OMNI-PERSONA: SYSTEMATIC BENCHMARKING AND IMPROVING OMNIMODAL PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While multimodal large language models have advanced across text, image, and audio, personalization research has remained primarily vision-language, with unified omnimodal benchmarking that jointly covers text, image, and audio still limited, and lacking the methodological rigor to account for absent-persona scenarios or systematic grounding studies. We introduce Omni-Persona, the first comprehensive benchmark for omnimodal personalization. We formalize the task as cross-modal routing over the *Persona Modality Graph*, encompassing 4 task groups and 18 fine-grained tasks across $\sim$750 items. To rigorously diagnose grounding behavior, we propose *Calibrated Accuracy (*Cal*)*, which jointly evaluates correct grounding and appropriate abstention, incorporating absent-persona queries within a unified evaluation framework. On our dedicated experiments, three diagnostic findings emerge: (i) recent open-weight models show a consistent audio-vs-visual grounding gap that RLVR partially narrows via dense rule-based supervision; (ii) recall score and parameter scale are incomplete diagnostics, since strong recall can coexist with absent-persona hallucination and larger models do not always achieve higher Cal, exposing calibration as a separate evaluation axis; and (iii) SFT is limited by the scalability of ground-truth annotation. RLVR improves recall but not calibrated abstention under our reward design, leaving Cal at or below the base model. Omni-Persona thus serves as a diagnostic framework that surfaces the pitfalls of omnimodal personalization, guiding future post-training and reward design.

1 INTRODUCTION

The landscape of large generative models has expanded rapidly toward *omnimodal* systems capable of processing or even generating across text, image, and audio within a single model (Xu et al., 2025b; OpenBMB, 2026; Comanici et al., 2025; Gemma Team et al., 2026; OpenAI et al., 2024). This convergence of modalities broadens the task scope that a single model can handle and moves the community closer to the vision of a personal AI assistant, one that can recognize a user's face and voice, recall their biographical context, and ground responses in individual identity.

Despite this momentum, multimodal personalization research has remained primarily focused on vision-language settings (Nguyen et al., 2024; Hao et al., 2025; Oh et al., 2025; 2026), leaving three key gaps that limit progress toward true omnimodal deployment. First, existing benchmarks have rarely provided unified coverage across all three modalities: while vision and text are well represented, systematic treatment of audio signals such as voice identity, emotional tone, and conversational context remains limited. Second, real-world retrieval is inherently noisy, often yielding contexts where the queried identity is completely *absent*. Yet, personalization is typically evaluated under *well-controlled settings*, such as explicit identity naming (Alaluf et al., 2024; Nguyen et al., 2024; Oh et al., 2025) or carefully designed caption-based distractors (Oh et al., 2026), that assume the target is always present. Consequently, these artificial setups and their recall-only protocols fail to expose this critical failure mode. Third, *realistic personalization scenarios* (for example, identifying a person from a face image or voice clip and then answering a query about that individual) have not been systematically studied. Without a benchmark that addresses all three gaps, the community lacks a principled way to diagnose *when* and *how* current omnimodal models fail at personal grounding. While recent studies (Mei et al., 2026; Kim et al., 2025a) each address important aspects of the multimodal personalization problem, substantial gaps remain in audio grounding, absent-persona coverage, and realistic evaluation.

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.

To this end, we introduce *Omni-Persona*, the first evaluation-only benchmark for omnimodal personalization, offering systematic cross-modal coverage with full support for *audio* as a persona modality and *absent-persona* cases. We formalize each user's multimodal profile through the *Persona Modality Graph* (PMG). In this graph-based abstraction, individual user profiles (comprising a profile image, biographical text, and personal audio) act as context nodes. We frame omnimodal personalization as a cross-modal routing problem: the model must evaluate incoming queries and correctly establish a directed linkage (edge) to the matching context node to ground its response.

Omni-Persona spans 4 task groups and 18 fine-grained tasks over $\sim$750 evaluation items, enabling evaluation of both perceptual matching and grounded retrieval. To reflect real-world retrieval imperfections, we include absent-persona samples, where the ground-truth persona is missing from the retrieved context. This setting introduces retrieval noise and captures a crucial challenge overlooked by prior multimodal personalization benchmarks. Finally, because recall alone cannot capture hallucination and over-abstention, we employ *Calibrated accuracy* (Cal) as our primary metric, equally scoring correct grounding for answerable items and correct abstention for absent-persona items.

Beyond benchmarking, we investigate which post-training regimes best align current omnimodal models with personalization. While previous studies (Oh et al., 2025; 2026) have highlighted the efficacy of RLVR for multimodal personalization in image captioning tasks, we broaden this investigation to omnimodal personalization. Specifically, we rigorously compare supervised fine-tuning (SFT) and reinforcement learning with verifiable rewards (RLVR) to reveal which post-training regime is most suitable, and which specific aspects drive improvements in *omnimodal personalization*. Prior work establishes that SFT is heavily influenced by data quality (Zhou et al., 2023) and scale (Dong et al., 2024), whereas recent RLVR methods rely on carefully specified verifiable reward signals, such as rule-based accuracy and format rewards (Guo et al., 2025). Motivated by this distinction, we conduct SFT on our rigorously curated ground-truth annotation corpora at two different scales (1K and 10K). We contrast this with an RLVR (without SFT warmup) recipe that jointly optimizes perception and retrieval. This RLVR approach utilizes rule-based perceptual verification, alongside LLM-as-a-judge retrieval verification for free-form QA.

Our comparative analysis reveals a clear trade-off. SFT is limited by the difficulty of constructing high-quality, in-domain ground-truth supervision for diverse open-ended scenarios; consequently, broader task coverage does not always translate into gains in Cal. Our RLVR approach alleviates this limitation by using verifiable reward signals to optimize task-level correctness directly, without requiring reference responses for every training instance. However, it introduces a separate trade-off: under our reward design, models tend to favor answering over abstaining. We validate these findings comprehensively across both Qwen2.5-Omni and Gemma4 (Xu et al., 2025b; Gemma Team et al., 2026) architectures.

Our contributions are as follows:

(1) **Omni-Persona Benchmark and PMG Formulation.** We introduce *Omni-Persona*, the first comprehensive evaluation-only benchmark for omnimodal personalization. Built on the *Persona Modality Graph* (PMG), it formalizes contextual grounding over retrieved persona evidence and the integration of raw-form multimodal contexts, spanning 4 task groups and 18 fine-grained tasks across image, text, and vocal audio.

(2) **Addressing Recall Blind Spots with Absent-Persona Evaluation.** While prior personalization benchmarks heavily rely on answerable-only recall, we elevate absent-persona queries to a first-class evaluation dimension. By coupling these unanswerable queries with hard distractors and retrieval noise, we propose a calibrated accuracy metric that jointly assesses correct grounding and appropriate abstention. This balanced approach exposes critical hallucination and over-abstention behaviors often masked by recall-only protocols.

(3) **Diagnostic Analysis of Omnimodal Personalization and Post-Training.** We systematically evaluate closed-source models and open-source models, with post-training analysis conducted on the latter. Our analysis reveals a visual-over-audio grounding asymmetry in open-source models and identifies distinct failure modes across SFT and RLVR. Together, these findings provide a model-specific diagnostic map to guide future research on omnimodal personalization.

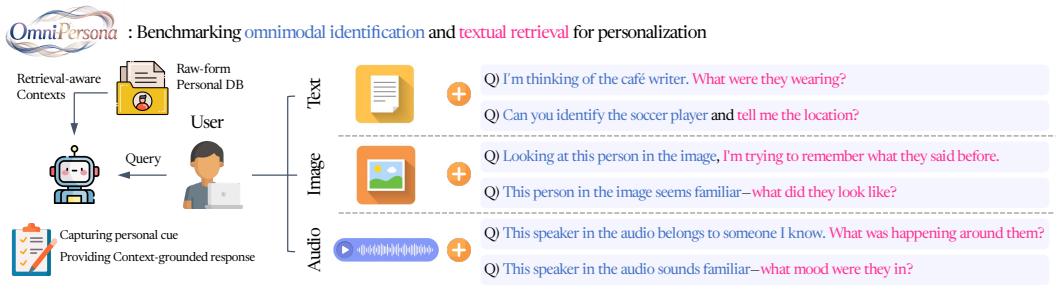


Figure 1: **Formulation of omnimodal personalization in the Omni-Persona benchmark.** As illustrated, a user query consists of a target modality and a text prompt, which we decompose into two main axes: blue elements denote the process of perceptual identification, whereas pink elements highlight the realistic textual retrieval desired by the user. Note that our underlying assumption is that the raw-form contexts are pre-retrieved; we do not evaluate the retrieval process itself.

2 Related Works

Multimodal Personalization Methods. Early personalized vision-language models (VLMs) (Alaluf et al., 2024; Nguyen et al., 2024; An et al., 2025) repurpose off-the-shelf models to recognize user-defined concepts via zero- or few-shot retrieval, yet remain brittle when new concepts must be incorporated dynamically into a user's memory. Post-training-based approaches subsequently emerged to mitigate this rigidity. Hao et al. (2025) first demonstrated that SFT over retrieval-augmented user contexts enables coherent personalized response generation, but its reliance on costly large-scale caption annotations limits practical scalability. To alleviate this annotation burden, Oh et al. (2025; 2026) introduced RLVR-based methods, validating their utility in multi-concept image captioning (Oh et al., 2025) and contextualized scenarios (Oh et al., 2026). Despite this progress, audio has received comparatively limited attention throughout this evolution: visual identity and biographical text (Hong et al., 2026; Mei et al., 2026) have served as the predominant persona modalities, and speaker voice audio has rarely been integrated within a single unified personalization framework.

Evaluation Protocols for Multimodal Personalization. The evaluation protocols accompanying these methods, including those of (Nguyen et al., 2024; Oh et al., 2025; 2026; Alaluf et al., 2024), rely heavily on recall-centric metrics. Such metrics primarily reward surface-level signals, such as name recall and contextual dialogue snippets, that can be directly reinforced during post-training. As a result, generation quality, calibration under absent-persona queries, and the trade-offs introduced by post-training remain largely unmeasured. This limitation is further compounded by existing benchmarks, which often operate under tightly controlled settings and abstract away realistic retrieval noise. To overcome such limitations, our benchmark unveils failure modes that are otherwise hidden beneath recall-only evaluation in multimodal personalization. Specifically, it exposes hallucination and over-abstention behaviors that conventional recall-centric metrics fail to capture. To the best of our knowledge, no prior work has unified interleaved omnimodal contexts, absent-persona evaluation, and a rigorous diagnostic protocol within a single benchmark. Further related work is in Appendix A.

3 Problem Formulation

As illustrated in Figure 1, we formally define *omnimodal personalization*, extending the vision-language personalization paradigm (Nguyen et al., 2024; Hao et al., 2025; Oh et al., 2025; 2026; Alaluf et al., 2024; Mei et al., 2026; Hong et al., 2026) to incorporate audio as a persona modality alongside vision and text.

Formal Definition. We formalize omnimodal personalization as follows. Let a user's personal memory be denoted by $\mathcal{M} = \{(v_i, a_i, t_i)\}_{i=1}^{N}$, where each entry is a triplet comprising a visual identity v_i, an audio sample a_i, and an associated text descriptor t_i. Specifically, v_i may represent a profile image or an appearance snapshot, a_i a 5–15s voice sample or a conversational recording, and t_i dialogue or biographical information. Given a new query comprising a user prompt alongside visual, audio, and textual inputs (v_q, a_q, t_q), the scenario determines which input provides the relevant cue. The relevant entries are retrieved from the memory $\mathcal{M}$ to construct the aggregated top-K contexts $\{\mathcal{C}_i\}_{i=1}^{K}$ where $\mathcal{C}_i = (v_i, a_i, t_i)$. Following this retrieval, the model must, at inference time:

1. *Recognize* which specific entry $\mathcal{C}_j$ within the aggregated contexts $\{\mathcal{C}_i\}_{i=1}^{K}$ corresponds to the relevant query cue, while disregarding the irrelevant query modalities; and
2. *Selectively extract and integrate* the specific details pertinent to the query from the associated text t_j of the identified entry $\mathcal{C}_j$ into a contextually grounded response.

The model must first accurately identify the relevant query cue and then ground its personalized response in the corresponding context. Furthermore, the retrieved contexts $\{\mathcal{C}_i\}_{i=1}^{K}$ arrive in an interleaved format, where the components of each entry $\mathcal{C}_i$ appear in an ordered sequence.

Why Raw Context Matters. Previous textual-memory-based multimodal personalization works (Mei et al., 2026; Hong et al., 2026; Long et al., 2025) convert multimodal signals into compact textual descriptions, creating an information bottleneck that can discard fine-grained identity cues that are difficult to verbalize, such as vocal timbre and facial geometry. We therefore derive personalization directly from *raw omnimodal context*, preserving images and audio as perceptual signals rather than reducing them to text.

Contextual Grounding over Retrieval. We decompose omnimodal personalization into two conceptually distinct sub-problems: (i) *retrieval*, identifying which memories in a user's history are relevant to a given query, and (ii) *contextual grounding*, integrating retrieved multimodal evidence into a faithfully personalized response. This work focuses on the latter, which our evaluation separates from retrieval quality by construction. We refer to the ability to identify query-relevant evidence and faithfully integrate the associated persona information into a response as *grounding expressiveness*. Our goal is to define, measure, and systematically improve this capability. Further details are provided in Appendix C.

Task Composition and Controlled Post-retrieval Evaluation. To evaluate grounding expressiveness under post-retrieval conditions, we provide all models with the same pre-retrieved, person-related contexts. The model must align the query with the relevant visual, textual, or acoustic evidence and ground its free-form response in the associated persona information when such evidence is available. The supplied contexts include visually similar identities and gender-matched voice samples as hard distractors, approximating retrieval noise while keeping the input fixed across models. Omni-Persona therefore evaluates post-retrieval grounding and generation, rather than standalone retrieval quality.

Positioning within Retrieval-Augmented Personalization. Although Omni-Persona does not directly evaluate retrieval, its findings inform end-to-end retrieval-augmented personalization in two respects. First, they expose generator-side limitations that retrieval improvements alone cannot resolve, as correctly retrieved evidence remains ineffective if the model cannot interpret and use it. Second, they indicate whether the generator remains reliable under imperfect retrieval, using relevant evidence when available while avoiding unsupported personalization when it is absent.

4 Benchmarking Omnimodal Perception and Retrieval

We instantiate *Omni-Persona* through the Persona Modality Graph (PMG), where each node is defined as a triplet representing an individual's omnimodal data. In this framework, personalization scenarios are modeled by the interconnections established between these nodes. Building upon this formulation, we propose a novel benchmark that simulates realistic personalization challenges, specifically focusing on modality matching (*i.e.*, graph linkage) within the PMG.

Persona Modality Graph (PMG) and Task Formulation. We formalize omnimodal personalization as a cross-modal routing problem over a PMG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertices $\mathcal{V}$ consist of a query node $\mathcal{Q}$ and retrieved context nodes $\mathcal{C}_1, \ldots, \mathcal{C}_K$, where each node can encompass visual (v), audio (a), and textual (t) modalities, as represented in Figure 2.

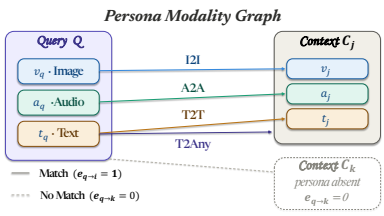


Figure 2: PMG illustration.

The core task is to determine whether a retrieved context contains the target persona and to establish a directed linkage (edge) $e_{q\to j} \in \mathcal{E}$ accordingly. Based on the provided query modality, we categorize the routing process into four primary matching scenarios:

(1) *Image-to-Image (I2I)*: matching visual identity to an image query (*i.e.*, visual identification);

Table 1: **Omni-Persona evaluation scenario groups**. Each group tests whether a model can retrieve the correct persona information from a query in a different modality-matching setting.

