VAPO: End-to-end Slide-Enhanced Speech Recognition with Omni-modal Large Language Models

Rui Hu[1,2], Delai Qiu[2], Yining Wang[2], Shengping Liu[2], Jitao Sang[1,3]*
[1]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University
[2]Unisound AI Technology Co., Ltd.
[3]State Key Laboratory of AI Safety, Beijing
{rui.hu, jtsang}@bjtu.edu.cn
https://github.com/isruihu/SlideASR-Bench

Abstract

Omni-modal large language models (OLLMs) offer a promising end-to-end solution for slide-enhanced speech recognition due to their inherent multimodal capabilities. However, we found a fundamental issue faced by OLLMs: *Visual Interference*, where models show a bias towards visible text over auditory signals, causing them to hallucinate slide content that was never spoken. To address this, we propose Visually-Anchored Policy Optimization (**VAPO**), which aims to reshape models' inference process to follow the human-like "*Look-then-Listen*" inference chain. Specifically, we design a temporally decoupled policy: the model first extracts visual priors in a <*think*> block to serve as semantic anchors, then generates the transcription in an <*answer*> block. The policy is optimized via multi-objective reinforcement learning. Furthermore, we introduce SlideASR-Bench, a comprehensive benchmark designed to address the scarcity of entity-rich data, comprising a large-scale synthetic corpus for training and a challenging real-world test set for evaluation. We conduct extensive evaluations demonstrating that VAPO effectively eliminates visual interference and achieves state-of-the-art performance on SlideASR-Bench and public datasets, significantly reducing entity recognition errors in specialized domains.

1 Introduction

Current Automatic Speech Recognition (ASR) models, such as Whisper (Radford et al., 2023), have demonstrated impressive performance in general domains. However, recognition accuracy often deteriorates significantly in specialized scenarios, such as academic lectures or technical presentations where domain-specific terminologies and rare entities are prevalent (Sinhamahapatra and Niehues, 2025). While previous works have improved ASR accuracy by incorporating lip movement information (Afouras et al., 2022; Ma et al., 2021, 2023; Shi et al., 2022), they ignore the rich semantic context present in presentation slides. In these scenarios, critical keywords and proper nouns are often explicitly displayed on the slides, serving as strong visual cues for the spoken content (Wang et al., 2024a,b). For clarity, we refer to the task of improving ASR accuracy by incorporating visual context from presentation slides as *SlideASR*.

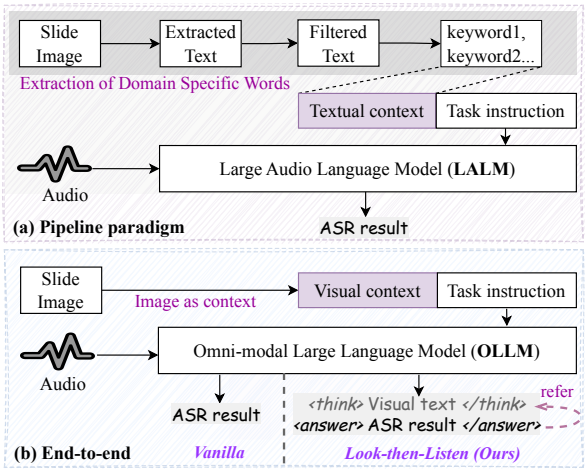


Figure 1: Comparison of paradigms for the SlideASR task. **(a) Pipeline paradigm:** Cascades independent modules which is complex. **(b) End-to-end paradigm:** Illustrates two approaches using an OLLM. The ***Vanilla*** path directly generates the ASR result, often leading to visual interference. In contrast, our ***Look-then-Listen*** path enforces a structured inference chain, explicitly decoupling visual perception from auditory processing.

Currently, the dominant strategy for the SlideASR task is the pipeline paradigm (Wang et al., 2024b). As shown in Fig. 1(a), this paradigm extracts text from slide images and selects domain-specific words, which are then fed as textual context into Large Audio Language Models (LALMs) (Chu et al., 2024; Bai et al., 2024; Dinkel et al., 2025) for ASR. This cascading process in-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9417–9432
July 2-7, 2026

volves multiple modules, suffering from complexity and error accumulation.

Given these limitations, a natural question arises: **Can we establish an end-to-end paradigm** that directly utilizes the slide image as visual context, thus bypassing the complex pipeline? To answer this question, we argue that the recently emerging Omni-modal Large Language Models (OLLMs) (Xu et al., 2025a,b; Yao et al., 2024; Li et al., 2025a) offer a promising solution. OLLMs are capable of simultaneously processing textual, visual, and auditory modalities. Thus, they are inherently well-suited to accomplish the SlideASR task in an end-to-end manner.

However, our preliminary investigation reveals a critical gap between this theoretical potential and practical performance (Sec. 2). Specifically, we observe that current OLLMs suffer from severe **visual interference**. Instead of using the slide image as auxiliary information, the models exhibit a strong tendency to hallucinate, i.e., incorrectly transcribe visible slide text while this text is absent from the spoken audio. We attribute this failure to the **lack of an intermediate reasoning phase.** Considering human behavior, when listening to professional presentations, humans typically adopt a "*Look-then-Listen*" strategy, which temporally decouples the processing of the two modalities. We first scan the slide to establish a contextual prior of the topic and then anchor the auditory input within this context. In contrast, vanilla OLLMs process visual and auditory signals simultaneously (left path in Fig. 1(b)). Lacking this sequential guidance, strong visual cues tend to suppress auditory inputs, resulting in the observed interference.

Inspired by this, we propose **V**isually-**A**nchored **P**olicy **O**ptimization (**VAPO**) which centers on reshaping the model's inference process to align with the human-like "*Look-then-Listen*" workflow (right path in Fig. 1(b)). Specifically, we design a <*think*><*answer*> format to explicitly decompose the multimodal task. In the <*think*> phase, the model is required to first perform Optical Character Recognition (OCR) to establish a visual context prior. Subsequently, in the <*answer*> phase, it generates the transcription by attending to the audio while referencing the extracted content as a reliable anchor. To instantiate this policy, we employ four complementary rewards: a) *Format Reward* to ensure structural compliance; b) *OCR Reward* to promote precise visual perception; c) *ASR Reward* to maintain the overall transcription accuracy; and d) *Visual Anchoring Reward* to encourage the model to effectively leverage key entities identified within the <*think*> phase.

Furthermore, to address data bottlenecks, we construct ***SlideASR-Bench***. Existing datasets, e.g., SlideSpeech (Wang et al., 2024b), primarily focus on general scenarios and lack sufficient entity density. SlideASR-Bench is specifically tailored to promote the utilization of visual cues for specialized terms and comprises two subsets: 1) ***SlideASR-S***, a large-scale synthetic corpus derived from ContextASR-Bench (Wang et al., 2025) providing both training and test sets; and 2) ***SlideASR-R***, a small real-world test set for evaluation in complex presentation environments.

We conduct extensive experiments on SlideSpeech (Wang et al., 2024b) and our proposed SlideASR-Bench. Empirical results demonstrate that our approach significantly outperforms state-of-the-art models, e.g., Qwen3-Omni (Xu et al., 2025b), particularly on entity-related metrics. The main contributions are summarized as follows:

- **Analysis:** We identify visual interference as the primary bottleneck in current OLLMs for SlideASR, revealing why naive end-to-end approaches fail.
- **Method:** We propose VAPO, which reshapes the inference process into a structured, human-like "*Look-then-Listen*" workflow, explicitly temporally decoupling visual perception from auditory transcription to eliminate visual interference.
- **Data:** We construct SlideASR-Bench, comprising synthetic and real-world subsets, to train and evaluate SlideASR task in entity-rich scenarios.

