research uses, restrictions against surveillance or privacy-invasive applications, and instructions for reproducing experiments without using the system for sensitive personal identification or automated high-stakes decision-making. The released code will also separate training-time validation components from inference-time deployment to make clear that MA-VAL is intended for supervision and auditing rather than covert monitoring.

I Existing Assets and Licenses

OmniVAL uses existing benchmark datasets, pretrained models, and training-time scaffolding tools. We cite the original creators of all assets in the main paper and use each asset according to its corresponding license, terms of use, or access policy. OmniVAL-CoT is derived from established benchmark sources and validator-generated annotations; redistribution will follow the licenses and terms of the underlying datasets. Table 26 summarizes the assets used in this work. We will include license files, attribution, and usage restrictions in the released repository.

Table 26: Existing assets used in OmniVAL. We follow the official licenses, access policies, and terms of use associated with each asset.

Asset type	Assets	Use in this work	License / terms
Benchmarks	DocVQA, VisualMRC, FUNSD, CORD, SROIE, OCRBenchV2-En, MMLongBench-Doc, CharXiv	Training splits contribute document and chart examples to $\mathcal{D}_1$ / OmniVAL-CoT; held-out validation/test splits are used exclusively for evaluation	Per official dataset releases (e.g., DocVQA/FUNSD terms; Hugging Face dataset cards where hosted)
Benchmarks	ScreenSpot-Pro, OSWorld, Video-MME, WorldSense, DailyOmni, VoiceBench, HF Open ASR, DIHARD III	Training splits, where available, contribute GUI, video, audio-visual, voice, ASR, and diarization examples to $\mathcal{D}_1$ / OmniVAL-CoT; held-out validation/test splits are used exclusively for evaluation	Per official releases; DIHARD III via LDC/Data Hub agreements; HF assets under dataset/repo licenses
Student / baselines	Nemotron 3 Nano Omni, Qwen3-Omni, NExT-OMNI; Claude Opus 5.5, GPT-6 Sol, Qwen3.8-Max, Kimi K3, Claude Sonnet 5, GPT-6 Luna, Gemini 3.1 Pro, Gemini 3.8 Flash	Student model, open-source omni baselines, and frontier comparison systems	Open weights under their model licenses (NVIDIA Open Model Agreement for Nemotron 3 Nano Omni; Apache-2.0 for Qwen3-Omni-30B-A3B; NExT-OMNI repository license; Kimi K3 License; Qwen3.8-Max License); closed models via provider API terms
Teacher / ablation models	Gemini 3.1 Pro (default), the frontier models above, Claude Sonnet 4.5, GPT-5, Gemini 2.5 Flash, Qwen3-VL-235B-A22B-Thinking, Llama 4-400B-A17B	Teacher generation and teacher-choice ablations	API or downloadable weights per vendor license and acceptable-use policies
Training-time scaffolding tools	DB-ResNet, SAM 2.1, Whisper v3, SigLIP, pyannote, TableFormer	Validation and feedback generation only; not used at inference	Open-source licenses per upstream repositories (e.g., MIT, Apache-2.0)

J New Asset Documentation

This paper introduces **OmniVAL-CoT**, a corpus of 110K validator-verified multimodal chain-of-thought traces spanning documents, charts, GUI screens, video, audio, speech, and diarization. Each

example contains a multimodal input reference, a question, an answer, a modality-routed reasoning trace, and modality-appropriate grounding supervision such as a bounding box, time interval, cell index, or speaker turn. The corpus is constructed from established benchmark sources and teacher-generated reasoning traces filtered by MA-VAL; no private user data or newly scraped personal data is introduced.

The released asset will include documentation describing dataset composition, data fields, modality coverage, construction pipeline, validation criteria, train/refinement/test splits, intended research uses, limitations, and license or terms-of-use constraints inherited from the underlying benchmark datasets. The release will also include instructions for reproducing the MA-VAL filtering process and for using OmniVAL-CoT in Stage B1 supervised fine-tuning and Stage B2 validation-driven refinement. To preserve anonymity during review, the release is provided through an anonymized repository.