Group	Setting	What the Model Must Match	Core Challenge
1: I2I	Image query	Match a person's face to the same visual identity in the persona context	Visual identity recognition
2: A2A	Audio query	Match a person's voice to the same speaker identity in the persona context	Voice identity recognition
3: T2T	Text query	Match a textual cue to the relevant textual persona attribute	Same-modal semantic retrieval
4: T2Any	Text query	Match the meaning of a text cue to relevant text, image, or audio information	Cross-modal semantic grounding

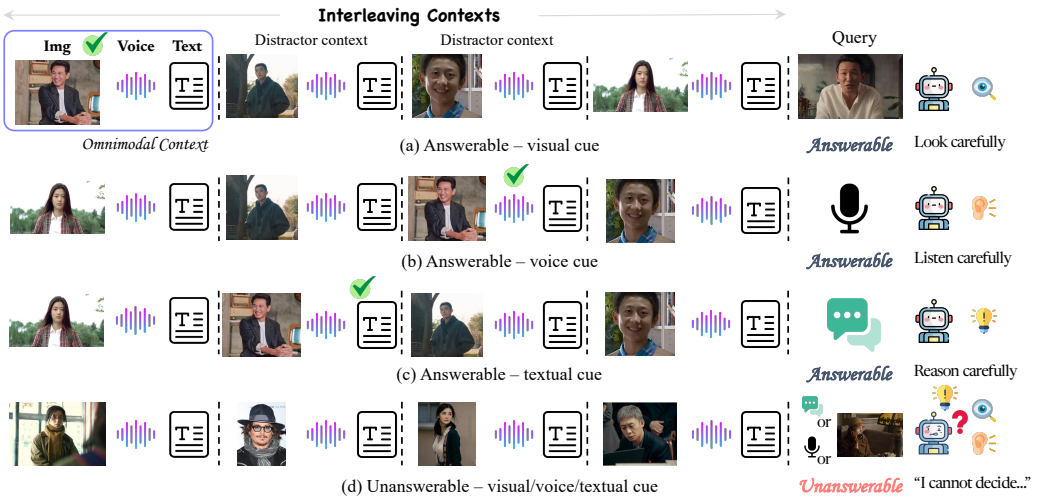


Figure 3: Qualitative examples of context construction, personalization cues, distractors, and unanswerable cases in Omni-Persona.

(2) *Audio-to-Audio (A2A)*: matching voice identity to an audio query (*i.e.*, voice identification);

(3) *Text-to-Text (T2T)*: matching textual attributes to a text query (*i.e.*, same-modal semantic); and

(4) *Text-to-Any (T2Any)*: aligning the semantic meaning of a text query with the cross-modal content of text, image, or audio (*i.e.*, cross-modal semantic).

Crucially, this formulation natively handles absent-persona calibration. If a context $\mathcal{C}_j$ contains the target persona, an active edge is formed ($e_{q\to j} = 1$), making the grounded details in the associated text available to the model. Conversely, if the target persona is absent from all provided contexts, no edge is formed ($e_{q\to j} = 0$), prompting the model to abstain confidently. This unified framework yields the four scenario groups in Table 1 and 18 fine-grained tasks detailed in Appendix J.

Benchmark Design Principles. Designed around natural, human-centric interaction scenarios, Omni-Persona is, to our knowledge, the first personalization benchmark to incorporate *audio* as a full persona modality alongside images and text, while systematically treating *unanswerable* items—where the queried persona is absent from the retrieved context—as a primary evaluation dimension. Unlike prior personalization benchmarks (Nguyen et al., 2024; Oh et al., 2025; 2026), which measure only recall—that is, whether the model retrieves the correct persona when present—Omni-Persona jointly evaluates *grounding recall and abstention*. This setting reflects the dual challenge faced by real-world retrieval systems, where retrieval noise may cause the queried person to be absent from the retrieved context altogether. Furthermore, our cross-modal task design requires models to align audio evidence with visual descriptions, or vice versa, enabling the measurement of modality-specific grounding biases that unimodal tasks cannot reveal. Please see Appendix C for further explanations.

Robustness Under Retrieval Imperfection. Because real retrieval pipelines are noisy, Omni-Persona introduces two classes of perturbation into the evaluation benchmark. The first, *hard distractors*, involves context entries from individuals who share visual or vocal similarities with the target. The

Table 2: **Systematic Omni-Persona benchmark results.** Complementary metrics include 1-FalseAbs and TrueAbs. FA and TA stand for False Abstention and True Abstention, respectively. Green indicates improvement, and Red indicates regression relative to the base model after RLVR.

	Model	Overall Ans	Overall Cal	I2I			A2A			T2T			T2Any			Add. Metrics (×100)		
				Ans	Unans	Cal	Ans	Unans	Cal	Ans	Unans	Cal	Ans	Unans	Cal	1-FA	TA	Avg
	Closed-source Models																	
	Gemini-3.6-Flash (Medium)	**60.1**	**35.7**	**71.3**	**6.0**	**38.5**	**40.0**	**10.9**	**23.9**	69.9	**9.8**	**50.0**	61.4	**10.8**	39.2	99.7	**9.2**	**54.5**
	Gemini-3.6-Flash (High)	51.2	28.8	57.4	1.7	29.4	30.9	4.4	16.2	**71.1**	7.3	**50.0**	49.4	7.7	31.1	**100.0**	4.5	52.2
	Gemini-3.5-Flash-lite	51.4	29.7	65.2	5.2	35.1	12.7	8.0	10.1	62.7	2.4	42.7	**72.3**	6.2	**43.2**	99.5	6.1	52.8
	Open-source Models																	
	MiniCPM-o 4.5 (Think)	**41.7**	**39.3**	22.6	**40.5**	**31.6**	**17.3**	**38.7**	**29.1**	**77.1**	24.4	**59.7**	65.1	**33.8**	**51.4**	87.5	**36.8**	**62.1**
	Phi-4 Multimodal	40.7	33.2	**26.1**	25.9	26.0	13.6	26.3	20.6	69.9	14.6	51.6	**67.5**	27.7	50.0	**91.8**	25.1	58.4
	Interactive-Omni	32.7	29.6	21.7	19.8	20.8	9.1	31.4	21.5	55.4	**29.3**	46.8	56.6	24.6	42.6	84.9	26.2	55.5
Qwen-Omni series	Qwen2.5-Omni-3B	34.0	36.7	27.8	44.0	35.9	18.2	**32.1**	**25.9**	45.8	**53.7**	48.4	51.8	38.5	45.9	74.9	39.6	57.2
Qwen-Omni series	+ SFT (1K)	33.5	36.4	27.9	**44.1**	**36.0**	17.3	**32.1**	25.5	44.6	**53.7**	47.6	52.0	38.3	45.8	75.1	39.4	57.3
Qwen-Omni series	+ SFT (10K)	34.0	**36.9**	27.6	43.8	35.7	18.2	31.4	25.5	45.8	**53.7**	48.4	51.6	**38.7**	**46.1**	74.7	**39.8**	**57.1**
Qwen-Omni series	+ RLVR	**40.9**	34.1	**29.6**	31.9	30.7	**28.2**	21.2	24.3	**54.2**	39.0	**49.2**	**60.2**	21.5	43.2	**84.7**	26.7	55.7
Qwen-Omni series	Δ vs. Base (RLVR)	+6.9	-2.5	+1.7	-12.1	-5.2	+10.0	-10.9	-1.6	+8.4	-14.6	+0.8	+8.4	-16.9	-2.7	+9.7	-12.8	-1.5
Qwen-Omni series	Qwen2.5-Omni-7B	38.6	30.9	40.9	**26.7**	33.8	21.8	**13.1**	17.0	38.6	**48.8**	41.9	57.8	**18.5**	40.5	82.1	**22.6**	52.3
Qwen-Omni series	+ SFT (1K)	39.6	31.5	40.7	26.9	33.7	21.8	**13.1**	17.0	42.2	**48.8**	**44.4**	59.0	**18.5**	41.2	82.3	22.4	52.3
Qwen-Omni series	+ SFT (10K)	42.7	**31.9**	**41.1**	26.5	**34.0**	22.7	**13.1**	**17.4**	48.2	36.6	**44.4**	61.4	16.9	**41.9**	81.9	22.8	52.1
Qwen-Omni series	+ RLVR	**47.8**	28.4	43.5	7.8	25.5	**30.9**	5.8	17.0	**55.4**	12.2	41.1	**68.7**	6.2	41.2	**98.0**	7.2	**52.6**
Qwen-Omni series	Δ vs. Base (RLVR)	+9.2	-2.5	+2.6	-19.0	-8.2	+9.1	-7.3	+0.0	+16.9	-36.6	-0.8	+10.8	-12.3	+0.7	+15.9	-15.3	+0.3
Qwen-Omni series	Qwen3-Omni-30B	51.2	33.7	53.0	21.6	37.2	25.5	8.0	15.8	65.1	19.5	50.0	68.7	13.8	44.6	94.4	14.8	54.6
Gemma4-series	Gemma4-E2B	41.7	**32.1**	45.2	7.8	26.4	15.5	43.8	31.2	56.6	**4.9**	39.5	56.6	**10.8**	36.5	88.7	21.7	**55.2**
Gemma4-series	+ SFT (1K)	41.2	31.9	45.4	**7.7**	26.5	14.5	43.8	30.8	54.2	**4.9**	37.9	57.8	**10.8**	**37.2**	88.9	**21.9**	**55.2**
Gemma4-series	+ SFT (10K)	40.7	31.3	45.0	8.0	26.2	15.5	**44.5**	**31.6**	53.0	**4.9**	37.1	57.8	7.7	35.8	88.5	21.5	55.1
Gemma4-series	+ RLVR	**44.5**	28.1	**48.7**	6.0	**27.3**	**18.2**	18.2	18.2	**59.0**	2.4	**40.3**	**59.0**	6.2	35.8	**96.7**	10.3	53.5
Gemma4-series	Δ vs. Base (RLVR)	+2.8	-4.0	+3.5	-1.7	+0.9	+2.7	-25.5	-13.0	+2.4	-2.4	+0.8	+2.4	-4.6	-0.7	+7.9	-11.4	-1.7
Gemma4-series	Gemma4-E4B	46.8	31.2	50.4	19.0	34.6	15.5	17.5	16.6	**62.7**	0.0	**41.9**	67.5	**7.7**	**41.2**	92.8	14.2	**53.5**
Gemma4-series	+ SFT (1K)	43.5	26.9	41.7	5.2	23.4	13.6	16.1	16.2	57.8	**2.4**	39.5	67.5	4.6	39.9	**96.9**	8.9	52.9
Gemma4-series	+ SFT (10K)	46.8	**33.7**	**55.7**	**25.0**	**40.3**	13.6	**26.3**	**20.6**	61.4	**2.4**	**41.9**	63.9	6.2	38.5	85.7	**19.5**	52.6
Gemma4-series	+ RLVR	**48.6**	30.4	52.2	14.7	33.3	**18.2**	13.1	15.4	**62.7**	0.0	**41.9**	**69.9**	4.6	**41.2**	94.4	10.6	52.5
Gemma4-series	Δ vs. Base (RLVR)	+1.8	-0.8	+1.7	-4.3	-1.3	+2.7	-4.4	-1.2	+0.0	+0.0	+0.0	+2.4	-3.1	+0.0	+1.5	-3.6	-1.0

second, *no-GT retrieval*, omits the ground-truth persona from the context, demanding structured abstention instead of hallucinated matching. This rigorous setup guarantees a comprehensive evaluation across diverse omnimodal tasks. With approximately 50% of the evaluation samples being no-GT, the benchmark systematically probes the model's resistance to hallucination, an essential desideratum when integrating with RAG systems (Xu et al., 2025c; Chhikara et al., 2025; Li et al., 2026).

5 Experiments

Our training study investigates which post-training regime most effectively aligns current omnimodal models for personalization. To this end, we systematically evaluate diverse models on our benchmark, elucidate the underlying behaviors surfaced by our evaluation metrics, and conduct an in-depth model debugging analysis to identify what is fundamentally required to advance omnimodal personalization. Details on data curation and implementation for the experiments are deferred to Appendix G.

5.1 Experimental Setup

Used Models. We evaluate four open-source omnimodal backbones (Gemma4-E2B/E4B-it and Qwen2.5-Omni-3B/7B) under three training regimes: SFT-1K/10K and RLVR. As an upper-bound reference, we include Gemini-3 Flash family (Google DeepMind, 2026), together with four open-source baselines: Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025a), Phi-4-multimodal-Instruct (Microsoft, 2025), Interactive-Omni-8B (Tong et al., 2025), and MiniCPM-o 4.5 (OpenBMB, 2026).

SFT Training Setup. We construct a 10K-sample SFT dataset spanning 12 distinct task types, complemented by a 1K subset for efficient ablation studies. This corpus encompasses foundational grounding, audio-centric scenarios, and absent-persona cases designed to promote calibrated abstention. Crucially, we curate this dataset for broad modality alignment across image, audio, and text, rather than narrow, benchmark-specific optimization. We emphasize that constructing a training corpus for SFT is fundamentally constrained by several factors: (i) the *inherent noise* in synthesizing high-quality ground truth responses for diverse personalization scenarios (Hao et al., 2025; Oh et al., 2025); (ii) the *unpredictability* of the test-time query distribution (Oh et al., 2025; Jiang et al., 2025a; Hong et al., 2026); and (iii) the *scarcity* of large-scale, paired real-world multimodal data (Hu et al., 2025; Liu et al., 2025), necessitating a reliance on synthetic samples that may introduce domain bias. These limitations collectively make SFT alone insufficient for ensuring predictable personalization

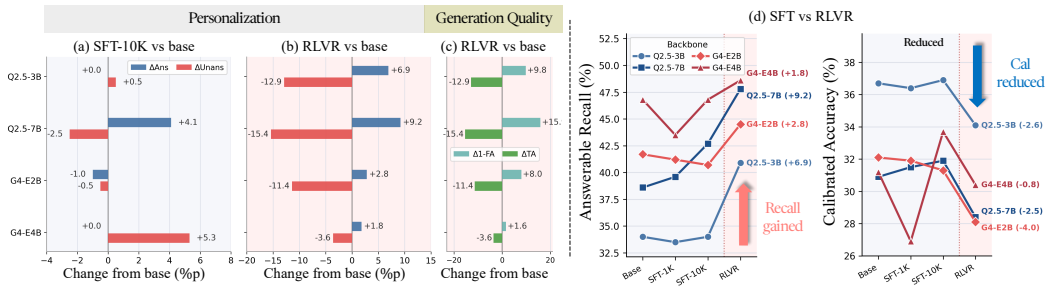


Figure 4: **Performance across post-training regimes**. (a) SFT scaling does not directly translate to gains in open-ended personalization scenarios. (b) RLVR directly improves answerable recall over SFT. (c) RLVR shifts models toward answering, broadly decreasing abstention (1 − FA up, TA down). (d) RLVR exhibits a recall–calibration trade-off, lowering Cal below the base model.

coverage at test time, consistent with recent omnimodal post-training studies showing that RL-based objectives can substantially outperform SFT under matched data and compute budgets (Rouditchenko et al., 2025; Li et al., 2025; Yang et al., 2025). Consequently, we treat SFT as a comparative baseline for RLVR to rigorously analyze how the two post-training regimes shape performance trends.

RLVR Training Setup. To ensure a fair comparison, we perform RLVR on synthetic persona contexts produced by the same pipeline as the SFT corpus. Thus, both regimes use the same type of image–audio–text context triplets and queries. Unlike SFT, however, RLVR does not require reference GT responses. Instead, the model is trained with verifiable feedback that checks whether its response satisfies the intended capability, such as perceptual matching or grounded retrieval. We use three complementary reward components:

(1) *Perception (Rule-based):* Evaluates visual or auditory persona matching by comparing the model's binary decision (*i.e.*, `yes/no`) with the GT label derived from the query–context pairing.