2 Visual Interference of OLLMs in End-to-end SlideASR

Although OLLMs theoretically have the capability to process auditory and visual modalities simultaneously, our preliminary investigation reveals a critical failure mode in their practical application. We present a motivating example in Fig. 2 using Qwen2.5-Omni-7B (Xu et al., 2025a). In the audio-only setting, the model accurately transcribes the speech. However, when the corresponding slide image is introduced, the model's behavior shifts drastically: instead of utilizing the visual information as auxiliary context, it ignores the auditory signal and directly outputs the text visible on the slide.

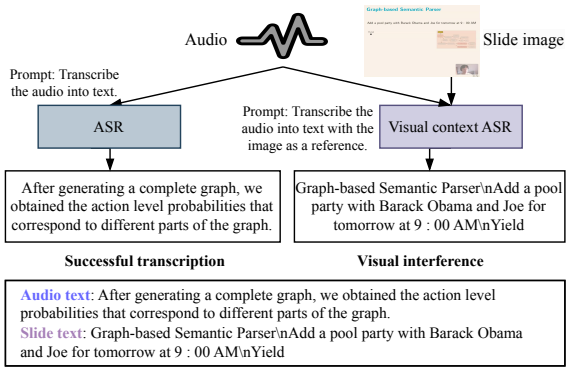


Figure 2: An illustrative example of **Visual Interference** in the end-to-end SlideASR task. **Left:** the model correctly transcribes the speech in audio-only mode. **Right:** the model erroneously outputs the text on the slide when the image is introduced as visual context.

We term this specific failure mode as **Visual Interference**. Formally, it is defined as a phenomenon where the model generates text present in the visual context while it is absent from the speech.

To systematically quantify this phenomenon, we introduce a metric termed the **Visual Interference Rate (VIR)** which measures the tendency of the model to hallucinate text from the slide that was never spoken. We evaluate four representative OLLMs: Qwen2.5-Omni (7B/3B) (Xu et al., 2025a), MiniCPM-o-2.6 (Yao et al., 2024), and Megrez-Omni (Li et al., 2025a) on the SlideSpeech (Wang et al., 2024b) dataset. The calculation follows a set-difference logic:

- **Step 1: Context Isolation.** Let V_{slide} be the set of words on the slide and V_{audio} be the set of words in the ground-truth transcript. We first identify the *slide-exclusive vocabulary* $V_{exclusive} = V_{slide} \setminus V_{audio}$. These are words that appear visually but are not spoken.
- **Step 2: Interference Detection.** Let V_{pred} be the set of words in the model's prediction. We calculate the intersection $I = V_{pred} \cap V_{exclusive}$.
- **Step 3: Metric Calculation.** If $I \neq \emptyset$, the sample is flagged as exhibiting visual interference. The VIR is defined as the percentage of such flagged samples in the dataset.

The quantitative results are presented in Table 1. All evaluated OLLMs exhibit a high VIR, ranging from 12.87% to as high as 63.28% on the test set. Even the strongest baseline, Qwen2.5-Omni-7B, fails to suppress visual interference in over 12% of the samples. This consistent failure across different models indicates that this is a widespread issue rather than an isolated case. This phenomenon highlights a fundamental limitation in current paradigms: OLLMs lack an explicit mechanism to temporally decouple visual perception from auditory processing.

Table 1: Comparison of Visual Interference Rate across different OLLMs on the SlideSpeech dataset.

Model	**Dev set Num=1,801**	**Test set Num=3,053**
MiniCPM-o-2.6	57.96%	63.28%
Megrez-Omni	45.14%	44.90%
Qwen2.5-Omni-3B	15.43%	16.54%
Qwen2.5-Omni-7B	13.71%	12.87%

3 Method

To mitigate the *Visual Interference* issue identified in Sec. 2, we propose Visually-Anchored Policy Optimization (VAPO). Inspired by the human perception process, VAPO addresses this limitation by reshaping the model's inference process to achieve a temporal decoupling of modalities. Specifically, it establishes a structured <*think*><*answer*> inference chain: the model first establishes a visual prior in the <*think*> phase, and then references this content as an anchor to guide speech transcription in the <*answer*> phase. This mechanism mimics the human-like "*Look-then-Listen*" workflow. The overall framework is illustrated in Fig. 3.

3.1 The "Look-then-Listen" Inference Chain

We design a structured output format to explicitly enforce the necessary temporal decoupling between visual perception and auditory processing. Drawing inspiration from Chain-of-Thought (CoT) reasoning (Jaech et al., 2024; Ma et al., 2025), the model is mandated to generate its output within a sequential <*think*><*answer*> structure.

The <*think*> block corresponds to the "Look" phase: the model first processes the visual input and is tasked with extracting textual information from the slide image. This operation establishes a critical visual context prior before the auditory processing begins. Subsequently, the <*answer*> block corresponds to the "Listen" phase: the model generates the final transcription. In this phase, the model is required to reference the content anchored in the <*think*> block. This mechanism allows specialized terms from the slide to serve as semantic anchors, thereby assisting the model in resolving ambiguous audio and enhancing overall transcription accuracy.

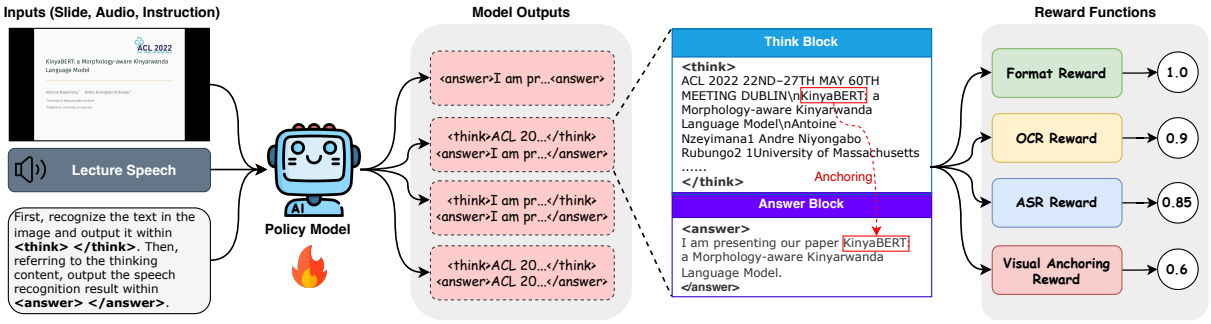


Figure 3: Overview of the Visually-Anchored Policy Optimization (VAPO) framework. The model takes audio, slide, and instruction as inputs and generates a structured <*think*><*answer*> sequence. In the <*think*> phase, the model extracts visual context, which serves as a semantic anchor (indicated by the red box) to guide the transcription in the <*answer*> phase. The policy is optimized via four rewards (Format, OCR, ASR, and Visual Anchoring).

3.2 Multi-Objective Policy Optimization

To make the model follow the proposed inference chain, we design four complementary reward functions to guide the model's learning. We employ the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024b) for model training.