(2) *Location (Rule-based):* Evaluates whether the model correctly identifies the GT context (*i.e.*, MCQA) index, thereby encouraging precise localization within the context.

(3) *Retrieval (LLM-as-a-Judge):* Evaluates whether the response is grounded in retrieved persona evidence. When the target persona is present, the judge verifies factual support against the GT answer and context; when absent, it checks whether the model correctly abstains.

For the retrieval reward, we use `GPT-5.4` (Singh et al., 2025) to generate training queries and GT answers from the benchmark scenarios, rather than full reference responses, while enforcing a disjoint split to prevent evaluation overlap (see Figure S.7 for an example of the RLVR training framework).

Evaluation Metrics. We treat visual and auditory perception jointly: a model that *"sees, hears, and reasons well"* must *consistently match identical entities and distinguish distinct ones across both modalities.* To rigorously evaluate these capabilities, a faithful assessment of omnimodal personalization must simultaneously account for two critical failure modes that standard recall metrics conflate: *hallucinating an identity absent from the context* and *wrongly answering when the identity is absent*. Accordingly, we evaluate models on a fixed pool of 750 queries, balanced between answerable and unanswerable scenarios. Further judge-reliability analyses are detailed in Appendix F.6. Our evaluation relies on the following key metric groups:

- *Calibrated Accuracy*: Our primary metric is the item-weighted accuracy over all queries, $\mathrm{Cal} = \frac{n_a \mathrm{Ans} + n_u \mathrm{Unans}}{n_a + n_u}$, where n_a and n_u are the numbers of answerable and unanswerable items. *Answerable Recall* (Ans) is evaluated for binary correctness using an LLM-as-a-judge (`GPT-5.4-mini` (Singh et al., 2025), while *Unanswerable Recall* (Unans) is measured via abstention-keyword matching (Chen et al., 2024). Cal serves to operationalize a model's true expressiveness and grounding in omnimodal personalization.

- *Anti-Hallucination Accuracy:* To prevent models from artificially inflating Cal through blanket abstention, we rigorously assess their anti-hallucination reliability. Specifically, we report two complementary metrics: the answerable-side complement of False Abstention (1 − FA), and the True Abstention (TA) rate on unanswerable items. These metrics capture generation quality, complementing Cal, and higher values indicate better performance across all reported metrics.

5.2 Main Results

Closed-Source Models. The Cal score exposes a sharp divide that recall alone hides. No closed-source model achieves both strong grounding and reliable abstention. Gemini-3.6-Flash leads in Ans (60.1%) but collapses on Unans, yielding only 35.7% Cal. This low Cal mainly reflects hallucination on absent-persona items, evidenced by its especially low TA score.

Open-Source Baselines. Among open-source models, parameter scale alone does not guarantee high Cal in our benchmark: Q3-30B *underperforms* Q2.5-3B on Cal (33.7% vs. 36.7%), and Phi-4 trails the 3B Qwen baseline, while MiniCPM attains the highest open-source Cal (39.3%) via strong abstention. Among our four RLVR backbones, Q2.5-3B retains the highest Cal score (34.1%). We further discuss the discrepancy between scaling and Cal score in Section 6 (Key Finding 2).

Scaling SFT Fails to Bridge the Distribution Gap in Open-Ended Personalization. Expanding the SFT dataset from 1K to 10K samples fails to yield consistent Cal improvements and even leads to performance degradation (Figure 4a,d) for most backbones. While the augmented corpus was curated to strengthen omnimodal matching across diverse contexts, this broader coverage does not translate into superior benchmark performance. This discrepancy highlights the rigorous complexity of the open-ended personalization required at test time and provides empirical evidence for the inherent limitations of SFT data construction: simply scaling training volume for general modality alignment is insufficient to bridge the distributional gap in complex, personalized reasoning tasks.

RLVR Delivers Recall Gains through Verifiable Supervision. Conversely, RLVR directly and substantially improves answerable recall (Figure 4b), extending the findings of RePIC and CoViP (Oh et al., 2025; 2026) to the omnimodal personalization domain. Recall improves on every backbone, with the largest gains on the Qwen models (Ans $+6.9$ for 3B and $+9.2$ for 7B) and smaller gains on Gemma ($+2.8$ and $+1.8$). However, these gains do not translate into calibration: RLVR fails to induce abstention, and Cal falls below base for every backbone (-0.8 to -4.0); MiniCPM-o 4.5 retains the highest Cal and anti-hallucination average ($1 - \text{FA}$ and TA) among the evaluated models, as shown in Table 2. The detailed analyses of the recall–Cal trade-off are in Appendix F, including paired significance tests and confidence intervals for this trade-off in Appendix F.2.

The Answer-Inducing Trade-off of RLVR. Notably, our anti-hallucination accuracies reveal that *RLVR shifts the backbones toward answering*. As shown in Figure 4 (c), all four backbones raise $1 - \text{FA}$ and drop TA below the baseline (most pronounced for the Qwen backbones), indicating fewer false abstentions on answerable cases but a collapse of abstention on unanswerable ones. We interpret this as a *reward-induced answering bias*: since our outcome rewards directly target grounding and reasoning, attempting a grounded answer becomes the dominant signal even when the target persona is absent, so abstention is rarely reinforced. This effect appears across model families and underlies the Cal decline observed across the backbones. More sophisticated reward shaping or cold-start SFT (Guo et al., 2025) may restore calibrated abstention, which we leave to future work.

6 Toward Omnimodal Personalization: In-Depth Analyses

Key Finding 1: Audio Grounding Lags Behind Visual Grounding

Open-source models consistently underperform on audio grounding relative to visual grounding, exhibiting an approximately 5–35 percentage-point gap in answerable recall between A2A and I2I tasks. Although RLVR improves recall in the audio modality, Cal reveals a key limitation: models trained with naive RLVR are biased away from abstention.

When evaluated separately on visual and auditory tasks, open-source models exhibit substantially stronger visual than auditory perception, resulting in a pronounced performance gap between I2I and A2A tasks. Although the closed-source Gemini models achieve the strongest A2A performance, reaching up to 40.0%, they still exhibit a clear cross-modal perceptual gap. By providing explicit perceptual supervision through rule-based rewards, RLVR improves A2A answerable recall across all four backbones. The gains are substantial for the Qwen models, reaching $+10.0$ and $+9.1$ points for the 3B and 7B variants, respectively, while G4-E2B and G4-E4B improve by $+2.7$ and $+2.5$ points, respectively. Notably, the A2A recall gains for Qwen exceed those observed on I2I, demonstrating that RLVR can effectively elicit audio-perception capabilities. However, Cal reveals a key limitation of RLVR alone: without an SFT warm-up, the models become biased against abstention, leaving the

trade-off between recall and calibrated abstention only partially resolved. Consequently, RLVR yields lower Cal scores than the base models on I2I and A2A tasks for most backbones.

Key Finding 2: Recall and Model Scale Miss the Calibration Axis

Answerable recall and model scale are useful but incomplete diagnostics for personalization. A model with strong recall may still hallucinate when the target persona is absent (low TA), or over-abstain even when the relevant persona is present (low $1 - \text{FA}$). Likewise, Cal exposes this missing axis by jointly measuring grounding on answerable cases and abstention on unanswerable cases.

A closer look at Q3-30B illustrates this point. While the 30B model achieves higher answerable recall on text-grounded tasks, it performs worse on unanswerable cases than smaller Q2.5 variants, resulting in lower overall Cal and the lowest TA score among the Qwen models, indicating greater hallucination. Similarly, Gemini-3.6-Flash achieves the strongest answerable recall (Ans 60.1%) yet pairs it with a low TA (9.2%), capping its calibrated accuracy at 35.7%. Increasing its thinking budget further degrades performance, suggesting that additional reasoning does not necessarily improve personalization. Recall-only comparisons often obscure these discrepancies, whereas Cal and anti-hallucination accuracy effectively expose them.

Key Finding 3: SFT and RLVR Reveal Different Post-training Trade-offs

SFT and RLVR fail in different ways. SFT is limited by the difficulty of constructing annotated GT supervision at scale, making it challenging to consistently improve evaluation performance. In contrast, RLVR can directly target perceptual grounding and textual retrieval through verifiable reward signals, yielding answerable-recall gains on every backbone, largest for the Qwen models. However, it fails to induce calibrated abstention: the backbones shift toward answering ($1 - \text{FA}$ up, TA down), and Cal fails to improve for any backbone.

SFT is limited by the difficulty of constructing high-quality supervision data at scale. Scaling the data from 1K to 10K does not reliably improve Cal, suggesting that a broader data mixture does not directly translate into better benchmark performance in open-ended personalization scenarios. This can be interpreted as a training-evaluation mismatch (Chu et al., 2025). Conversely, RLVR avoids this bottleneck by replacing reference-response imitation with outcome-level supervision. Perception reward reinforces binary visual/audio matching, location reward reinforces correct localization within a context, while retrieval reward reinforces responses grounded in textual evidence. This enables RLVR to substantially improve answerable recall, especially for the Qwen backbones, while bypassing the need for GT responses. However, because these rewards directly target grounding and reasoning, the models drift toward answering (higher $1 - \text{FA}$, lower TA), and calibrated abstention is not induced: Cal fails to improve for any backbone. Overall, these analyses reaffirm the potential of RLVR-based frameworks while highlighting the need for careful post-training to mitigate the inherent trade-off between recall and calibrated abstention.

7 Conclusion

We introduce *Omni-Persona*, the first comprehensive benchmark for omnimodal personalization. By formalizing contextual grounding over retrieved persona evidence and the integration of raw-form omnimodal context, it enables systematic analysis of personalized expressiveness, treats audio as a persona modality alongside images and text, and adopts absent-persona as a core evaluation dimension. We propose calibrated accuracy and anti-hallucination accuracy, showing that recall-only metrics can obscure hallucination under retrieval noise. Extensive benchmarking shows that, on omnimodal tasks evaluated independently, open-source models process visual cues substantially more reliably than audio cues; scaling SFT does not reliably improve performance, reflecting the difficulty of constructing data aligned with open-ended personalization; and RLVR substantially improves recall but fails to induce calibrated abstention, leaving calibrated accuracy below base. Overall, Omni-Persona provides a realistic diagnostic framework for analyzing the strengths and failure modes of omnimodal personalization.

Limitations. Our benchmark uses synthetic audio and text with rigorous model-based filtering, leaving large-scale human verification of the generated text as a future refinement. Free-form LLM-as-a-judge evaluation may introduce residual bias; see Appendix H for further discussion.

AI USE STATEMENT

Large language models (LLMs) are both the object of study and a research tool in this work, and we disclose their use as follows. **Synthetic data generation.** All textual elements of the Omni-Persona benchmark were synthesized with `GPT-5.5`: extracting explicit personal attributes from the source dialogues, imputing missing traits, restructuring them into concise persona profiles, and generating query prompts with ground-truth answers for answerable items and structured abstention targets for unanswerable items. The training queries and their ground-truth answers used for the retrieval reward were generated with `GPT-5.4` (Section 5.1). Synthetic speech was produced with `chatterbox`, and the synthetic images used in SFT and RLVR training depict no identifiable real individuals. The full prompt templates for every generation stage are released in Appendix K.

LLM-based evaluation. Model responses were scored by an LLM judge, `GPT-5.4-mini`; `GPT-5.6-sol` served as an independent second judge for the cross-judge agreement analysis in Appendix F.5, and judge reliability is quantified in Appendix F.6.

Literature discovery. LLM assistants were used to support literature search and identify potentially relevant prior work. All cited works were independently reviewed and verified by the authors.

Writing. The authors used LLM assistants to improve the readability and clarity of author-written text, refine wording, and assist with reference formatting.

Methodology and evaluation assistance. LLM assistants were used to provide feedback on the design and refinement of the evaluation protocol and proposed metrics. Final design decisions were made and validated by the authors.

Data processing. LLM assistants were used for dataset cleaning, reformatting, and translation where applicable.

Code assistance. LLM assistants were used to assist with the implementation and debugging of experimental and evaluation code. All AI-assisted code was reviewed and tested by the authors.

Not used. LLMs were not used to formulate the research question, propose or refine hypotheses, develop the theoretical framing or mathematical claims, or independently determine the study's conclusions.

The authors reviewed all AI-assisted output. Benchmark items were verified through the sanity-checking and integrity procedures described in Appendix D, and the authors take full responsibility for the contents of this submission.

ETHICS STATEMENT

This work introduces a benchmark for personalization, which by construction involves persona attributes. We did not directly collect new data from real, identifiable individuals. All persona profiles, dialogues, and training images are synthetic, while the evaluation images and real-world audio are drawn from existing publicly available sources and used under their original licenses. Because personalization systems can amplify harms when they misattribute an attribute to the wrong person, our evaluation explicitly measures absent-persona hallucination and rewards appropriate abstention rather than recall alone. Broader risks and misuse considerations are discussed in Appendix H.

REPRODUCIBILITY STATEMENT

We report the complete experimental setup needed to reproduce our results. Training and inference configurations for both SFT and RLVR, including hyperparameters, are given in Appendix G; benchmark construction, task taxonomy, and evaluation-scenario configurations are described in Appendices D, E, and J; and the prompt templates used for data generation and for the LLM judge are reproduced verbatim in Appendix K. Judge reliability, including cross-judge agreement, is reported in Appendix F.6. The benchmark and evaluation code will be released publicly upon acceptance.

REFERENCES

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pp. 73–91. Springer, 2024.

Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, Bocheng Zou, Chaoqun Yang, and Wentao Zhang. UniCTokens: Boosting personalized understanding and generation via unified concept tokens. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.14671.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 177–198, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. RAP: Retrieval-augmented personalization for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2410.13360.

Rongpei Hong, Jian Lang, Ting Zhong, Yong Wang, and Fan Zhou. Tameing long contexts in personalization: Towards training-free and state-aware mllm personalized assistant. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 452–463, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

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

Jesin James, Li Tian, and Catherine Inez Watson. An open source emotional speech corpus for human robot interaction applications. In *Interspeech*, pp. 2768–2772, 2018.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025a.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. PersonaMem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025b.

Jaeik Kim, Woojin Kim, Woohyeon Park, and Jaeyoung Do. MMPB: It's time for multi-modal personalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. arXiv:2509.22820.

Tae Soo Kim, Yoonjoo Lee, Yoonah Park, Jiho Kim, Young-Ho Kim, and Juho Kim. CUPID: Evaluating personalized and contextualized alignment of LLMs from interactions. In *Conference on Language Modeling (COLM)*, 2025b. arXiv:2508.01674.

Doyeop Kwak, Jaemin Jung, Kihyun Nam, Youngjoon Jang, Jee-Weon Jung, Shinji Watanabe, and Joon Son Chung. Voxmm: Rich transcription of conversations in the wild. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12551–12555. IEEE, 2024.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Che Liu, Yingji Zhang, Dong Zhang, Weijie Zhang, Chenggong Gong, Yu Lu, Shilin Zhou, Ziliang Gan, Ziao Wang, Haipang Wu, et al. Nexus-o: An omni-perceptive and-interactive model for language, audio, and vision. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10787–10796, 2025.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Xinyu Hou, Margaret Li, and Bill Byrne. According to me: Long-term personalized referential memory qa. *arXiv preprint arXiv:2603.01990*, 2026.

Microsoft. Phi-4 Mini Technical Report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'LLaVA: Your personalized language and vision assistant. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.09400.

Yeongtak Oh, Dohyun Chung, Juhyeon Shin, Sangha Park, Johan Barthelemy, Jisoo Mok, and Sungroh Yoon. RePIC: Reinforced post-training for personalizing multi-modal language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.18369.

Yeongtak Oh, Sangwon Yu, Junsung Park, Han Cheol Moon, Jisoo Mok, and Sungroh Yoon. Contextualized visual personalization in vision-language models. *arXiv preprint arXiv:2602.03454*, 2026.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenBMB. MiniCPM-o 4.5: A gemini 2.5 flash level mllm for vision, speech, and full-duplex multimodal live streaming on your phone. https://github.com/OpenBMB/MiniCPM-o, 2026.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? *arXiv preprint arXiv:2505.09439*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025a.

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

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025c.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Yuze Zhao, Jintao Qin, Haibo Niu, Wentao Cheng, Bingning Tang, Xingjun Yang, Hailin Zou, Yanping Li, Shen Liu, et al. SWIFT: A scalable lightweight infrastructure for fine-tuning. *arXiv preprint arXiv:2408.05517*, 2024.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

APPENDIX CONTENTS

A Further Related Works **15**

B Preliminaries **15**

C Benchmark Design Rationale and Integrity **16**

D Details on Evaluation Dataset and Metrics **17**

E Evaluation Dataset Configurations **17**

F Additional Analysis **19**

F.1 Supervision-Induced Recall–Abstention Trade-off 19

F.2 Statistical Significance of the Recall–Calibration Trade-off 20

F.3 Reward Asymmetry on the Recall–Abstention Trade-off 20

F.4 Reward-Component Ablation 22

F.5 Additional Cross-Judge Robustness 22

F.6 Alignment of Evaluation Metrics 23

G Additional Experimental Configurations **25**

G.1 Details on SFT Implementations 25

G.2 Details on RLVR Implementations 25

H Limitations and Broader Impacts **29**

I Further Analysis on Lexical Metrics **31**

J Detailed Task Taxonomy and Granularities **32**

K Used Templates for Dataset Construction **34**

Table S.1: Comparison of previous personalization benchmarks against real-world scenarios.

Dataset	Context Nature	Acquisition	Scenarios	Tasks	Modalities	Cost	Evaluation
CoViP (Oh et al., 2026)	Homogeneous	Synthetic only	Text dialogue	3	I, T	Low (API)	Standard (MCQA)
Omni-Persona (Ours)	Heterogeneous (distractors, no GT)	Mixed (Real & Syn.)	Dialogue, biography	18	I, T, A	Moderate	Multi-hop (Free form)
Real-world	Noisy & incomplete	Manual (Privacy risks)	Emails, logs, histories, memory update	$\gg 18$	I, T, A, V	Prohibitive	Complex (Agentic)

A Further Related Works

LLM Personalization. Recent text-only personalization benchmarks evaluate how well LLMs profile and respond to users. For instance, Jiang et al. (2025a) focus on dynamic profiling, Kim et al. (2025b) use real user interaction logs for alignment, and Jiang et al. (2025b) infer latent traits from conversational history. While foundational, these works remain confined to the textual modality, lacking the detailed analyses for multimodal grounding necessary for omnimodal personal assistants.

Comparison with CoViP. Table S.1 contrasts Omni-Persona with CoViP (Oh et al., 2026), its closest predecessor in the multimodal personalization benchmark space. Relative to CoViP, Omni-Persona: (i) incorporates audio as a persona modality alongside image and text; (ii) incorporates ground-truth-absent queries under multi-distractor contexts as a core evaluation axis; and (iii) broadens the task scope from image captioning to open-ended QA and cross-modal identity matching. Critically, CoViP's answerable-only multi-choice question answering (MCQA) protocol cannot surface model overconfidence or abstention collapse. By exposing these failure modes through a dual-axis design and extending absent-persona evaluation to audio and cross-modal scenarios, Omni-Persona offers a comprehensive diagnostic framework for omnimodal personalization.

B Preliminaries

Omnimodal and Multimodal Foundation Models. Omnimodal models extend MLLMs toward any-to-any generation across text, vision, and audio. Proprietary systems such as GPT-4o (OpenAI et al., 2024) and Gemini-3 (Google DeepMind, 2026) have shown strong cross-modal understanding and real-time dialogue capabilities. Among open-source omnimodal models, Qwen2.5-Omni (Xu et al., 2025b) adopts a Thinker–Talker architecture: the Thinker performs multimodal understanding and textual reasoning, while the Talker generates speech responses. MiniCPM-o (OpenBMB, 2026) unifies vision, speech, and language within an end-to-end framework, employing a time-division multiplexing mechanism that interleaves parallel omnimodal streams into periodic time slices, thereby supporting real-time bidirectional speech interaction. Phi-4-Multimodal (Microsoft, 2025) integrates text, vision, and speech/audio into a single 5.6B-parameter model through a Mixture-of-LoRAs design, in which modality-specific LoRA adapters and routers are attached to a frozen Phi-4-Mini backbone, enabling competitive multimodal reasoning without cross-modal interference while fully preserving the base language capability. Interactive-Omni (Tong et al., 2025) combines omni-modal pre-training with interaction-oriented post-training to support text-and-speech generation in multi-turn audio-visual dialogue with long-term contextual memory.

We further use Gemma4 E-series models (E2B and E4B) (Gemma Team et al., 2026) as our main analysis models. Gemma4 is not a native omnimodal model, but an MLLM designed for constrained DRAM and flash-memory environments through Per-Layer Embeddings and Grouped-Query Attention, which reduce KV-cache pressure. This makes it a useful on-device-oriented counterpart to larger server-class omnimodal systems.

Table S.2: Modality ablation measured by strict answerable recall across task groups. Parentheses in the *Both dropped* rows indicate changes relative to the corresponding full-context results.

Condition	I2I	A2A	T2T	T2Any	$1 - \text{FA}$
Qwen2.5-Omni-3B					
Full context	28	19	45	52	75
Image shuffled	12	13	33	49	68
Audio shuffled	22	11	41	57	80
Image dropped	8	14	27	41	44
Audio dropped	50	11	52	65	80
Both dropped	**3** (-25)	**4** (-15)	**50** ($+5$)	**62** ($+10$)	**26** (-49)
Gemma4-E2B					
Full context	44	15	57	57	89
Image shuffled	12	12	49	49	81
Audio shuffled	39	9	54	63	92
Image dropped	4	14	48	46	72
Audio dropped	57	10	63	64	89
Both dropped	**3** (-41)	**2** (-13)	**66** ($+9$)	**72** ($+15$)	**47** (-42)

C Benchmark Design Rationale and Integrity

Construction and Data Provenance. The benchmark comprises approximately 750 instances spanning 18 fine-grained tasks. Training images are synthetic, while training and evaluation audio are drawn from disjoint source splits to prevent contamination. Each instance is manually verified after constructing person-consistent query–context histories and introducing modality-appropriate hard distractors. To prevent text-only shortcuts in the visual- and voice-identity tasks, we ensure that all four persona descriptions state the queried subfield, each with a distinct, gender-consistent value. We further use `GPT-5.5` to rewrite the descriptions in a uniform natural style while preserving the gold semantics. This construction removes cues based on attribute presence, template structure, or writing style, requiring the model to identify the relevant persona from the query image or audio.

Role of Absent-persona Evaluation. Approximately half of the instances omit the target persona from the context, modeling retrieval failures in which relevant evidence is unavailable. These cases test whether the model abstains rather than producing an unsupported answer from a distractor, with failure patterns evaluated across visual, acoustic, and textual evidence. Answerable recall alone does not capture this behavior or distinguish appropriate abstention from systematic over-abstention. We therefore jointly report Cal, $1 - \text{FA}$, and TA to evaluate grounding and abstention behavior.

Mitigating Textual Shortcuts in Contextual Memory. We further enforce within-description consistency in names, gender references, and pronouns. Topic- and quote-related candidates are also expressed in the same neutral third-person register, removing first-person or shared-memory cues that could distinguish the gold context. These controls minimize residual lexical and stylistic asymmetries between the gold persona and distractors.

Modality Corruption and Grounding-Dependence Ablation. In Table S.2, we empirically test whether performance remains dependent on visual and auditory evidence after controlling for textual shortcuts. The textual inputs are held fixed, while the selected perceptual modality is either shuffled or removed. Under *shuffling*, the query and context inputs are replaced with those associated with different personas; under *dropping*, the selected modality is removed entirely. We evaluate Qwen2.5-Omni-3B and Gemma4-E2B, reporting strict answerable recall by task group together with $1 - \text{FA}$.

The perceptual tasks show clear modality-specific dependence. Removing both image and audio reduces I2I recall from 28 to 3 for Qwen2.5-Omni-3B and from 44 to 3 for Gemma4-E2B; A2A recall similarly decreases from 19 to 4 and from 15 to 2, respectively. The targeted ablations support the same conclusion: image corruption primarily degrades I2I, whereas audio corruption degrades A2A. Because the textual inputs remain unchanged, these losses show that the perceptual tasks cannot be recovered from text alone.

Removing both modalities also reduces $1 - \text{FA}$ from 75 to 26 for Qwen2.5-Omni-3B and from 89 to 47 for Gemma4-E2B, indicating increased false abstention when perceptual evidence is unavailable.

In contrast, T2T and T2Any improve for both models after removing perceptual inputs, suggesting that irrelevant image and audio inputs can interfere with text-grounded retrieval tasks. Overall, the ablation confirms that *performance depends on the modality relevant to each task rather than on a shared textual shortcut*.

D Details on Evaluation Dataset and Metrics

Task Overview. Omni-Persona instantiates the four PMG scenarios into 18 fine-grained sub-tasks organized into four distinct groups, each covering a different combination of query and context modalities. The groups are structured as follows:

- *Group 1 (I2I; 5 tasks):* A face-image query is matched against the persona's stored face images to ground visual identity, and the model then retrieves one of five persona attributes: *Biography* (text), *Dialogue* (voice), *Appearance* (image), *Emotion* (voice), or *Environment* (voice). All sub-tasks except *Appearance* require a visual-to-text or visual-to-audio bridge.
- *Group 2 (A2A; 5 tasks):* A voice query is matched against stored speaker samples to identify the persona, then sweeps the same five retrieval targets as Group 1. Sub-task 2-c (*Appearance*) further requires an audio-to-visual bridge, a capability absent in purely unimodal retrieval.
- *Group 3 (T2T; 4 tasks):* Identity is resolved through semantic matching between a textual query and the persona's stored textual profile, covering four retrieval targets: *Biography* (text), *Appearance* (image), *Emotion* (text), and *Environment* (text). Sub-task 3-b additionally crosses modality by retrieving a visual appearance description from a purely textual match. The *Dialogue* target is intentionally omitted, as text-to-text conversational matching risks collapsing into shallow keyword overlap.
- *Group 4 (T2Any; 4 tasks):* This group forms the benchmark's most demanding regime: a textual description of conversational content must be matched to the persona whose stored conversational audio semantically corresponds to it, without any explicit speaker cues. Retrieval targets span *Biography* (text), *Appearance* (image), *Emotion* (voice), and *Environment* (text). Sub-task 4-b requires a three-hop cross-modal path from the text query through audio-based identity matching to visual appearance retrieval, analogous to 2-c.

Across all four groups, every sub-task is paired with a no-GT (absent-persona) variant that requires the model to perform structured abstention, as summarized in Table S.14. The elucidations of additional metrics used are in Table S.3.

Benchmark Complexity and Design Principles. Our proposed benchmark is systematically more challenging in several respects. The evaluation distractors are deliberately curated to be highly similar to the target, sharing specific vocal characteristics or visual resemblances, thereby requiring fine-grained cross-modal discrimination. Moreover, each evaluation context contains four interleaved image-audio-text entries, exposing models to dense and heterogeneous multimodal signals. Detailed per-scenario prompt templates and query-construction procedures are provided in Appendix J. Crucially, we introduce a deliberate distribution shift between training and evaluation to assess true generalization: training uses three context entries paired with synthesized CoViP images, whereas evaluation uses four context entries paired exclusively with real images.

E Evaluation Dataset Configurations

Construction of Image–Text Pairs. To formulate diagnostic tasks for omnimodal personalization, it is essential to have multiple images of the same individual to facilitate cross-context identity reasoning. To this end, we sample person-related contexts from the evaluation split of CoViP, ensuring each individual appears as the query subject across diverse scenarios. This process yields 250 context–query pairs, each comprising a query image and four interleaved context entries, where each entry consists of an identity image paired with descriptive text.

To systematically construct the textual elements of the Omni-Persona benchmark, we refine the pre-generated dialogues through a multi-stage pipeline. First, a model extracts explicit personal

Table S.3: Summary of evaluation metrics used in the Omni-Persona benchmark. All scores are reported with a higher-is-better polarity; MeanLen is descriptive and is excluded from Avg.

Metric	Target Subset	Description
Overall	All	Item-weighted accuracy across both answerable and unanswerable scenarios.
Ans	Answerable	LLM-as-a-judge correctness rate.
Unans	Unanswerable	Successful abstention rate via keyword matching.
ROUGE-L	Answerable	Average LCS-based recall (sequence-sensitive lexical overlap).
Tok-F1	Answerable	Average token-level F1 (bag-of-words lexical overlap).
BERTScore	Answerable	Average BERTScore F1 (semantic embedding similarity).
Abs-F1	All	Harmonic mean of abstention precision and recall; penalizes over-abstention.
$1 - \mathrm{FA}$	Answerable	Complement of FalseAbs: fraction of answerable queries the model does *not* answer with an abstention keyword.
TA	Unanswerable	TrueAbs: fraction of unanswerable queries on which the model correctly abstains.
Avg	Composite	Simple mean of the nine higher-is-better scores above (Overall, Ans, Unans, Abs-F1, ROUGE-L, Tok-F1, BERTScore, $1 - \mathrm{FA}$, TA).
MeanLen	All	Mean response length in whitespace-tokenized words (descriptive; not part of Avg).

Evaluation Metrics

Dual-Axis Recall

We measure basic task completion via a dual-axis accuracy protocol that decouples successful answering from appropriate abstention.

- **Ans:** Proportion of the 391 answerable queries judged correct by an LLM-as-a-judge.
- **Unans:** Proportion of the 359 unanswerable queries for which the model triggers an abstention keyword (*e.g.*, "I can't decide").
- **Overall:** Combined accuracy across all 750 items.

Generation Quality (Answerable Only)

To assess the lexical and semantic fidelity of generated responses against the ground truth, we apply three established metrics on the answerable subset.

- **ROUGE-L Recall:** Structural overlap via the Longest Common Subsequence (LCS); strict on token order, penalizing valid paraphrasing.
- **Token F1:** Harmonic mean of token-level precision and recall under a bag-of-words representation, order-independent.
- **BERTScore F1:** Cosine similarity in the BERT embedding space, rewarding semantic equivalence beyond exact lexical matching.

Calibrated Abstention (Abs-F1). Unanswerable recall alone is gameable by always abstaining. **Abs-F1** closes this loophole by treating unanswerable queries as the positive class: TP are correct abstentions, FP are erroneous abstentions on answerable items, and Abs-F1 is the harmonic mean of the resulting precision and recall. A high Abs-F1 thus reflects calibrated confidence, rejecting unanswerable queries without sacrificing valid answers.

Anti-Hallucination Guardrails ($1 - \mathrm{FA}$, TA). We report the two raw abstention rates jointly so that the *direction* of a calibration failure is visible: $1-\mathrm{FA}$ drops when the model *over-abstains* on answerable items (*e.g.*, RL abstention-shortcut hacking), while TA drops when the model *hallucinates* a grounded answer on unanswerable items. A model that maximizes one through blanket behavior is therefore automatically penalized by the other.

Composite (Avg) and Length (MeanLen). Avg is the simple mean of the nine higher-is-better scores listed above; it flags single-axis policies (e.g., universal abstention) that boost one metric while collapsing others. **MeanLen** is the mean response length in whitespace-tokenized words and is reported descriptively only, since it has neither a $[0, 1]$ range nor a clear quality polarity.