Format Reward. This reward aims to ensure that the model's output strictly adheres to the <*think*><*answer*> format. A positive reward is assigned only if the model generates the complete structural tags. The reward function is defined as:

$$R_{\text{Format}} = \begin{cases} 1, & \text{if the format is correct,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

OCR Reward. This reward evaluates the visual perception quality in the "Look" phase by comparing the text generated in the <*think*> block (T_t) with the ground truth slide text (T_s). We use the Word Error Rate (WER) as the metric, treating each Chinese character as a word where applicable. The reward is normalized and clipped to ensure non-negativity:

$$R_{\text{OCR}} = \max(1 - \text{WER}(T_t, T_s), 0). \quad (2)$$

ASR Reward. This reward assesses the transcription quality in the "Listen" phase by comparing the output in the <*answer*> block (T_a) with the ground truth speech transcription (T_g). Similarly, it is derived from the WER:

$$R_{\text{ASR}} = \max(1 - \text{WER}(T_a, T_g), 0). \quad (3)$$

Visual Anchoring Reward. This reward bridges the "Look" and "Listen" phases by incentivizing the model to ground spoken entities in the visual context. Let E_{target} be the set of critical entities present in both the ground truth slide and the ground truth speech transcript. Let E_{gen} be the subset of E_{target} that successfully appears in both the generated <*think*> and <*answer*> blocks. The reward is defined as recall of these target anchors:

$$R_{\text{VA}} = \begin{cases} \frac{|E_{gen}|}{|E_{target}|}, & \text{if } |E_{target}| > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

This formulation encourages the model to explicitly capture and utilize the visual cues that are relevant to the current speech.

Finally, the total reward is defined as a weighted sum of the four components, where λ_i denotes the weighting hyperparameter for each reward:

$$R_{\text{total}} = \lambda_1 R_{\text{Format}} + \lambda_2 R_{\text{OCR}} + \lambda_3 R_{\text{ASR}} + \lambda_4 R_{\text{VA}}. \quad (5)$$

4 SlideASR-Bench: A Benchmark for Entity-rich SlideASR Task

Our primary goal is to enhance ASR accuracy for domain-specific entities within visual presentation contexts. While existing datasets like SlideSpeech (Wang et al., 2024b) and ChineseLips (Zhao et al., 2025), provide valuable benchmarks for general-domain scenarios, we observe that they often lack a sufficient density of domain-specific named entities. This scarcity creates a significant bottleneck for both training visually-anchored models and evaluating their capabilities in specialized scenarios. To bridge this gap, we introduce SlideASR-Bench, which comprises two distinct subsets: SlideASR-S, a large-scale synthetic corpus for training and evaluation, and

Table 2: Results on the SlideSpeech, a real-world English SlideASR dataset. † represents results from the original paper. The best and second-best results are in **bold** and underlined, respectively.

Model	Dev set				Test set			
	WER↓	B-WER↓	U-WER↓	Recall↑	WER↓	B-WER↓	U-WER↓	Recall↑
Contextless								
Qwen2-Audio	12.56	12.85	8.72	91.43	13.19	13.59	7.53	92.91
MiniCPM-o-2.6	16.09	16.68	8.14	91.98	18.71	19.41	8.90	91.50
Qwen2.5-Omni-3B	15.53	16.22	6.30	93.76	12.00	12.45	5.72	94.41
Qwen2.5-Omni-7B	11.75	12.20	5.39	94.78	11.75	12.20	5.39	94.78
Qwen3-Omni-30B-A3B	10.87	11.31	5.02	95.04	11.71	12.21	4.64	95.50
Slide text as context (Pipeline)								
Qwen2-Audio	139.81	145.05	69.94	85.40	146.08	152.41	56.99	88.98
Mi-Dasheng	33.67	35.18	13.56	93.02	47.21	49.34	17.21	91.00
Qwen3-Omni-30B-A3B	50.43	52.85	18.05	96.45	57.12	59.27	26.75	96.34
LCB-net†	18.80	18.11	27.90	72.09	19.21	18.89	23.70	76.48
MaLa-ASR†	11.14	11.36	8.92	91.44	11.26	11.52	7.67	92.50
Slide image as context (End-to-End)								
MiniCPM-o-2.6	182.96	192.83	51.07	86.26	210.37	220.96	60.92	83.22
Qwen2.5-Omni-3B	12.22	12.74	5.26	95.17	19.99	20.71	9.80	94.44
Qwen2.5-Omni-7B	13.65	14.13	7.19	92.84	14.97	15.58	6.33	93.99
Qwen3-Omni-30B-A3B	19.85	20.64	9.30	95.59	24.13	24.88	13.44	94.74
VAPO-3B (Ours)	9.84	10.31	3.61	96.54	10.73	11.24	3.55	96.57
VAPO-7B (Ours)	**8.62**	**9.08**	**2.48**	**97.61**	**10.31**	**10.84**	**2.87**	**97.32**

SlideASR-R, a small real-world test set for stress-test. Detailed statistics are presented in Table 7.

SlideASR-S. To train and evaluate models for entity-rich scenarios, we constructed SlideASR-S by extending the ContextASR-Bench (Wang et al., 2025) dataset. ContextASR-Bench leverages LLMs, such as DeepSeek-R1 (Guo et al., 2025), to generate colloquial text rich in named entities based on seed text. The seed text is sourced from Named Entity Recognition (NER) datasets across multiple domains (e.g., medicine, culture, and ecology). Text-to-speech models (Du et al., 2024) are then used to convert the generated text into natural and fluent speech.

We extract metadata for each sample from ContextASR-Bench, including the domain label L_{domain} and the list of domain-specific entities E. Using the L_{domain} and E as input, we employ an LLM (e.g., Qwen2.5-14B-Instruct[1] (Team et al., 2024)) to generate a short non-colloquial, formal text in slide style. The prompt guides the LLM to include a title and key points, ensuring that all entities from the original audio are naturally embedded in the generated text. Finally, we utilize Python's Matplotlib library to render the generated text into slide images, generating a total of 8,467 samples (6,413 for training set, 2,054 for test set). We provide the prompt and data example in Appendix A.

SlideASR-R. Furthermore, to assess the model's generalization ability in real, complex environments, we manually constructed a small-scale, high-quality, and challenging test set. We collected 60 real presentation audio clips and corresponding slide images from publicly available academic report videos, covering four specialized domains: chemistry, medicine, biology, and artificial intelligence. For each sample, we manually annotated the data by carefully comparing the speech and slide image, identifying the domain-specific entities that appear in both. We named this dataset SlideASR-R, which contains 200 domain-specific entities from real-world scenarios.

5 Experiment

5.1 Experimental Setup

Implementation Details. We fine-tune Qwen2.5-Omni (3B/7B) on the training set of SlideASR-S using VAPO for a total of 800 training steps. The training employs the AdamW (Loshchilov and Hutter, 2019) optimizer (learning rate $1e^{-6}$, global batch size 32) on 4×A100 GPUs. We set the group size to 4 and use a sampling temperature of 1.0 to encourage exploration during policy updates with

[1]https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

Table 3: Results on the SlideASR-Bench. The best and second-best results are in **bold** and underlined, respectively.

Model	SlideASR-S (en)			SlideASR-S (zh)			SlideASR-R	
	WER↓	NE-WER ↓	NE-FNR↓	WER↓	NE-WER↓	NE-FNR↓	NE-WER↓	NE-FNR↓
Contextless								
Qwen2-Audio	11.90	36.29	47.84	6.02	22.83	40.36	74.56	76.73
MiniCPM-o-2.6	11.19	27.51	30.93	10.35	25.00	41.62	55.85	65.37
Qwen2.5-Omni-3B	8.37	24.15	31.04	4.47	19.89	38.08	61.31	66.83
Qwen2.5-Omni-7B	8.15	23.44	27.77	4.34	17.54	32.80	53.68	63.37
Qwen3-Omni-30B-A3B	9.06	14.61	15.53	20.77	23.31	22.49	40.43	41.09
Slide text as context (Pipeline)								
Qwen2-Audio	92.16	66.38	24.82	39.09	50.58	31.52	59.04	21.29
Mi-Dasheng	78.98	49.85	30.58	66.88	56.30	32.30	47.52	26.73
Qwen3-Omni-30B-A3B	34.65	32.35	8.56	9.76	15.85	13.54	34.01	28.22
Slide image as context (End-to-End)								
MiniCPM-o-2.6	112.90	49.65	15.01	89.53	61.25	45.67	63.73	66.83
Qwen2.5-Omni-3B	100.08	53.19	18.72	86.86	65.62	9.62	49.00	53.47
Qwen2.5-Omni-7B	57.21	35.76	15.04	91.83	54.04	3.36	41.77	35.15
Qwen3-Omni-30B-A3B	101.45	59.64	12.08	79.21	46.45	5.54	32.26	24.75
VAPO-3B (Ours)	4.90	3.19	3.73	2.47	4.21	2.22	27.28	19.31
VAPO-7B (Ours)	**4.60**	**2.83**	**2.97**	**2.13**	**3.78**	**1.36**	**26.48**	**15.35**

a KL penalty coefficient of 0.01. The weights of the reward functions, λ_1 to λ_4, are all set to 1.

Baselines & Settings. We benchmark mainstream LALMs: Qwen2-Audio (Chu et al., 2024) and Mi-Dasheng (Dinkel et al., 2025) and OLLMs: MiniCPM-o-2.6 (Yao et al., 2024) Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b) on SlideSpeech (Wang et al., 2024b) and SlideASR-Bench across three settings:

1. ***Contextless***, which only use audio as inputs;
2. ***Slide text as context*** (Pipeline), employing PaddleOCR (Cui et al., 2025) for textual context extraction following Zhao et al. (2025);
3. ***Slide image as context*** (End-to-end).

We provide more evaluation details in Appendix B. Additionally, we report results on the ChineseLips (Zhao et al., 2025) dataset in Appendix C to demonstrate the generalization of VAPO.

Metrics. We utilize two sets of metrics consistent with prior works. For SlideSpeech (Wang et al., 2024b), we employ four metrics: WER, which measures overall transcription; U-WER, the unbiased WER on non-keyword segments for general transcription quality; B-WER, the biased WER focusing on keyword spans; and Recall, which is the percentage of correctly recognized keywords. Furthermore, for SlideASR-Bench, consistent with ContextASR-Bench (Wang et al., 2025), we focus on three metrics: WER; NE-WER, the WER of the named entity portion; and NE-FNR, the False Negative Ratio of named entities, which measures the proportion of missed ground-truth entities. See Appendix B.2 for specific calculation details of the metrics.

5.2 Main Results and Analysis

Results on SlideSpeech. Table 2 presents the results on the real-world SlideSpeech dataset. Ideally, visual context should enhance recognition; however, we observe a detrimental effect in baseline models. Most baselines exhibit performance degradation when incorporating slide information compared to the contextless setting. For instance, the Qwen3-Omni (Xu et al., 2025b) shows increased WER in both pipeline and end-to-end settings compared to using audio alone. In stark contrast, our VAPO method effectively leverages visual cues, achieving the best performance. VAPO-7B reaches a WER of 10.31 and a Recall of 97.32, significantly outperforming the Qwen3-Omni baseline and the previous SOTA MaLa-ASR (Yang et al., 2024).

Results on SlideASR-Bench. Table 3 reports the performance on our benchmark. Two critical observations highlight the superiority of VAPO:

1) Audio-only models struggle with specialized entities. On the challenging SlideASR-R subset, even the strongest audio-only model (Qwen3-Omni) suffers from a high NE-FNR of 41.09, underscoring the necessity of visual presentation context for domain-specific terms.

2) Naive visual integration leads to catastrophic failure. On SlideASR-S, baseline OLLMs fail to

Table 4: Ablation results of rewards on SlideASR-R.

ASR Reward	OCR Reward	VA Reward	NE-WER ↓	NE-FNR↓
Qwen2.5-Omni-3B				
✘	✘	✘	49.00	53.47
✔	✘	✘	37.23	31.19
✔	✔	✘	29.97	22.28
✔	✔	✔	**27.28**	**19.31**
Qwen2.5-Omni-7B				
✘	✘	✘	41.77	35.15
✔	✘	✘	28.63	20.30
✔	✔	✘	26.75	18.32
✔	✔	✔	**26.48**	**15.35**

Table 5: Sensitivity analysis of reward weights (λ_1:λ_2:λ_3:λ_4) on SlideASR-S. The balanced 1:1:1:1 strategy achieves the best robustness across metrics (N-W: NE-WER, N-F: NE-FNR).

Weights	SlideASR-S (en)			SlideASR-S (zh)		
	WER	N-W	N-F	WER	N-W	N-F
1:1:1:1	**4.90**	**3.19**	**3.73**	**2.47**	**4.21**	2.22
1:1:1:2	5.27	3.34	3.78	2.50	4.30	2.09
1:1:2:1	5.32	4.12	3.91	2.48	4.38	2.09
1:2:1:1	5.17	3.45	3.80	2.51	4.23	**1.99**

utilize visual context effectively. Instead of improving, they suffer from severe visual interference. For example, Qwen3-Omni in the end-to-end setting yields an exploded WER of 101.45 and a worsened NE-WER of 59.64 (vs. 14.61 in audio-only), indicating that the model is hallucinating slide content rather than transcribing speech. Conversely, on SlideASR-S (en), VAPO-3B achieves a remarkable WER of 4.90. Furthermore, on the SlideASR-R, VAPO-7B reduces the NE-FNR from the best baseline's 28.22 to 15.35. These results confirm that VAPO successfully enforces the "*Look-then-Listen*" paradigm, accurately anchoring visual entities without succumbing to interference. We present case study in Appendix D to provide visual comparison.

5.3 Ablation Study

Ablation on Reward Functions. Table 4 details the stepwise contributions on SlideASR-R. Compared to the baseline, introducing the ASR Reward yields the most significant boost, drastically reducing NE-WER by stabilizing generation. Adding the OCR Reward further refines the visual prior quality. Finally, incorporating the Visual Anchoring Reward achieves the best performance. This confirms that while ASR and OCR guarantees individual modality perception, the VA reward is essential

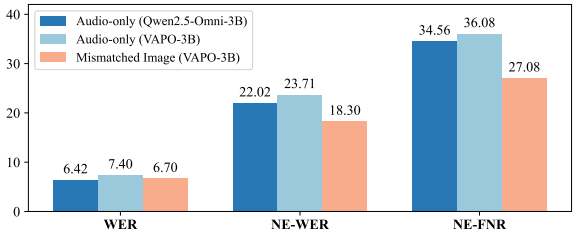


Figure 4: Robustness analysis of VAPO-3B under mismatched visual context on SlideASR-S.

for effective referencing. We provide additional ablation results in Appendix E.

Sensitivity to Reward Weights. To investigate hyperparameter sensitivity, we evaluate varying weight configurations on SlideASR-S, as detailed in Table 5. The results confirm that the balanced 1:1:1:1 scheme yields the most robust overall performance. We observe an inherent trade-off: doubling the VA reward (λ_4=2) enhances entity recall on the Chinese subset (improving NE-FNR to 2.09), but at the cost of increased overall WER due to over-aggressive anchoring. Conversely, double the ASR reward (λ_3=2) proves counterproductive, suppressing the model's reliance on visual cues and significantly degrading NE-WER (e.g., rising from 3.19 to 4.12 on the English subset). Consequently, we adopt the equal weighting strategy for its simplicity and effectiveness.