Figure S.1: **Qualitative image examples.** Sample images utilized during the training and evaluation phases. Note that synthetic images are strictly employed for SFT and RL training, whereas real images are reserved exclusively for the Omni-Persona benchmark.

attributes from the raw dialogue. It then plausibly imputes any missing traits and restructures the extracted information into concise personal profiles (*e.g.*, in biography or dialogue format), strictly limited to 1–2 sentences per concept. By bifurcating scenarios into answerable and unanswerable (no-GT) cases, we synthesize targeted query prompts accordingly: for answerable cases, a GT answer is generated based on the target context; for unanswerable cases, the target is formatted to elicit structured abstention. All synthesis procedures are conducted using the proprietary frontier model `GPT-5.5`.

Audio Modality Construction. The audio modality consists of a balanced mixture of synthetic voice samples and real-world recordings, covering a total of 450 distinct speakers.

Synthetic Data. For synthetic samples, we use `chatterbox`[1] to generate high-fidelity audio (*24-bit, 24 kHz*). For each speaker, we construct pairs of distinct clips that share the same voice identity while varying emotion, conversational setting, and metadata such as age and accent.

Real-world Data. Real-world conversational audio is curated from diverse corpora, including VoxMM (Kwak et al., 2024), MELD (Poria et al., 2019), JL-corpus (James et al., 2018), and RAVDESS (Livingstone & Russo, 2018). We extract 4–15 second segments and ensure expressive variation across two roles:

- **Reference Audio:** Uses relatively neutral or mild emotions, such as *neutral*, *calm*, and *happy*.
- **Emotional Utterance:** Uses more expressive and distinct emotions, such as *angry*, *sad*, *fearful*, *disgust*, and *surprised*.

Context–Query Pair Construction. To construct challenging audio distractors for visually grounded pairs, we use `wav2vec2` (Baevski et al., 2020) for automated gender detection. Each target voice is then paired with a *gender-aligned distractor* voice from either the synthetic or real-world subset. This setup prevents the model from relying on coarse gender cues and instead requires more fine-grained recognition of speaker-specific vocal characteristics.

In summary, the final Omni-Persona evaluation benchmark comprises approximately 750 items spanning 18 tasks across 4 scenario groups. This benchmark serves strictly as a held-out test set.

F Additional Analysis

F.1 Supervision-Induced Recall–Abstention Trade-off

To assess the effect of labeled supervision composition during RLVR, we conduct all experiments using `Qwen2.5-Omni-3B` while varying the answerable-to-unanswerable training ratio from 1:0 to 1:1. As shown in Figures S.2 and S.3, the guarded evaluation reveals a clear recall–abstention trade-off as the proportion of unanswerable examples increases. Answerable recall declines from 43.0 at a

[1] https://github.com/resemble-ai/chatterbox

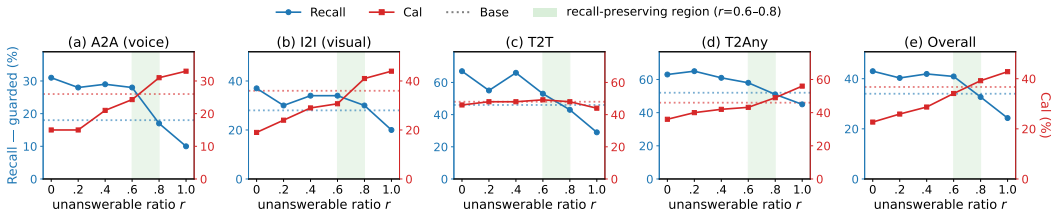


Figure S.2: **Ablation of the answerable-to-unanswerable training ratio.** Varying the proportion of unanswerable examples reveals the trade-off between strict recall and calibration relative to the base model. Here, $r = 0$ denotes training without unanswerable examples, consistent with the settings used in RePIC and CoViP. The model used is `Qwen2.5-Omni-3B`.

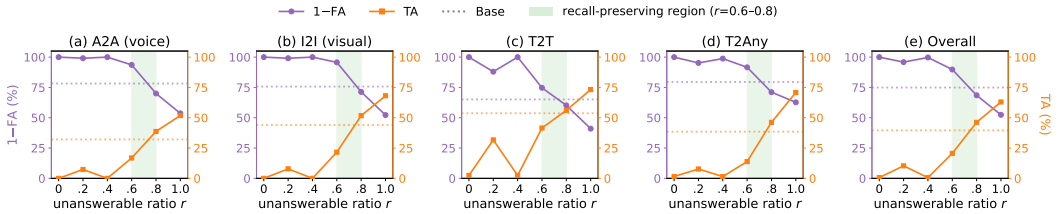


Figure S.3: **Trade-off between $1 - \mathrm{FA}$ and TA.** Varying the answerable-to-unanswerable ratio during post-training reveals a trade-off between correctly abstaining on unanswerable examples (high TA) and avoiding false abstention on answerable examples (high $1 - \mathrm{FA}$). Here, $r = 0$ denotes training without unanswerable examples, consistent with the settings of RePIC and CoViP. The model used is `Qwen2.5-Omni-3B`.

1:0 ratio to 24.3 at approximately 1:1, while true-abstention accuracy rises from 0.6 to 63.0. This pattern suggests that abstention is primarily induced by explicit abstain-labeled supervision rather than emerging from grounding rewards or prompt-level instructions alone. In particular, the 1:0 setting achieves the highest recall yet almost never abstains, demonstrating that strong grounding supervision is insufficient to produce calibrated abstention without explicit unanswerable examples.

A balanced 1:1 ratio achieves the highest aggregate calibration score but reduces recall below the base model by 9.7 points, showing that calibration is obtained at a substantial retrieval cost. In contrast, asymmetric supervision offers a more favorable operating point: the 1:0.6 setting restores recall to near-base performance while retaining substantial abstention accuracy, and ratios around 0.6–0.8 form a recall-preserving region on the empirical frontier. These results suggest that a relatively small but non-negligible amount of unanswerable supervision can mitigate the recall–abstention trade-off more effectively than fully balanced training.

F.2 Statistical Significance of the Recall–Calibration Trade-off

We test whether the observed changes after RLVR reflect a consistent change in model behavior rather than random variation. Because the base and RLVR checkpoints are evaluated on the same items, we use McNemar's exact test for paired comparisons and report 95% confidence intervals for each change. We conduct this analysis on the two `Qwen2.5-Omni` backbones for which paired item-level results are available.

As shown in Table S.4, RLVR significantly improves Ans and $1 - \mathrm{FA}$ while significantly reducing TA on both backbones. This indicates that RLVR makes the model more likely to answer, improving recall on answerable queries but reducing appropriate abstention on unanswerable ones. Because these effects oppose each other, the resulting decrease in Cal is modest and not statistically significant for either backbone. Overall, the results support a clear shift toward answering, but not an improvement in calibrated behavior.

F.3 Reward Asymmetry on the Recall–Abstention Trade-off

We examine whether the recall–abstention trade-off can be adjusted through asymmetric reward assignments. All experiments use `Qwen2.5-Omni-3B` with a fixed answerable-to-unanswerable train-

Table S.4: Paired comparison of base and RLVR checkpoints on the `Qwen2.5-Omni` backbones. Changes and 95% confidence intervals are reported in percentage points. McNemar's exact test is used for significance testing.

Metric	Base	RLVR	Δ	95% CI	p
Qwen2.5-Omni-3B					
Ans	34.0	40.9	+6.9	[+2.8, +11.0]	0.002
Cal	36.7	34.1	−2.6	[−5.3, +0.3]	0.093
1-FA	74.9	84.7	+9.8	[+6.5, +12.9]	$< 10^{-8}$
TA	39.6	26.7	−12.9	[−16.3, −9.4]	$< 10^{-13}$
Qwen2.5-Omni-7B					
Ans	38.6	47.8	+9.2	[+4.3, +14.1]	$< 10^{-3}$
Cal	30.9	28.4	−2.5	[−5.8, +0.7]	0.151
1-FA	82.1	98.0	+15.9	[+12.2, +19.5]	$< 10^{-17}$
TA	22.6	7.2	−15.4	[−19.2, −11.4]	$< 10^{-14}$

Table S.5: **Reward asymmetry ablation.** Left: reward configurations used to separate the effects of abstention-related rewards from the penalty for unsupported answers; the TP reward is fixed at +1.0 in all configurations. Right: the corresponding effect on answerable performance, abstention behavior, and calibrated accuracy.

	Reward configuration			Result			
Configuration	TN	FN	FP	Recall ↑	1 − FA ↑	TA ↑	Cal. ↑
Base	–	–	–	34.0	74.9	**39.6**	**36.7**
Symmetric	+1.0	−1.0	−1.0	30.9	89.2	28.8	30.1
Abstention-softened	+0.5	−0.5	−1.0	39.1	**95.7**	8.4	24.4
Error-softened (Ours)	+1.0	−0.5	−0.5	**40.9**	84.7	26.7	34.1

ing ratio of 1:0.6. The training data, optimization procedure, and all other configurations are held constant; only the rewards assigned to different response outcomes are varied.

For an answerable query, a correct response is treated as a true positive (TP), whereas abstention is a false negative (FN), corresponding to false abstention. For an unanswerable query, abstention is a true negative (TN), whereas providing an unsupported answer is a false positive (FP), corresponding to false identification. Correct answers receive a fixed reward of +1.0 in all configurations, while the rewards or penalties for FN, TN, and FP outcomes are varied as summarized in Table S.5.

Asymmetric Reward Ablation. The reward configurations are designed to distinguish improved answerability discrimination from a general tendency to answer more frequently. Comparing *Abstention-softened* with *Symmetric* isolates the effect of reducing both the reward for correct abstention and the penalty for false abstention, while keeping the false-identification penalty fixed. Comparing *Error-softened (Ours)* with *Symmetric* instead halves both error penalties while retaining the full reward for correct abstention.

A symmetric assignment penalizes false abstention as strongly as false identification, which limits systematic bias toward either answering or abstaining but also suppresses answering overall. Our final configuration therefore softens both error penalties while retaining the full reward for correct abstention. The ablation evaluates whether this asymmetric shaping provides a more favorable balance between answering supported queries and abstaining from unsupported ones.

Table S.5 shows that all post-training configurations shift the model toward answering more frequently. Relative to the base model, the symmetric configuration raises 1 − FA from 74.9% to 89.2% while TA decreases from 39.6% to 28.8%, and recall drops from 34.0% to 30.9%: penalizing both error types equally increases answering attempts without converting them into correct answers.

Softening the abstention-related rewards further amplifies this tendency. Under *Abstention-softened*, 1 − FA rises to 95.7%, while TA falls to 8.4%. This change therefore reflects a general decrease in abstention rather than improved discrimination between answerable and unanswerable inputs.

Table S.6: **Reward-component ablation.** Localization drives answerable recall, whereas removing it collapses attempts into abstention and inflates Cal.

	Overall				Perceptual Ans	
Configuration	Recall ↑	1 − FA ↑	TA ↑	Cal. ↑	I2I	A2A
Base (no RLVR)	34.0	74.9	39.6	36.7	27.8	18.2
Full (Ours)	**40.7**	**84.9**	26.7	34.0	**29.6**	**28.2**
w/o Verification	37.6	74.4	39.0	35.3	26.6	20.5
w/o Localization	31.5	60.9	**61.3**	**45.7**	21.7	18.2
Text-QA only	31.4	73.4	43.5	32.7	22.1	12.7

In contrast, our *Error-softened* configuration, which halves both error penalties while keeping the full abstention reward, attains the highest recall (40.9%) while retaining TA at 26.7% and the best calibrated accuracy among the post-training configurations (34.1%).

Overall, how the softening is distributed matters. Softening only the abstention-related rewards trades TA away for a higher attempt rate, whereas softening both error penalties equally, as in our configuration, yields the best recall–calibration balance among the post-training configurations. Even so, its calibrated accuracy remains below the base model, indicating that reward asymmetry alone does not fully resolve the recall–abstention trade-off.

Compared with a symmetric non-negative reward, this formulation assigns a penalty to false negatives and a larger reward to correctly recognized positive matches. It therefore discourages a policy from improving its apparent reliability by indiscriminately rejecting candidate identities. False positives remain explicitly penalized to expose the opposite failure mode of indiscriminate positive prediction.

F.4 Reward-Component Ablation

We next isolate the contribution of each verifiable reward component. Starting from the full reward (localization, verification, and text QA; Appendix G), we remove one component at a time and additionally train a text-QA-only variant. All runs use `Qwen2.5-Omni-3B` under identical settings.

Table S.6 shows that the reward components jointly govern the trade-off between recall and abstention. *Localization* is the primary driver of answerable recall: removing it reduces overall recall from 40.7% to 31.5%, below the base model, while I2I recall falls from 32.2% to 21.7% and A2A recall returns to the base level (26.4% → 18.2%). At the same time, TA rises sharply from 26.7% to 61.3%, while 1 − FA falls from 84.9% to 60.9%, indicating a strong shift toward abstention on both unanswerable and answerable examples. This shift yields the highest Cal score (45.7%), but the gain reflects increased abstention rather than improved perceptual grounding.

Removing *verification* produces a more moderate decline, reducing overall recall by 3.1 points and perceptual recall by 5.6 points on I2I and 5.9 points on A2A. Although TA returns close to the base level (39.0% versus 39.6%), Cal remains below the base model (35.3% versus 36.7%). The *text-QA-only* variant performs worse than the base model in overall recall (31.4% versus 34.0%), Cal (32.7% versus 36.7%), and perceptual recall (I2I 22.1% versus 27.8%; A2A 12.7% versus 18.2%), demonstrating that text-only supervision does not account for the gains of the full reward.

Overall, the full reward achieves the highest overall and perceptual recall, whereas removing localization yields the highest Cal by substantially increasing abstention. These results identify localization as the main source of the recall–abstention tension and verification as an important contributor to perceptual grounding. They also show that the reward components are complementary: removing either component weakens recall, while text-QA supervision alone fails to improve either recall or calibration over the base model.

F.5 Additional Cross-Judge Robustness

To further assess the robustness of the LLM-based evaluation, we re-evaluated all answerable instances using GPT-5.6-sol with medium reasoning effort. The system prompt and evaluation procedure were identical to those used for the primary GPT-5.4-mini judge, with only the judge model changed. This

Table S.7: Judge robustness under alternative LLM judges, measured on an earlier snapshot of the model pool (11 models). Rank correlations are computed against the primary GPT-5.4-mini judge.

Alternative judge	Cal. rank correlation	Ans. rank correlation	Top model preserved
Gemini-3.1-Pro	$\rho = 0.89$ ($p < 0.001$)	$\rho = 0.63$ ($p = 0.039$)	Yes
GPT-5.6-sol	$\rho = 0.91$ ($p < 0.001$)	$\rho = 0.66$ ($p = 0.026$)	Yes

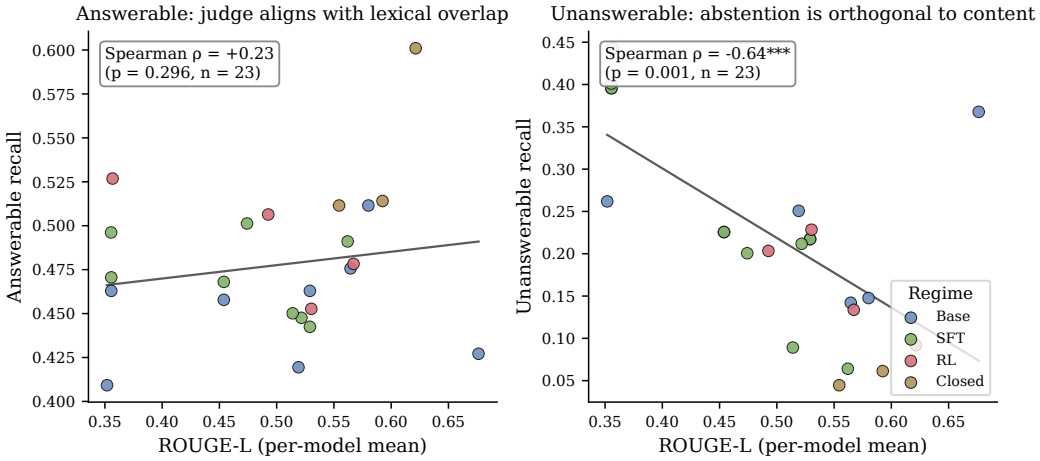


Figure S.4: Metric alignment across post-training regimes. The plot compares judge scores and lexical overlap for answerable and unanswerable queries. Crucially, it demonstrates the absence of content-induced bias in unanswerable cases, showing that abstention behavior is independent of content overlap.

check was run on an earlier snapshot of the model pool (11 models, 391 answerable instances each, no API failures), so it probes the reliability of the judge rather than re-scoring the final checkpoints reported in Table 2. Unanswerable instances remained keyword-scored and were therefore unaffected by the choice of LLM judge.