5.4 Robustness against Mismatched Slides

To verify that VAPO performs valid visual anchoring rather than blindly copying slide content, we evaluated the model's performance when fed with randomly mismatched slide images. As illustrated in Fig. 4, the model exhibits strong robustness: the overall WER remains stable at 6.70%, which is comparable to the audio-only baselines, including Qwen2.5-Omni-3B (6.42%) and the audio-only variant of VAPO-3B (7.40%). This indicates that VAPO effectively mitigates visual interference, safely falling back to auditory perception when visual context is unreliable.

5.5 Impact of Training Paradigm: VAPO vs. Supervised Fine-Tuning (SFT)

To disentangle the contributions of the structured format from the reinforcement learning (RL) optimization, we compare three training settings on SlideASR-S: SFT w/o think (direct mapping), SFT w/ think (structured but SFT-only), and VAPO. Results in Table 6 highlights two key findings:

Audio text: Proscia's Concentriq platform lives at the center of the pathology ecosystem.
Slide text: Concentriq: the center of the digital pathology ecosystem\nCONCENTRIQ by PROSCIA\nAWS

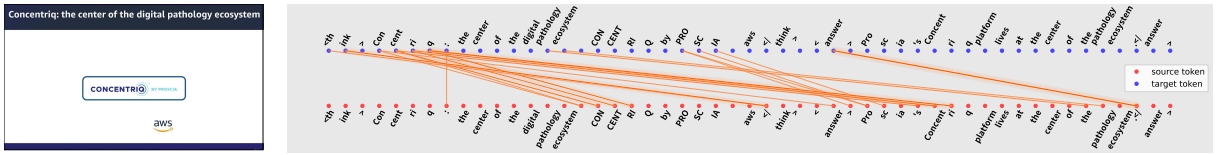


Figure 5: Attention visualization of VAPO-7B. The dense orange connections demonstrate the visual anchoring mechanism, showing that the model explicitly attends to extracted entities in <*think*> to guide the final transcription.

Table 6: Comparison between VAPO and SFT baselines on SlideASR-S. **"SFT w/o think"** is trained to directly predict the transcription, whereas **"SFT w/ think"** is supervised using the structured inference chain.

Model	WER↓	NE-WER ↓	NE-FNR↓
Qwen2.5-Omni-3B			
+ SFT w/o think	8.44	8.18	9.30
+ SFT w/ think	6.60	6.47	7.10
+ VAPO	**3.88**	**3.84**	**3.00**
Qwen2.5-Omni-7B			
+ SFT w/o think	10.58	10.51	12.77
+ SFT w/ think	6.73	5.56	6.53
+ VAPO	**3.37**	**3.07**	**2.00**

1) Efficacy of the Inference Chain. Introducing a structured inference chain yields significant gains, reducing the WER of Qwen2.5-Omni-7B from 10.58 to 6.73. This confirms that temporally decoupling the "Look" and "Listen" phases inherently mitigates visual interference.
2) Necessity of RL. VAPO further lowers the WER to 3.37. This indicates that while SFT teaches the format, it fails to establish a sufficient connection between <*think*> and <*answer*>. VAPO, driven by the multi-dimensional rewards, ensures the <*think*> block is effectively utilized for the final transcription.

5.6 Attention Visualization

To empirically validate the effectiveness of the visual anchoring mechanism, we visualize the attention weights of VAPO-7B on a sample from SlideSpeech, as shown in Fig. 5. The visualization reveals a distinct semantic reference pattern: during the generation of the <*answer*> block, the model heavily attends to the corresponding extracted text within the <*think*> block.

Specifically, take the proper noun "*Concentriq*" as an example. After generating the token "*Concent*" in <*answer*>, the model pays significant attention to the "*ri*" token in <*think*> and subsequently generates it. It then refers to the "*q*" token in the <*think*> block, enabling accurate and complete transcription of "*Concentriq*". A similar process occurs for the entity "*proscia*". This qualitative evidence confirms that VAPO successfully enforces the "*Look-then-Listen*" paradigm, where the model explicitly references the visual context prior. We provide more cases of attention visualization in Appendix F.

6 Related Works

Contextual ASR. The objective of contextual ASR is to incorporate contextual information, including domain labels, entity lists, and conversational history, into the speech recognition system in order to improve the recognition accuracy of domain-specific terminology (Bai et al., 2024; Xiao et al., 2025; Zhou and Li, 2025). Besides textual information, researchers have focused on leveraging visual information to enhance the performance of ASR models. For example, integrating lip movement information during the recognition process (Ma et al., 2023; Rouditchenko et al., 2024; Shi et al., 2022). This study focuses on the SlideASR task (Zhao et al., 2025; Wang et al., 2024b,a), which involves utilizing slide content as contextual information to support the model, given that slides in presentation scenarios generally contain information closely related to the spoken content. Most existing methods for SlideASR are based on the pipeline paradigm (Wang et al., 2024a; Zhao et al., 2025; Wang et al., 2024b; Yu et al., 2024; Yang et al., 2024), which results in relatively complex systems. The objective of this study is to accomplish the task using an end-to-end approach.

Omni-modal Large Language Models. Recently, OLLMs (Hurst et al., 2024; Fu et al., 2025; Yao et al., 2024; Xu et al., 2025a; Li et al., 2025a; Hu et al., 2025; Li et al., 2025b) have emerged, integrating vision, audio, and text by aligning their encoders during training for end-to-end process-

ing. Models such as MiniCPM-o (Yao et al., 2024) and Qwen2.5-Omni (Xu et al., 2025a) have demonstrated strong multimodal performance. Benefiting from the unified modeling capability across visual and audio modalities, they are expected to solve the SlideASR task in an end-to-end manner. However, in practical scenarios, the models exhibit visual interference, for instance, they sometimes reproduce the textual content from the slides instead of generating the expected speech transcription.

Chain-of-Thought Reasoning. CoT reasoning is a breakthrough approach that enhances the reasoning capabilities of LLMs. Recent works (Jaech et al., 2024; Guo et al., 2025) have shown that by using reinforcement learning algorithms (Shao et al., 2024b; Rafailov et al., 2023; Schulman et al., 2017) to encourage models to generate intermediate reasoning steps before producing the final answer, performance on tasks involving arithmetic, commonsense, and symbolic reasoning can be significantly enhanced. This paradigm has also been extended to the multimodal domain (Shao et al., 2024a; Xu et al., 2024; Ma et al., 2025; Lin et al., 2025; Diao et al., 2025), demonstrating the general effectiveness of making reasoning processes explicit.

7 Conclusion

In this work, we uncover a critical *Visual Interference* phenomenon in OLLMs applied to the SlideASR task, where models succumb to modality dominance and ignore auditory inputs. To address this, we introduce Visually-Anchored Policy Optimization (VAPO), a novel post-training method that enforces a human-like "*Look-then-Listen*" inference chain. By leveraging a structured <*think*><*answer*> format and multi-dimensional reward optimization, VAPO effectively decouples visual perception from auditory transcription. Furthermore, we contribute SlideASR-Bench, a comprehensive benchmark designed to address the data scarcity in entity-rich scenarios. Extensive experiments demonstrate that VAPO not only eliminates visual interference but also sets a new state-of-the-art performance, particularly in recognizing specialized domain entities.