GPT-5.6-sol showed item-level agreement of 80.6–92.1% with the primary judge, with Cohen's κ ranging from 0.61 to 0.81. Although GPT-5.6-sol assigned lower answerable scores to most models, the overall model ordering remained similar: the Spearman rank correlation was $\rho = 0.91$ for calibrated accuracy ($p < 0.001$) and $\rho = 0.66$ for answerable recall ($p = 0.026$). Changing the judge therefore shifts absolute answerable scores more than it shifts the calibrated-accuracy ranking.

Table S.7 summarizes the results across the two alternative judges. In both cases, the calibrated-accuracy ranking was strongly correlated with that obtained using the primary judge, while the answerable-recall ranking was more judge-sensitive.

Overall, the consistency across two alternative judges indicates that the principal calibrated-accuracy ranking is reasonably robust to the choice of evaluator.

F.6 Alignment of Evaluation Metrics

LLM-as-a-judge evaluations and traditional lexical/semantic metrics capture fundamentally distinct signals. Because their correspondence varies substantially across query types, this benchmark reports answerable QA quality and unanswerable abstention behavior separately.

Why LLM-as-a-judge is Needed to Evaluate Free-Form Responses: Figure S.4 demonstrates that content-overlap metrics primarily explain performance on answerable queries, but fail to capture abstention behavior on unanswerable queries. That is, content quality and abstention capability are distinct evaluation axes. While answerable recall increases with ROUGE-L, indicating that lexical overlap partially reflects answer quality, unanswerable recall shows little association with it. This

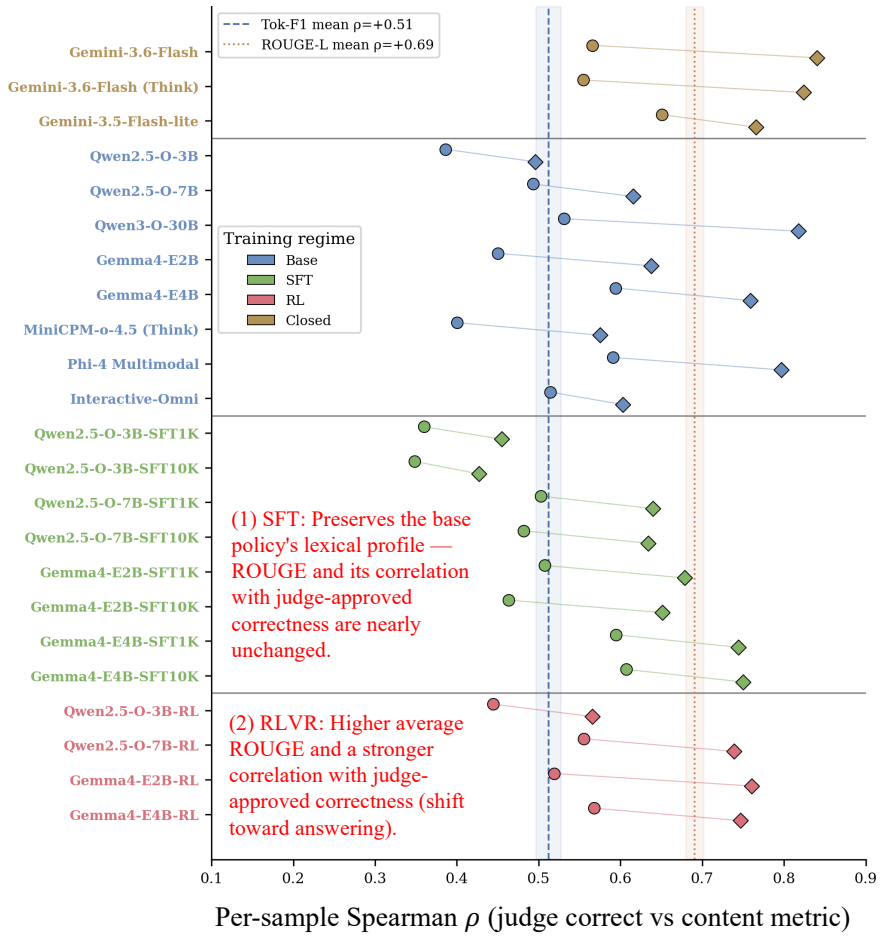


Figure S.5: Per-model agreement on answerable items only between judge and content metrics (circle = `Tok-F1`, diamond = `ROUGE-L`)

suggests that abstention behavior cannot be inferred from traditional content metrics and must be evaluated separately.

Key Finding 4: RLVR Improves Lexical Overlap and Judge–Metric Agreement

Unlike SFT, which largely preserves the lexical profile of the base policy, RLVR shifts model outputs toward answering. This raises conventional lexical overlap (ROUGE-L $+9.2$/$+10.0$ points on the Qwen backbones and modestly on the Gemma4 models) and, together with it, strengthens the per-sample agreement between the LLM judge and lexical metrics.

Per-Sample Metric Reliability: As illustrated in Figure S.5 and Table S.13, ROUGE-L exhibits stronger agreement with LLM-as-a-judge evaluations than token-level F1 on answerable items across all models. Strong closed-source and base models maintain high concordance between the judge and lexical metrics, and RL-tuned models attain the strongest alignment among the post-trained regimes (mean $\rho = 0.70$ versus 0.63 for the base policies). This indicates that reward optimization moves outputs toward grounded answering, which both raises corpus-level lexical overlap and makes lexical metrics track judge verdicts more closely. Conversely, SFT models largely preserve the judge–metric alignment of their base counterparts (mean $\rho = 0.62$), suggesting that supervised fine-tuning induces little lexical deviation from the original policy.

G Additional Experimental Configurations

G.1 Details on SFT Implementations

SFT Training Objective. We train the model on $\mathcal{D}_{\text{SFT}}$, a dataset of triplets $(q, \mathcal{C}, y^*)$, where q is a query, $\mathcal{C}$ is the omnimodal context, and y^* is the ground-truth response. We minimize the standard autoregressive negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(q,\mathcal{C},y^*)\sim\mathcal{D}_{\text{SFT}}} \sum_{t=1}^{|y^*|} \log \pi_\theta(y_t^* \mid q, \mathcal{C}, y_{<t}^*). \tag{S.1}$$

SFT Configurations. We fine-tune four backbones via ms-swift: `Qwen2.5-Omni` (3B/7B) and `Gemma4` (E2B/E4B). All SFT runs use $\text{lr} = 2 \times 10^{-5}$ over 3 epochs with vision/audio encoders frozen. We apply LoRA ($r = 64, \alpha = 128$) for the Qwen backbones and a more compact LoRA ($r = 32, \alpha = 64$) for the Gemma4 models. All post-training is performed with LoRA (Hu et al., 2022), using ms-swift (Zhao et al., 2024) for SFT. Within the Gemma4 series, audio processing is supported exclusively by the E2B and E4B variants.

SFT Data Corpus Construction. The 10K-scale SFT corpus (detailed in Table S.9) encompasses 12 distinct task types explicitly designed to enforce robust omnimodal alignment. The dataset comprehensively integrates foundational grounding tasks, audio-centric scenarios (*e.g.*, conversational and event sounds), and missing-modality variants that train the model to recognize and verbalize when required context is absent. To mitigate positional and cross-modal shortcut biases, we apply rigorous augmentations such as context reordering, distractor replacement, and modality swapping. Contextual memory is sourced from CoViP dialogues (Oh et al., 2026) and CUPID biographies (Kim et al., 2025b), with ground-truth visual captions rigorously filtered by an MCQA accuracy threshold ($\geq$50%). A downscaled 1K subset is also provided for experimental efficiency (Table S.8).

Training Data Corpus Overview. Our experimental pipeline comprises three distinct data stages (see Table S.11). The SFT corpus (1K and 10K variants) provides broad modality-alignment supervision, examining the impact of data scaling. Subsequently, the RLVR corpus supplies VR signals focused on two core grounding tasks: audio and image perception, and text-grounded retrieval. Notably, across both the SFT and RLVR regimes, we ensure that absent-persona (no-GT) samples account for approximately 37.5% of the training mixtures, following the 1:0.6 answerable-to-unanswerable ratio. This deliberate inclusion is intended to encourage calibrated abstention when confronted with unanswerable queries. The resulting trade-off between answerable recall and Cal is governed jointly by the reward design and the answerable-to-unanswerable data composition, as analyzed through the ratio sweep (Figures S.2 and S.3) and the reward-asymmetry ablation (Appendix F.3). The dataset distribution across each training stage is illustrated in Figure S.6.

G.2 Details on RLVR Implementations

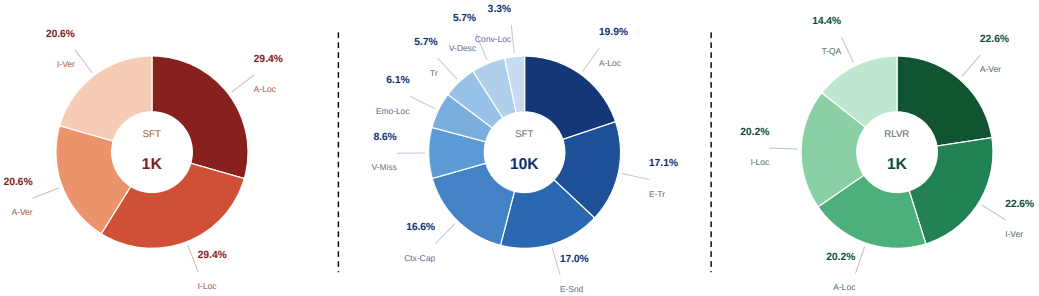
We adopt Group Sequence Policy Optimization (GSPO; Zheng et al., 2025) as the base optimizer for our RLVR, with a moderate KL regularization coefficient $\beta = 0.04$.

RLVR Training Objective. We optimize π_θ to maximize a verifiable reward while penalizing the KL divergence (Shao et al., 2024; Zheng et al., 2025) from the reference policy π_{ref}:

$$\max_\theta \mathbb{E}_{(q,\mathcal{C})\sim\mathcal{D}_{\text{tr}}} \mathbb{E}_{y\sim\pi_\theta(\cdot|q,\mathcal{C})} \left[r(y, q, \mathcal{C}) - \beta \mathrm{D}_{\text{KL}} \left(\pi_\theta(\cdot \mid q, \mathcal{C}) \| \pi_{\text{ref}}(\cdot \mid q, \mathcal{C}) \right) \right] \tag{S.2}$$

here, q is a query, $\mathcal{C}$ is omnimodal contexts, y is the model response. For each training instance, the verifiable reward $r(y, q, \mathcal{C})$ is specifically determined based on the success of y in the assigned task as follows:

RLVR Configurations. Without an SFT warmup stage, we optimize the RLVR objective using GSPO (Zheng et al., 2025) on 1K samples. Key hyperparameters include: $G = 4$ generations, $\beta = 0.04$, $\text{lr} = 1 \times 10^{-5}$, and LoRA $r = 32, \alpha = 64$. Similar to SFT, we freeze all perceptual encoders to maintain pre-trained representations. We note that our post-training hyperparameters



(a) SFT 1K dataset (b) SFT 10K dataset (c) RLVR 1K dataset

Figure S.6: **Post-training data strategy.** Visualization of the task coverage and data distributions across the SFT (1K), SFT (10K), and RLVR training regimes.

Table S.8: **SFT (1K) task distribution**. Overview of the task proportions and corresponding abbreviations utilized in our post-training ablations.

Task name	Sample query text
Audio localization	Select which reference audio panel contains the same speaker as the query audio.
Image localization	Select which reference image panel contains the same person/object as the query image.
Audio verification	Answer yes/no on whether the query audio matches the referenced speaker/audio identity.
Image verification	Answer yes/no on whether the query image matches the referenced person/object identity.

Table S.9: **SFT (10K) task distribution**. Overview of the task proportions and corresponding abbreviations utilized in our post-training ablations.

Task name	Sample query text
Audio localization	Select the reference speaker/audio panel that matches the query audio.
Event-aware transcription	Transcribe speech while handling an added environmental sound event.
Event sound MCQA	Identify the environmental sound mixed with speech from multiple choices.
Contextualized captioning	Describe the query image using the matching memory/context panel when relevant.
Missing-image visual description	Describe visual information when the target image modality is absent.
Emotion speaker localization	Find which reference speaker matches the query speech/emotion cue.
Speech transcription	Convert the target speech audio into text.
Visual description	Generate a grounded description of the target image.
Conversation speaker localization	Identify the matching speaker/person from dialogue-conditioned references.

Table S.10: **Task distribution in the RLVR (1K) dataset.** This table summarizes the task abbreviations used in the RLVR training distribution. For all referential tasks, the target responses are restricted to binary "Yes" or "No" labels.

Task name	Sample query text
Audio verification	Decide whether the query audio matches the referenced speaker/audio identity.
Image verification	Decide whether the query image matches the referenced person/object identity.
Audio localization	Select which reference audio panel contains the same speaker as the query audio.
Image localization	Select which reference image panel contains the same person/object as the query image.
Text QA	Answer a context-grounded question or abstain when the answer is not supported.

Table S.11: Overview of data configurations for training and evaluation phases.

Stage	Size	# of Contexts	Tasks	Source Modality		
				Image	Text	Audio
SFT Corpus	1K / 10K	3	12 foundational task types	Synthetic-only	Synthetic (`GPT-5.4`)	Real & Syn. (80%)
RLVR Corpus	1K	3	3 core grounding tasks (Visual ID, Voice ID, Text QA)	Synthetic-only	Synthetic (`GPT-5.4`)	Real & Syn. (80%)
Eval Benchmark	750	4	18 fine-grained tasks (4 groups)	Real-only	Synthetic (`GPT-5.5`)	Real & Syn. (20%)

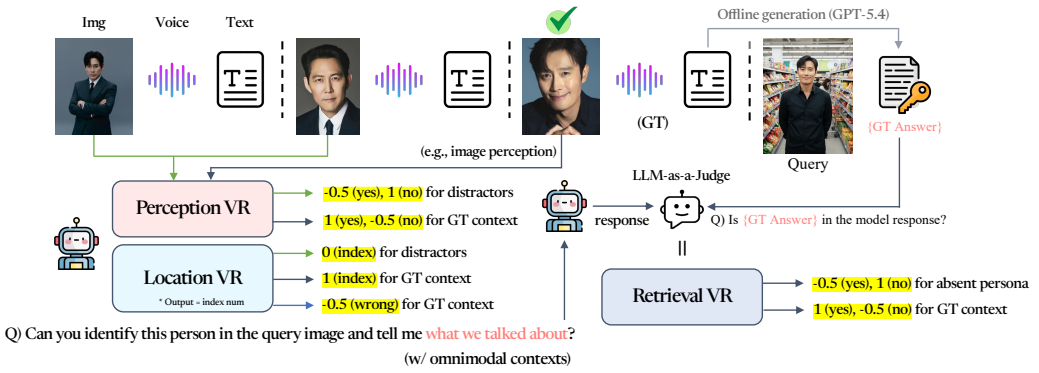


Figure S.7: **Overview of the perception, localization, and retrieval rewards used for RLVR.** The rewards use error-softened scoring rather than strictly binary signals. Perception and localization rewards supervise identification of the target context and its position among distractors, while the retrieval reward evaluates grounded answering when the target persona is present and appropriate abstention otherwise. Answer correctness is assessed against offline reference answers using an LLM judge. Training-data details are provided in Table S.11.

were tuned primarily on `Qwen2.5-Omni-3B` and reused across backbones, so they may be suboptimal for the Gemma4 models.