8 Limitations

While VAPO demonstrates significant improvements, this work has several limitations.

Task Generalization. Our current approach is highly specialized for leveraging textual information from presentation slides and does not incorporate other visual cues, such as images of entities (e.g., pictures of specific drugs). In the future, we will adapt our "Look-then-Listen" paradigm to handle more diverse multimodal environments, where various visual elements play a crucial role.

Real-World Robustness. While our training relies on the synthetic SlideASR-S dataset, we have validated its effectiveness on three real-world datasets, SlideSpeech, ChineseLips and SlideASR-R. Nonetheless, a subtle domain gap may still exist, as synthetic slides may not fully capture the stylistic diversity and visual noise (e.g., complex diagrams, low-quality images) of all real-world presentations.

Inference Efficiency. The structured reasoning process of VAPO introduces a computational overhead, resulting in higher inference latency compared to models of the same size (detailed in Appendix G). This makes VAPO most suitable for offline applications where accuracy is critical. However, this trade-off between latency and accuracy is often acceptable for offline transcription tasks where precision is paramount. We will explore strategies such as model distillation in our future research to improve efficiency, enabling its use in real-time applications.

9 Ethical Considerations

The primary societal benefit of our work is enhancing accessibility by improving the transcription accuracy of specialized terms, which can significantly aid individuals who are deaf or hard of hearing. However, this technology must be deployed with caution in high-stakes settings, such as medical transcription, where errors could lead to serious consequences. We advocate for responsible development and believe that human oversight is essential for any critical applications.

10 Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2023YFC3310700), the Open Funding Programs of State Key Laboratory of AI Safety, the National Natural Science Foundation of China (No. 2576030), and the Beijing Nova Program under Grant 20250484899.

References

Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2022. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8717–8727.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, Lu Gao, Yi Guo, Minglun Han, et al. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *CoRR*, abs/2407.04675.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

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

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. 2025. Soundmind: Rl-incentivized logic reasoning for audio-language models. *CoRR*, abs/2506.12935.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. 2025. Midashenglm: Efficient audio understanding with general audio captions. *CoRR*, abs/2508.03983.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. VITA-1.5: towards gpt-4o level real-time vision and speech interaction. *CoRR*, abs/2501.01957.

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

Rui Hu, Delai Qiu, Shuyu Wei, Jiaming Zhang, Yining Wang, Shengping Liu, and Jitao Sang. 2025. Investigating and enhancing vision-audio capability in omnimodal large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 7452–7463. Association for Computational Linguistics.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

Boxun Li, Yadong Li, Zhiyuan Li, Congyi Liu, Weilin Liu, Guowei Niu, Zheyue Tan, Haiyang Xu, Zhuyu Yao, Tao Yuan, Dong Zhou, Yueqing Zhuang, Shengen Yan, Guohao Dai, and Yu Wang. 2025a. Megrez-omni technical report. *CoRR*, abs/2502.15803.

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

Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. 2025. Mind with eyes: from language reasoning to multimodal reasoning. *CoRR*, abs/2503.18071.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-avsr: Audio-visual speech recognition with automatic labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7613–7617. IEEE.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *CoRR*, abs/2501.07246.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogério Feris, and James Glass. 2024. Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Supriti Sinhamahapatra and Jan Niehues. 2025. Do slides help? multi-modal context for automatic transcription of conference talks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16111–16121.

Qwen Team et al. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Hao Wang, Shuhei Kurita, Shuichiro Shimizu, and Daisuke Kawahara. 2024a. Slideavsr: A dataset of paper explanation videos for audio-visual speech recognition. *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 129–137.

Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, Shiliang Zhang, and Ming Li. 2024b. Slidespeech: A large scale slide-enriched audio-visual corpus. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11076–11080. IEEE.

He Wang, Linhan Ma, Dake Guo, Xiong Wang, Lei Xie, Jin Xu, and Junyang Lin. 2025. Contextasr-bench: A massive contextual speech recognition benchmark. *CoRR*, abs/2507.05727.

Cihan Xiao, Zejiang Hou, Daniel Garcia-Romero, and Kyu J. Han. 2025. Contextual ASR with retrieval augmented large language model. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *CoRR*, abs/2411.10440.

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. 2025a. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025b. Qwen3-omni technical report. *CoRR*, abs/2509.17765.

Guanrou Yang, Ziyang Ma, Fan Yu, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Mala-asr: Multimedia-assisted llm-based ASR. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800.

Fan Yu, Haoxu Wang, Xian Shi, and Shiliang Zhang. 2024. Lcb-net: Long-context biasing for audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10621–10625. IEEE.

Jinghua Zhao, Yuhang Jia, Shiyao Wang, Jiaming Zhou, Hui Wang, and Yong Qin. 2025. Chinese-lips: A chinese audio-visual speech recognition dataset with lip-reading and presentation slides. *CoRR*, abs/2504.15066.

Shilin Zhou and Zhenghua Li. 2025. Improving contextual ASR via multi-grained fusion with large language models. *CoRR*, abs/2507.12252.

Appendix

A Details of SlideASR-Bench

Statistics of SlideASR-Bench. Table 7 presents the key statistical metrics for SlideASR-Bench, including the number of samples, entities, and hours.

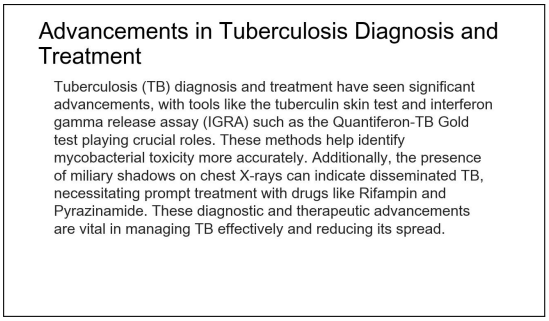
Table 7: Details of our proposed SlideASR-Bench.

Subset	Sample	Entity	Hour
SlideASR-S (Training set)	6,413	44,240	67.3
SlideASR-S (Test set)	2,054	13,895	18.5
SlideASR-R	60	200	0.35

Prompt for Generating SlideASR-S. The prompt for the LLM to generate text paragraphs based on a domain label and an entity list is as follows. Fig. 6 shows an example of SlideASR-S.

Prompt for Qwen2-14B-Instruct to generate slide text

Given a domain label and a list of entities, generate a title and a paragraph for use in a PPT report, with the requirement that the paragraph includes these entities, Keep paragraphs within 150 words.
Domain label:
{ }
List of entities:
{ }
Output format:
###
Title
###
Paragraph



Advancements in Tuberculosis Diagnosis and Treatment

Tuberculosis (TB) diagnosis and treatment have seen significant advancements, with tools like the tuberculin skin test and interferon gamma release assay (IGRA) such as the Quantiferon-TB Gold test playing crucial roles. These methods help identify mycobacterial toxicity more accurately. Additionally, the presence of miliary shadows on chest X-rays can indicate disseminated TB, necessitating prompt treatment with drugs like Rifampin and Pyrazinamide. These diagnostic and therapeutic advancements are vital in managing TB effectively and reducing its spread.

Figure 6: Data example of SlideASR-S.

B Evaluation Details

B.1 Prompts

The prompts for baseline models and our VAPO models are as follows.

Prompt for baseline models

Contextless
Convert the audio to text.

Slide text as context
The speech is the speaker's talk accompanied by a slide, with the text of the slide being: { }
Transcribe the speech into text by integrating the speech with the slide content.

Slide image as context
Taking the image content into account, convert the audio to text.

Prompt for VAPO model

Slide image as context

Role:System
Your task is to convert the speech into text, and the image serves as the reference content related to the speech.
Role:User
First, recognize the text in the image and output it within <think> </think>. Then, referring to the thinking content, output the speech recognition result within <answer> </answer>.

B.2 Metrics

For SlideSpeech, as in the original work (Wang et al., 2024b), we use four metrics

- **WER:** word error rate.
- **U-WER:** unbiased word error rate, computed on non-keyword segments, to evaluate model impact on general transcription.
- **B-WER:** unbiased word error rate, which measures errors on keyword spans.

- **Recall:** keyword recall, the percentage of keywords fully and correctly recognized.

For SlideASR-Bench, we maintain consistency with ContextASR-Bench (Wang et al., 2025) and use the following three evaluation metrics:

- **WER**: word error rate. For Chinese samples, we treat each character as a word.
- **NE-WER**: WER of named entity portion, we first perform a fuzzy match to identify key entities (with an edit distance tolerance of $\frac{2}{WordCountOfEntity} - 1$) in the model's output, and then calculate the WER based on the fuzzy-matched entities.
- **NE-FNR**: The false negative ratio of named entities, calculated as $1 - \frac{H}{N}$, where H and N denote the recognized and ground-truth entity counts.

B.3 Baselines

LALMs. For LALMs, we selected Qwen2-Audio (Chu et al., 2024) and Mi-Dasheng (Dinkel et al., 2025) as baselines, both with 7B parameters. ASR is a core capability of these models. Additionally, they have instruction-following abilities, making them suitable for the context-enhanced ASR task, i.e., *Slide text as context* setting. For SlideSpeech, we additionally selected LCB-net (Yu et al., 2024) and MaLa-ASR (Yang et al., 2024) as baselines. These models were trained on the SlideSpeech training set, and the results for Dev and Test sets are provided (Yang et al., 2024).

OLLMs. For OLLMs, we selected MiniCPM-o-2.6 (Yao et al., 2024), Qwen2.5-Omni-3B (Xu et al., 2025a), Qwen2.5-Omni-7B (Xu et al., 2025a) and Qwen3-Omni-30B-3B (Xu et al., 2025b) as baselines. Similarly, these models not only have ASR capabilities and instruction-following abilities, but they can also directly accept both image and audio as inputs.

C Results on ChineseLips

Table 8 presents the results on ChineseLips (Zhao et al., 2025). Similar to SlideSpeech (Wang et al., 2024b), ChineseLips is a real-world general-domain SlideASR dataset with low entity density both in the speech and the slides. Since ChineseLips does not provide text information for the slides, we report the CER metric on the transcribed text.

The results show that our method achieves the lowest CER, demonstrating its effectiveness in real-world general-domain scenarios.

Table 8: Results on ChineseLips, a real-world Chinese SlideASR dataset. The best and second-best results are in **bold** and underlined, respectively.

Model	CER↓
Contextless	
Qwen2-Audio	12.536
Mi-Dasheng	3.311
MiniCPM-o-2.6	2.252
Qwen2.5-Omni-3B	1.937
Qwen2.5-Omni-7B	2.243
Qwen3-Omni-30B-A3B	2.202
Slide text as context (Pipeline)	
Qwen2-Audio	84.291
Mi-Dasheng	65.505
Qwen3-Omni-30B-A3B	69.172
Slide image as context (End-to-End)	
MiniCPM-o-2.6	76.203
Qwen2.5-Omni-3B	24.847
Qwen2.5-Omni-7B	14.340
Qwen3-Omni-30B-A3B	41.930
VAPO-3B (Ours)	1.548
VAPO-7B (Ours)	**1.298**

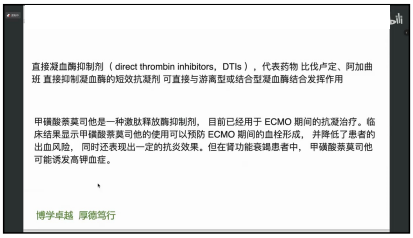
D Case Study

D.1 Successful Case

Fig. 7 shows a comparison of outputs from Qwen2.5-Omni-7B, and Qwen3-Omni-30B-A3B and our proposed VAPO-7B models on samples from the SlideASR-R dataset. Among them, Qwen2.5-Omni-7B uses audio-only input, Qwen3-Omni-30B-A3B uses OCR text extracted from the slide image as context, and VAPO-7B uses the slide image as context input. For Qwen2.5-Omni-7B, due to the lack of auxiliary information, the entity error rate is relatively high. For example, it misrecognized "ConVIRT" as "convert". For Qwen3-Omni-30B-A3B, although slide text is used as context, it fails to utilize it effectively. For example, it also misrecognized "ConVIRT" as "Convert". The VAPO-7B model achieves higher entity recognition accuracy thanks to its "Look before Transcription" reasoning structure.

D.2 Failure Case

Fig. 8 reveals a failure mode of our VAPO-7B model, originating from a visual perception error. The issue begins in the <*think*> block, where the



entity_list
"直接凝血酶抑制剂",
"比伐卢定",
"阿加曲班",
"甲磺酸萘莫司他",
"激肽释放酶抑制剂"

✓ **Qwen2.5-Omni-7B**:
Ti就是直接凝血酶抑制剂，它代表药物就是比伐芦定和阿加曲班啊。那在这个图我们也看到这个绿色的就是这个凝血酶啊，甘素是需要呃也是起到一个催化的作用，是AT和它结合。但是这个呃直接凝血酶抑制剂呢是可以直接和这个凝血酶进行结合，而且不光是游离型的结合型的凝，它都可以跟它结合发挥作用。那还有一个新的药物，就是这个甲磺酸奈莫司他，它是一种激肽释放酶抑制剂，目前已经用于ecmo期间的抗凝治疗。

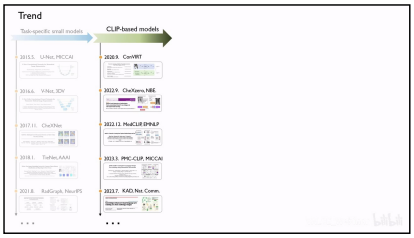
Entity Error Rate = 2/5

✓ **Qwen3-Omni-30B-A3B with OCR**:
ti 就是直接凝血酶抑制剂，它代表药物就是比伐卢定和阿加曲班啊，那在这个图我们已看到这个绿色的就是这个凝血酶，啊，肝素是需要呃也是起到一个催化的作用，是AT和它结合，但是这个呃直接凝血酶抑制剂呢是可以直接和这个凝血酶进行结合，而且不光是游离型的，结合型的凝它都可以跟它结合发挥作用，那还有一个新的

Entity Error Rate = 2/5

✓ **VAPO-7B (Ours)**:
<think>
\n直接凝血酶抑制剂 (direct thrombin inhibitors，DTIs)，代表药物 比伐卢定、阿加曲班 直接抑制凝血酶的短效抗凝剂 可直接与游离型或结合型凝血酶结合发挥作用\n甲磺酸萘莫司他是一种激肽释放酶抑制剂，目前已经用于 ECMO 期间的抗凝治疗。临床结果显示甲磺酸萘莫司他的使用可以预防 ECMO 期间的血栓形成， 并降低了患者的出血风险，同时还表现出一定的抗炎效果。但在肾功能衰竭患者中， 甲磺酸萘莫司他可能诱发高钾血症。\n博学卓越 厚德笃行\n
</think>
<answer>
T I 就是直接凝血酶抑制剂，它代表药物就是比伐卢定和阿加曲班。啊，那在这个图我们也看到这个绿色的就是这个凝血酶，肝素是需要呃也是起到一个催化的作用，是 A T 和它结合。但是这个呃直接凝血酶抑制剂呢是可以直接和这个凝血酶进行结合，而且不光是游离型的，结合型的凝它都可以跟它结合发挥作用。那还有一个新的药物就是这个甲磺酸萘莫司他，它是一种激肽释放酶抑制剂，目前已经用于 E C M O 期间的抗凝治疗。
</answer>