Verifiable Reward Designs. Mirroring the three reward components in Section 5.1, our reward function is dispatched by task type $t(q)$ into *perception*, *location*, and *retrieval*:

$$r_{\text{base}} = \begin{cases} r_{\text{perc}}(y, q, \mathcal{C}), & \text{if } t(q) = \text{PERC}, \\ r_{\text{loc}}(y, q, \mathcal{C}), & \text{if } t(q) = \text{LOC}, \\ r_{\text{qa}}(y, q, \mathcal{C}), & \text{if } t(q) = \text{QA}. \end{cases} \tag{S.3}$$

All three share the same *asymmetric* outcome scheme. Writing the four outcomes of a verifiable decision as true positive (TP; a correct grounded answer or a correct match), true negative (TN; a correct abstention or a correct rejection), false negative (FN; false abstention or a missed match), and false positive (FP; an unsupported answer or a false match), we assign

$$r(\text{TP}) = r(\text{TN}) = +1.0, \qquad r(\text{FN}) = r(\text{FP}) = -0.5, \tag{S.4}$$

with all remaining outcomes (*e.g.*, an incorrect grounded answer or an unparsable response) scored 0. Rewarding TP and TN equally prevents the policy from inflating its apparent reliability through blanket answering or blanket abstention, while the softened -0.5 penalties keep both error types costly without over-suppressing attempts; this asymmetry is ablated in Appendix F.3.

1) Perception VR. Perception VR supervises whether a target perceptual signal, such as an image or audio clip, matches a given context. Queries expose only the target modality to prevent cross-modal shortcutting. Correct matches and rejections receive $+1.0$, while false negatives and false positives receive -0.5.

2) Location VR. Location VR provides finer-grained supervision by requiring the model to select the matching context index or a "None" option when no match exists. Correct index selection and correct abstention receive $+1.0$; false negatives and false positives receive -0.5; incorrect indices and unparsable outputs receive 0.

3) Retrieval VR. Retrieval VR supervises grounded open-ended answering and appropriate abstention. When the target persona is present, an LLM judge evaluates answer correctness; when absent, abstention is checked using the evaluation-time lexical rule. Correct answers and abstentions receive $+1.0$, false negatives and false positives receive -0.5, and incorrect grounded answers receive 0. Although these rewards explicitly supervise abstention, the resulting policies still exhibit a bias toward answering.

Safeguards Against Reward Hacking. We include several safeguards to reduce reward hacking during RL fine-tuning. Outputs with clear degeneration patterns, such as repeated 4-grams, character-level spam, or sentence loops, receive zero reward. We also parse LLM-judge outputs with strict

exact matching, so that only the verdict `correct` is rewarded and strings such as `incorrect` cannot be mistakenly counted as positive. Invalid ground-truth labels are treated as explicit errors rather than ignored silently. These safeguards make the reward signal more reliable while keeping the training procedure simple.

On-Policy RLVR Algorithms. Given a prompt x, group size G, and sampled responses $\{y_i\}_{i=1}^{G}$ with group-normalized advantages $\hat{A}_i$, the two objectives differ in the granularity at which the importance ratio is computed and clipped:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\Big(r_{i,t}(\theta)\,\hat{A}_i,\ \text{clip}(r_{i,t}(\theta),\,1\pm\varepsilon)\,\hat{A}_i\Big)\right] - \beta\,\mathbb{D}_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}]\,, \tag{S.5}$$

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\Big(s_i(\theta)\,\hat{A}_i,\ \text{clip}(s_i(\theta),\,1\pm\varepsilon)\,\hat{A}_i\Big)\right], \tag{S.6}$$

where $r_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, y_{i,<t}) \,/\, \pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t})$ is the per-token ratio, and

$$s_i(\theta) \;=\; \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\text{old}}(y_i \mid x)}\right)^{1/|y_i|} \tag{S.7}$$

is the length-normalized sequence-level ratio. Both objectives share the same KL reference π_{ref}; the substantive distinction lies in Eq. (S.7) and in whether clipping is applied per-token (Eq. (S.5)) or per-sequence (Eq. (S.6)).

Why Sequence-Level Optimization Suits our RLVR. Our RLVR verifier combines rule-based binary verification (e.g., 'yes'/'no' compliance) with an LLM-as-judge component, producing a reward signal that is defined at the sequence outcome level rather than at the token level. Under GRPO (Shao et al., 2024), a single noisy high-reward rollout propagates its advantage through per-token ratios; consequently, tokens that happen to diverge from π_{old} receive disproportionately large updates, even if they are not causally responsible for the binary outcome. In contrast, GSPO (Zheng et al., 2025) aggregates the ratio over the full response before clipping, ensuring that per-token fluctuations cancel out and only the geometric-mean deviation of the sequence survives. This alignment between clipping and reward granularity significantly reduces gradient variance in the presence of judge-model false positives and rule-based shortcut exploitation.

Role of $\beta = 0.04$ in GSPO. We use $\beta = 0.04$ as a conservative choice given the LLM-as-a-judge reward noise, noting that GSPO's original formulation omits KL regularization. RLVR begins from a LoRA-adapted policy without any SFT warmup on the target persona distribution. At step 0, $\pi_\theta \approx \pi_{\text{ref}}$ and $\mathbb{D}_{\text{KL}} \approx 0$, so the choice of β governs how much cumulative drift is tolerated once advantages become informative. We found that $\beta = 0$ (unconstrained) permits rapid drift, which in preliminary runs led to format collapse and over-assertive answering on unanswerable prompts. In contrast, $\beta = 0.04$ preserved the group-relative advantage signal while bounding drift to a trust-region regime.

Variance-Aware Rollout Resampling. Group-relative policy optimization requires within-group reward variation to produce an informative relative-advantage signal. For a prompt x, let $\{r_{x,1}, \ldots, r_{x,G}\}$ denote the rewards of its G sampled completions. We classify the rollout group as *live* when

$$\text{Std}\left(r_{x,1}, \ldots, r_{x,G}\right) > \epsilon, \tag{S.8}$$

and as *dead* otherwise. Dead groups arise when all completions receive the same reward and therefore contribute little or no group-relative learning signal.

When a dead group is encountered, we regenerate completions for the same prompt with an increased sampling temperature. We perform at most two retries, multiplying the temperature by 1.25 at each retry up to a maximum of 1.6, starting from an initial temperature of 1.1. Training retains the first

Table S.12: Keyword and phrase list used for lexical abstention detection.

Abstention Keywords / Target Surface Forms	
cannot determine	cannot be determined
not enough information	insufficient information
cannot answer	unable to determine
don't know from	do not know from
the provided context does not	not provided in the context
no information in the context	context does not contain
cannot identify	i cannot tell

live group obtained; groups that remain uninformative after the final retry are excluded from the group-relative loss.

This procedure is inspired by dynamic sampling in DAPO (Yu et al., 2026) but is adapted to the small effective prompt batch used in our setting. Directly discarding all zero-variance groups would remove many all-wrong positive-match groups, even though these groups reveal precisely the grounding failure that we seek to correct. Prompt-level resampling instead attempts to recover informative variation before excluding a group.

H Limitations and Broader Impacts

Synthetic-Real Domain Bias in Post-Training. Real-world face images and voice recordings of consistent identities are difficult to obtain at the scale required for systematic SFT and RLVR training, particularly given privacy, consent, and licensing constraints. We therefore train on synthetic personas (TTS-generated voice clips and generated facial images) while reserving real images strictly for the Omni-Persona evaluation benchmark. This deliberate split preserves benchmark realism and also provides a natural testbed for studying how well post-training generalizes from synthetic personas to real-world distributions, a question we view as important for future omnimodal personalization research.

Difficulty of Evaluation-Aligned In-Domain SFT Data. One could alternatively use an SFT dataset that closely matches the evaluation distribution. However, constructing such a dataset at scale is a significant challenge. Generating high-quality answers without accidentally exposing test-set information is difficult, and using test-style queries during training risks benchmark contamination. Therefore, we treat our SFT results as an analysis of broad-coverage training. Developing a reliable method to create high-quality, in-domain data while maintaining benchmark integrity remains an important goal for future research.

Scope of Reward Design Ablations. Our RLVR pipeline includes several stability filters, such as 4-gram repetition, character diversity, and sentence-repetition checks, which effectively prevent severe degeneration patterns that can occur during RL training on thinking-style models. Fine-grained reward shaping for balancing grounding and abstention is a distinct and complementary direction; we leave such ablations to future work. Importantly, we view the answer-inducing abstention collapse surfaced by our TA and Cal metrics as a diagnostic strength of Omni-Persona, rather than as evidence that RLVR is inherently prone to this behavior, and we expect it to be addressable with more targeted reward design.

Limitations of Lexical Abstention Detection. To determine whether a model abstains on unanswerable queries, we use a simple lexical matching rule. A response is classified as an abstention if it contains any predefined phrase listed in Table S.12. The same rule is applied both during RLVR training, where it defines the abstention-related reward signal, and at evaluation time, making abstention evaluation directly comparable without requiring an additional judge model. This lexical design may undercount valid abstentions, particularly when base or SFT models use phrases outside the predefined keyword list. Although the evaluation prompt in Table S.15 instructs models to use "I cannot determine that from the provided context" for unsupported answers, models may still produce paraphrases or partial uncertainty statements that are not captured by the detector.

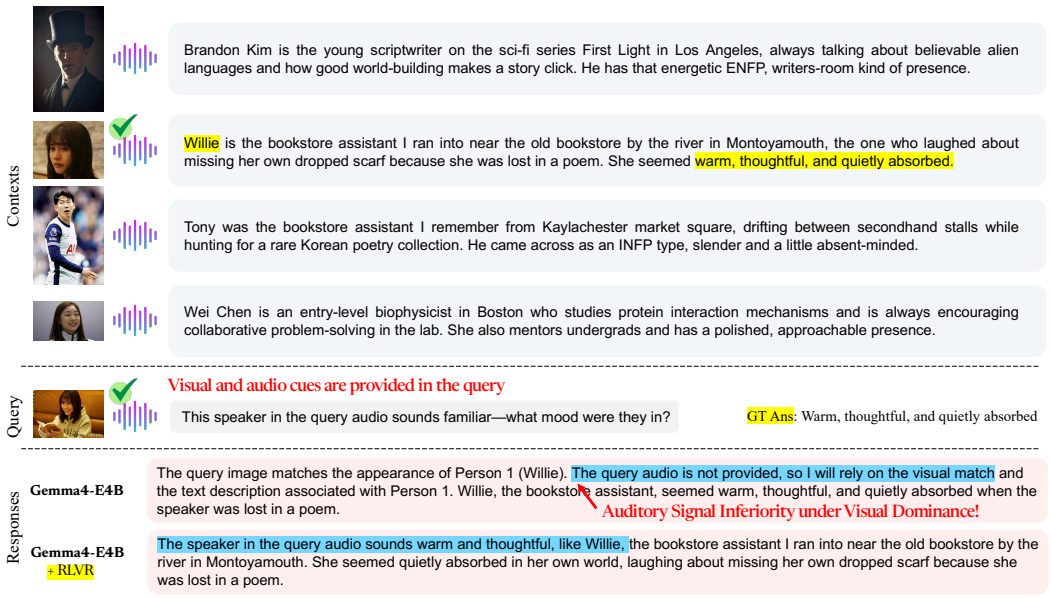


Figure S.8: **Qualitative comparison of model predictions given the query and its GT answer.** When the query includes both image and audio modalities and audio is the primary perceptual cue, Gemma4-E4B tends to overlook the audio signal and rely heavily on visual cues. In contrast, the RLVR-trained model grounds its response in the relevant audio signal.

Privacy and Consent. Omni-Persona is constructed from publicly available or synthetically generated user signals, including face images, voice samples, and biographical text, and does not include personally identifiable information from real individuals without consent. However, systems built for real-user personalization must obtain explicit consent for collecting, storing, and processing biometric signals such as voice and facial imagery.

Limited Coverage of Perceptual Uncertainty. The benchmark evaluates relatively fine-grained visual and auditory discrimination within a provided query–context setting; it does not fully represent open-ended uncertainty in unconstrained real-world environments. The instances nevertheless include several realistic sources of difficulty, including background speech, conversational voices, perceptually similar distractors, and missing textual information. These factors are not separately annotated according to a controlled difficulty taxonomy such as ambiguous evidence, noisy audio, occlusion, partial evidence, or highly similar candidates. The current benchmark therefore cannot quantify the degradation attributable to each source of perceptual uncertainty. A controlled analysis that isolates these uncertainty types is an important direction for future work.

Bias and Representation. Synthetic persona generation may underrepresent certain demographic groups, leading to uneven personalization quality across users. While our benchmark includes hard distractors and retrieval noise to test robustness, fairness across multilingual voice, facial, and linguistic attributes remains an important direction for future evaluation.

Future Directions. Our work remains limited to the post-retrieval grounding and generation stage. A natural next step is therefore to integrate Omni-Persona with end-to-end retrieval-augmented pipelines that jointly evaluate retrieval quality, multimodal grounding, generation, and calibration, while carefully distinguishing retriever errors from downstream grounding failures. This framework could further be extended to on-device personalization benchmarks constructed from unstructured omnimodal user signals, such as daily voice memos, image galleries, and textual interaction histories, as well as to agentic tool-use settings in which models perform complex cross-modal actions grounded in locally stored user context.

Table S.13: **Complementary lexical and semantic metric results.** Note: BS = BERTScore, 1-FA = 1-FalseAbs, TA = TrueAbs, MLen = Mean Length.

Model	Accuracy			Generation Quality				Avg	Add. Metrics (×100)			MLen
	Overall(Cal)	Ans	Unans	Abs-F1	ROUGE-L	Tok-F1	BS		1-FA	TA	Avg	
Closed-source Models												
Gemini-3.6-Flash	**35.7**	**60.1**	**9.2**	**16.8**	**62.1**	**17.2**	**84.7**	**40.8**	99.7	**9.2**	**54.5**	35.6
Gemini-3.6-Flash (Think)	28.8	51.2	4.5	8.5	55.5	15.8	84.3	35.5	**100.0**	4.5	52.2	32.5
Gemini-3.5-Flash-lite	29.7	51.4	6.1	11.5	59.2	16.2	84.7	37.0	99.5	6.1	52.8	**40.2**
Open-source Models												
Phi-4 Multimodal	33.2	40.7	25.1	37.4	51.9	12.5	83.8	40.7	**91.8**	25.1	58.5	77.2
MiniCPM-o 4.5 (Think)	**39.3**	**41.7**	**36.8**	**48.9**	**67.6**	5.4	79.8	**45.6**	87.5	**36.8**	**62.1**	**716.0**
Interactive-Omni	29.6	32.7	26.2	36.7	35.2	**14.3**	**85.1**	37.1	84.9	26.2	55.6	20.3
Qwen3-Omni-30B	33.7	51.2	14.8	24.4	58.0	13.2	84.2	39.9	94.4	14.8	54.6	59.0
Qwen2.5-Omni-3B	36.7	34.0	39.6	47.4	35.6	15.3	85.3	42.0	74.9	39.6	57.2	18.1
SFT (1K)	36.4	33.5	39.4	47.2	35.8	15.2	85.4	41.8	75.1	39.4	57.3	18.2
SFT (10K)	**36.9**	34.0	**39.8**	**48.1**	35.5	15.4	85.3	**42.1**	74.7	**39.8**	**57.1**	18.1
RLVR	34.1	**40.9**	26.7	37.3	**44.8**	**19.2**	**85.9**	41.3	**84.7**	26.7	55.7	**19.6**
Qwen2.5-Omni-7B	30.9	38.6	22.6	**31.8**	45.4	14.6	84.8	38.4	82.1	22.6	52.3	35.1
SFT (1K)	31.5	39.6	22.4	31.6	45.6	14.5	84.9	38.6	82.3	22.4	52.3	35.3
SFT (10K)	**31.9**	42.7	**22.8**	29.2	47.4	14.7	84.9	**38.7**	81.9	**22.8**	52.1	35.8
RLVR	28.4	**47.8**	7.2	13.2	**55.4**	**17.8**	**85.5**	36.5	**98.0**	7.2	**52.6**	**36.7**
Gemma4-E2B	**32.1**	41.7	21.7	**32.4**	52.9	9.9	**80.4**	**38.7**	88.7	21.7	**55.2**	72.8
SFT (1K)	31.9	41.2	**21.9**	32.2	53.1	9.8	80.5	38.6	**88.9**	**21.9**	**55.2**	72.6
SFT (10K)	31.3	40.7	21.5	31.8	52.2	9.9	80.3	38.2	88.5	21.5	55.1	72.7
RLVR	28.1	**44.5**	10.3	19.0	**54.9**	**10.3**	80.3	35.3	**96.7**	10.3	53.5	**76.9**
Gemma4-E4B	31.2	46.8	14.2	23.3	56.5	15.2	84.2	38.8	92.8	14.2	**53.5**	45.1
SFT (1K)	26.9	43.5	8.9	15.9	51.4	15.7	**84.7**	35.3	**96.9**	8.9	52.9	34.2
SFT (10K)	**33.7**	46.8	**19.5**	**28.9**	**58.4**	15.2	83.6	**40.9**	85.7	**19.5**	52.6	**57.5**
RLVR	30.4	**48.6**	10.6	22.4	58.4	**16.1**	84.6	38.7	94.4	10.6	52.5	42.9

I Further Analysis on Lexical Metrics

Lexical Metrics Validate, but Cannot Replace, the LLM Judge. We first examine whether our LLM-as-a-judge verdicts are consistent with reference-based lexical signals. For each model, we partition answerable predictions ($N = 391$) by judge verdict and recompute lexical scores within each partition. Across all evaluated backbones, predictions judged as CORRECT consistently achieve substantially higher ROUGE-L than those judged as WRONG, typically by a factor of 2–4. This pattern suggests that the judge is aligned with the same reference-overlap signal captured by conventional lexical metrics. However, lexical overlap remains too sparse for free-form personalized QA. Per-item lexical overlap remains sparse: threshold-based accuracy using ROUGE-L ≥ 0.5 stays below 14% across models. Thus, lexical metrics provide a useful sanity check for judge reliability, but they cannot serve as the primary evaluation criterion. This motivates a calibrated, behavior-aware metric that accounts for both grounded answering and appropriate abstention.

Post-training Regimes Show Different Behavioral Fingerprints. From this calibrated perspective, SFT and RLVR produce qualitatively different behavioral shifts. SFT-10K largely preserves the base-model behavior: ROUGE-L changes only marginally, and the genuine hallucination rate remains nearly unchanged. This suggests that broader supervised data alone does not reliably reshape the model's abstention boundary in open-ended personalization settings. In contrast, RLVR more directly reshapes this boundary, but in the answer-inducing direction: it shifts every backbone toward attempting answers, reducing false abstention on answerable queries, while unsupported answers on unanswerable queries rise and Cal fails to improve over the base model. This reflects a clear trade-off under our current reward design: outcome-level rewards targeting grounding and reasoning yield large answerable-recall gains for the Qwen backbones (with Gemma recall near base), but fail to induce calibrated abstention. The evaluated models illustrate the same tension from another angle. `MiniCPM-o 4.5` achieves strong calibrated behavior, whereas `Gemini-3.6-Flash` tends toward over-answering. These extremes confirm that single-axis metrics are insufficient: lexical quality, recall, hallucination, and abstention behavior can move independently. Calibrated Accuracy is therefore an appropriate headline metric for personalized omnimodal QA because it jointly penalizes blanket abstention and unsupported guessing.

J Detailed Task Taxonomy and Granularities

Table S.14: Fine-grained task taxonomy of the Omni-Persona benchmark: 4 matching scenario groups and 18 sub-tasks derived from the PMG formulation. Each group represents a primary matching scenario defined by its query and context modalities; sub-tasks within a group sweep the retrieval target modality. Symbols: I (image), T (text), A^v (audio/voice sample). Every sub-task additionally has an absent-persona (no-GT, $e_{q\to j}$=0, target=None) variant that requires structured abstention.

ID	Query	Context	Target	Task	Sample question
Group 1 (I2I): Visual Identification (Query: I, Context: I)					
1-a	I	I	T	Biography	*Who is this person, and what is their job?*
1-b	I	I	A^v	Dialogue recall	*Looking at this person in the image, I'm trying to remember what they said before.*
1-c	I	I	I	Appearance	*This person in the image seems familiar. What did they look like?*
1-d	I	I	A^v	Emotion	*When I met this person in the image, what mood were they in?*
1-e	I	I	A^v	Environment	*Who is this person in the image, and where were we?*
Group 2 (A2A): Voice Identification (Query: A^v, Context: A^v)					
2-a	A^v	A^v	T	Biography	*Who is the speaker in the query audio, and what is their affiliation?*
2-b	A^v	A^v	A^v	Dialogue recall	*Whose voice is in the query audio, and what had they said earlier?*
2-c	A^v	A^v	I	Appearance	*I recognize this speaker from the audio. Can you describe their appearance?*
2-d	A^v	A^v	A^v	Emotion	*This speaker in the query audio sounds familiar. What mood were they in?*
2-e	A^v	A^v	A^v	Environment	*This speaker in the query audio belongs to someone I know. What was happening around them?*
Group 3 (T2T): Same-Modal Semantic (Query: T, Context: T)					
3-a	T	T	T	Biography	*The professional esports player, what was their occupation?*
3-b	T	T	I	Appearance	*I'm thinking of the bookstore café barista. What were they wearing?*
3-c	T	T	T	Emotion	*The freelance illustrator, how were they feeling?*
3-d	T	T	T	Environment	*The gallery assistant, where were we when that happened?*
Group 4 (T2Any): Cross-Modal Semantic (Query: T, Context: A^v)					
4-a	T	T	T	Biography	*The one who said that, what was their affiliation?*
4-b	T	T	I	Appearance	*The person who talked about a peaceful reunion over poetry and a dropped scarf, what did they look like?*
4-c	T	T	A^v	Emotion	*The person who talked about an unexpected rainy-day encounter outside an art gallery, how were they feeling then?*
4-d	T	T	T	Environment	*Can you identify the person who talked about a memorable fan encounter at a gaming convention, and tell me where we were?*

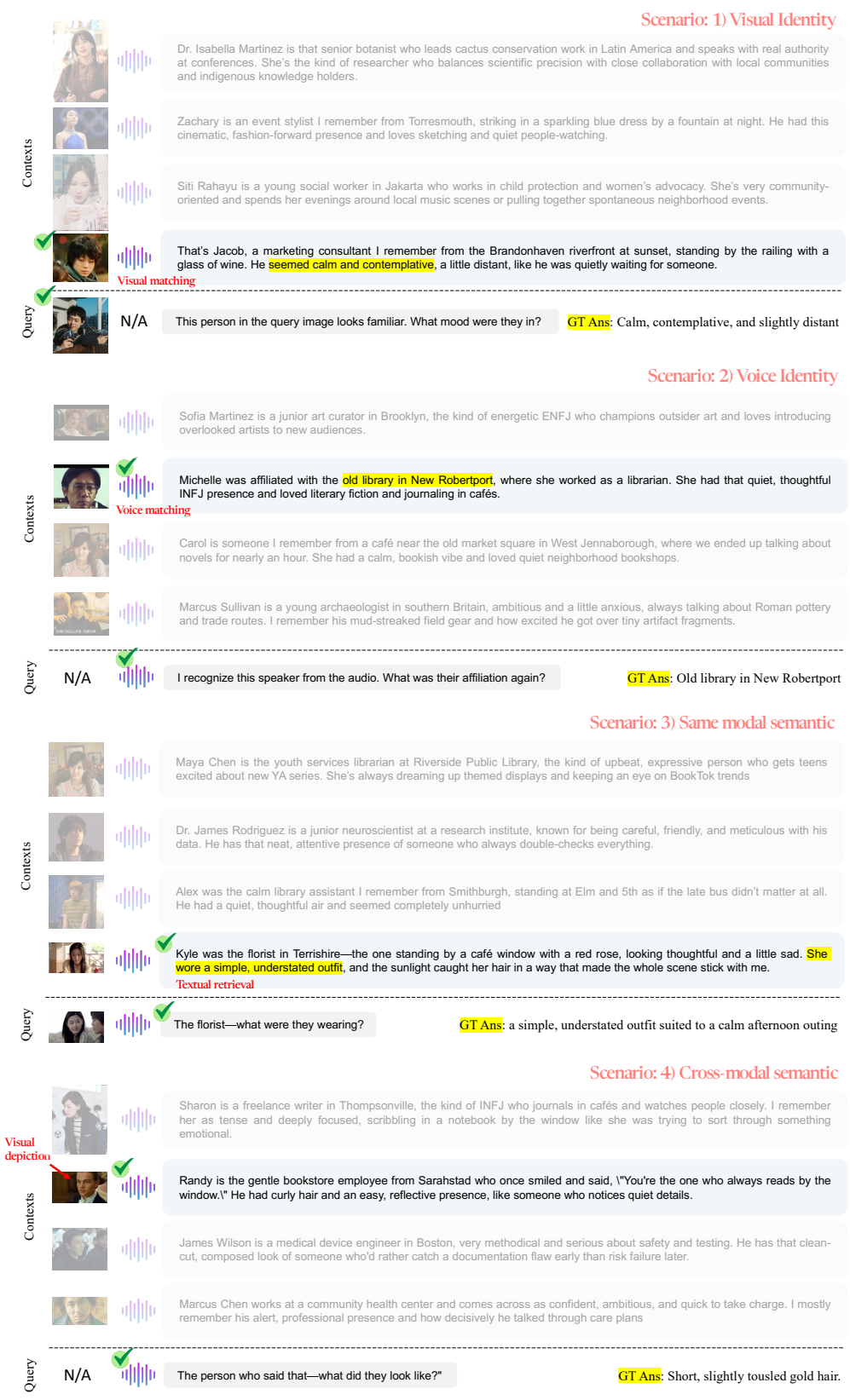


Figure S.9: Illustrative examples for each scenario within the Omni-Persona benchmark. Specifically, only answerable instances (where the target persona is present) are visualized in this figure.

K USED TEMPLATES FOR DATASET CONSTRUCTION

Table S.15: Fixed prompt template used for evaluation in our Omni-Persona benchmark.

User Prompt Prompt:
You are a personal memory assistant for a multimodal personalization benchmark.

[Context Setup]
You will receive memory contexts about four different people (`Context 0` through `Context 3`). Each context may contain some combination of:

- A photo of the person (`image`)
- A voice recording or spoken clip of the person (`audio`)
- A text description with personal details such as name, job, hobbies, personality, location, etc.

After the four memory contexts, you will receive a `Query` section that may include:

- A query image showing a person (use it to identify which context person this is)
- A query audio clip of a person speaking (use it to identify which context person this is)
- A question asking for specific information about that person

[Task]

1. Carefully examine all four contexts, attending to every available modality.
2. Use the query image, audio, or text to identify which context (`Person 0-3`) the query refers to.
   - If the query includes an image, match the face or appearance to the context images.
   - If the query includes an audio clip, match the voice to the context audio recordings.
   - If the query is text-only, use the semantic description to identify the correct person.
3. Answer the question using only the information from the matched context.

[Response Rules]

1. Before answering, briefly reason through which person matches the query and what relevant information their context contains.
2. Provide a specific and informative answer in **1–3 sentences**, including all relevant details available in the matched context.
3. Base the answer solely on the provided contexts; do not hallucinate or infer beyond what is explicitly given.
4. Do not repeat the question.
5. Use `I cannot determine that from the provided context.` only as a last resort, when the requested information is genuinely absent from all modalities of the matched person's context.
6. If any modality provides a partial or indirect answer, use it rather than abstaining.

[Output Behavior]
The response should first briefly identify the matched person and summarize the relevant evidence, and then provide the final answer grounded only in the corresponding context.

Table S.16: Judge prompt used for answer correctness.

Evaluation Judge Prompt:
You are an evaluation judge for a question-answering benchmark. You will be given a gold answer and a model prediction, and must decide whether the prediction is correct.

[Judging Rules]

1. For factual answers (*non-abstain*), the prediction is `CORRECT` if its final answer conveys the same core information as the gold answer, even if the wording differs.
   - Minor paraphrasing is acceptable.
   - The prediction is also `CORRECT` if the gold answer is explicitly contained in the prediction as a clear and stated fact.
2. The model receives multiple memory contexts and must identify the correct person before answering.
   - If the prediction lists multiple conflicting answers across different people, it is `WRONG`.
   - If the gold answer is only mentioned in passing while a different answer is presented as the model's conclusion, it is `WRONG`.
   - If the model hedges without committing to a final answer, it is `WRONG`.
3. For abstain gold answers ("`I cannot determine...`"), the prediction is `CORRECT` only if it also abstains.
   - Any concrete answer in this case is `WRONG`.
4. Output exactly one word: `CORRECT` or `WRONG`.
   - Do not provide any explanation.

[User Input Template]
Input:

```
Gold answer: {gold}
Model prediction: {pred}
Verdict:
```

Table S.17: Visualization of the structured prompt used for automated persona profiling and attribute enrichment.

Personal Profile Builder Prompt (JSON-only):
You are a personal profile builder for a multimodal personalization benchmark.

[Task]
Your job has two stages:

1. **EXTRACT**: pull out any facts explicitly stated in the source text (e.g., name, job, MBTI).
2. **ENRICH**: for every field still `null` after extraction, invent a plausible and realistic value consistent with the dialogue style.

Think like a person writing down memory notes: assign believable personal attributes such as a hobby, a past travel experience, a job, an MBTI type, a memorable quote, an emotional state, or a physical appearance so that the profile feels like a real person someone might remember.

[Constraints]

1. Never contradict explicitly stated facts in the source text.
2. Keep each value concise (1–2 sentences). Values for experience/preference fields should reflect realistic, personal anecdotes.
3. Output must be strict JSON matching the profile schema.

[Input]
`concept_id: {concept_id} source_type: {dialogue_type} SOURCE: {source_text}`

[Profile Schema]

```
{
  "hobby": "<string>",
  "major": "<string|null>",
  "affiliation": "<string|null>",
  "role": "<string>",
  "interest": "<string>",
  "topic": "<string>",
  "quote": "<string>",
  "location": "<string>",
  "environment": "<string>",
  "emotion": "<string>",
  "appearance": "<string>",
  "clothing": "<string>",
  "hairstyle": "<string>",
  "voice_identity": "<string|null>",
  "source_style": "<string>",
  "travel_experience": "<string|null>",
  "food_preference": "<string|null>",
  "life_event": "<string|null>"
}
```