Entity Error Rate = 0/5



entity_list
"ConVIRT",
"CheXzero"

✓ **Qwen2.5-Omni-7B**:
包括就是convert,还有checks是zero,这是发表的nature al medical engineering.然后包括我们自己团队的这个KAD,就是knowledge enhouse ast这个发表在呃nature communication上。后来呢不断的在去进展。

Entity Error Rate = 2/2

✓ **Qwen3-Omni-30B-A3B with OCR**:
...包括就是Convert, 还有 CheX-zero，这是发表在 Nature Biomedical Engineering, 然后包括我们自己团队的这个KAD，就是 knowledge enhanced，这个发表在呃 Nature Communication上。后来呢不断的再去进展。

Entity Error Rate = 1/2

✓ **VAPO-7B (Ours)**:
<think>
Trend\nTask-specific small models\n2015.5. U-Net, MICCAI\n2016.6. V-Net, 3DV\n2017.11. CheXNet\n2018.1. TieNeC, AAAI\n2021.8. RadGraph, NeurIPS\n...\nCLIP-based models\n2020.9. ConVIRT\n2022.9. CheXZero, NBE\n2022.12. MedCLIP, EMNLP\n2023.3. PMM-CLIP, MICCAI\n2023.7. KAD, Nat. Comm.\n...
</think>
<answer>
包括就是 ConVIRT，还有 CheXZero，就是发表在 Nature Biomedical Engineering。然后包括我们自己团队的这个 K A D，就是 knowledge enhanced，这个发表在 Nature Communications 上。后来呢不断地再去进展。
</answer>

Entity Error Rate = 0/2

Figure 7: Qualitative comparison on the SlideASR-R dataset. We compare the transcription results of the baseline (Qwen2.5-Omni-7B), a strong contextual baseline (Qwen3-Omni-30B-A3B with OCR), and our VAPO-7B. **Red text** indicates entity recognition errors (e.g., misrecognizing "CheXzero" as "checks zero"), while **blue text** highlights correctly transcribed entities.

.# entity_list
"中胸腹突",
"狭义花金龟",
"刻花金龟",
"鞘翅",
"侧缘显著内弯"

✓ **VAPO-7B (Ours)**:

<think>
\n(2) 中胸腹突发达（狭义花金龟，频花金龟族和部分巨花金龟族除外）\n(3) 鞘翅侧缘显著内弯（狭义花金龟）\n(4) 体态扁平（广义花金龟，部分斑金龟种类除外）\n
</think>
<answer>
比如说这个中胸腹突十分的发达，你看这个，还有这个，这不是有两个很明显的肩突这儿这儿。然后大家看其他的金龟这儿就看不见了，这儿就是一个空的，这儿就没有了。这个呢是所有的狭义花金龟，当然除了少部分的科花金龟和蝶花金龟以外都有的。第二，第三个特征呢是鞘翅的侧缘显著内弯，像这儿，这儿有。
</answer>

Entity Error Rate = 1/5

Figure 8: A failure case from SlideASR-R. OCR errors occurred due to low image resolution and small entity font size, leading to the loss of correct visual anchor points.

internal OCR component misidentifies a key entity. Specifically, the correct character (pronounced ke) is misrecognized as a different character (pronounced pin). This OCR error is highly plausible because the two characters are visually confusable due to their structural similarity. This type of resemblance is a known challenge for OCR systems, especially with low-resolution text, and it results in the absence of a correct visual anchor.

Due to the lack of a correct visual anchor in the <*think*> block, the entity transcription in the <*answer*> block ends up being incorrect. This case highlights that VAPO's performance is sensitive to low resolution and font size, particularly with visually similar characters. However, this doesn't diminish its overall advantage, as a contextless audio-only model would be equally, if not more, prone to failure when confronted with such inherent ambiguities in the source modalities.

E More Ablation Results on Reward Functions

Table 9 and Table 10 respectively present the ablation results of VAPO-3B (based on Qwen2.5-Omni-3B) on SlideSpeech (Wang et al., 2024b) and SlideASR-S. The results indicate that different reward functions have a positive impact on the final performance.

F More Cases of Attention Visualization

Fig. 9 further presents two cases of attention visualization. It can be seen that when transcribing key entities, the model is able to focus its attention on the same entities in the <*think*> block. This desirable property enables the model to accurately transcribe key entities in the speech.

G Inference Latency Analysis

To evaluate the practical inference efficiency of our proposed VAPO framework, we measured the average inference time per sample on SlideASR-R and compared it against several key baseline models. The results are presented in Table 11.

As shown, our VAPO-7B model has an inference time of 7.27 seconds per sample. This is slower than Qwen2.5-Omni-7B (2.51s), which is expected, as the structured <*think*><*answer*> generation process introduces a computational overhead. However, this moderate increase in latency is accompanied by a dramatic improvement in accuracy, with the NE-FNR dropping from 35.15 to 15.35.

More importantly, when compared to the best baseline model Qwen3-Omni-30B-A3B, our VAPO-7B is significantly more efficient and accurate. While not yet suitable for real-time applications, the latency of VAPO is a reasonable trade-off for its state-of-the-art performance, particularly for offline transcription tasks where accuracy is paramount.

Table 9: Ablation results of functions on the SlideSpeech dataset.

ASR Reward	OCR Reward	VA Reward	Dev set				Test set			
			WER↓	B-WER↓	U-WER↓	Recall↑	WER↓	B-WER↓	U-WER↓	Recall↑
✘	✘	✘	12.22	12.74	5.26	95.17	19.99	20.71	9.80	94.44
✔	✘	✘	10.22	10.68	4.01	96.08	11.02	11.52	3.92	96.17
✔	✔	✘	9.97	10.49	3.83	96.30	11.74	12.25	4.47	95.63
✔	✔	✔	**9.84**	**10.31**	**3.61**	**96.54**	**10.73**	**11.24**	**3.55**	**96.57**

Table 10: Ablation results of the rewards on the SlideASR-S dataset.

ASR Reward	OCR Reward	VA Reward	SlideASR-S (en)			SlideASR-S (zh)		
			WER↓	NE-WER ↓	NE-FNR↓	WER↓	NE-WER↓	NE-FNR↓
✘	✘	✘	100.08	53.19	18.72	86.86	65.62	9.62
✔	✘	✘	5.40	3.78	4.47	2.49	4.33	2.49
✔	✔	✘	4.98	3.71	4.23	2.58	4.54	2.82
✔	✔	✔	**4.90**	**3.19**	**3.73**	**2.47**	**4.21**	**2.22**

Table 11: Comparison of inference time and NE-FNR on the SlideASR-R dataset.

Model	Setting	Inference time per sample (s)	NE-FNR
Qwen3-Omni-30B-A3B	Slide text as context	105.98	28.22
Qwen3-Omni-30B-A3B	Slide image as context	172.95	24.75
Qwen2.5-Omni-7B	Slide image as context	2.51	35.15
VAPO-7B	Slide image as context	7.27	15.35

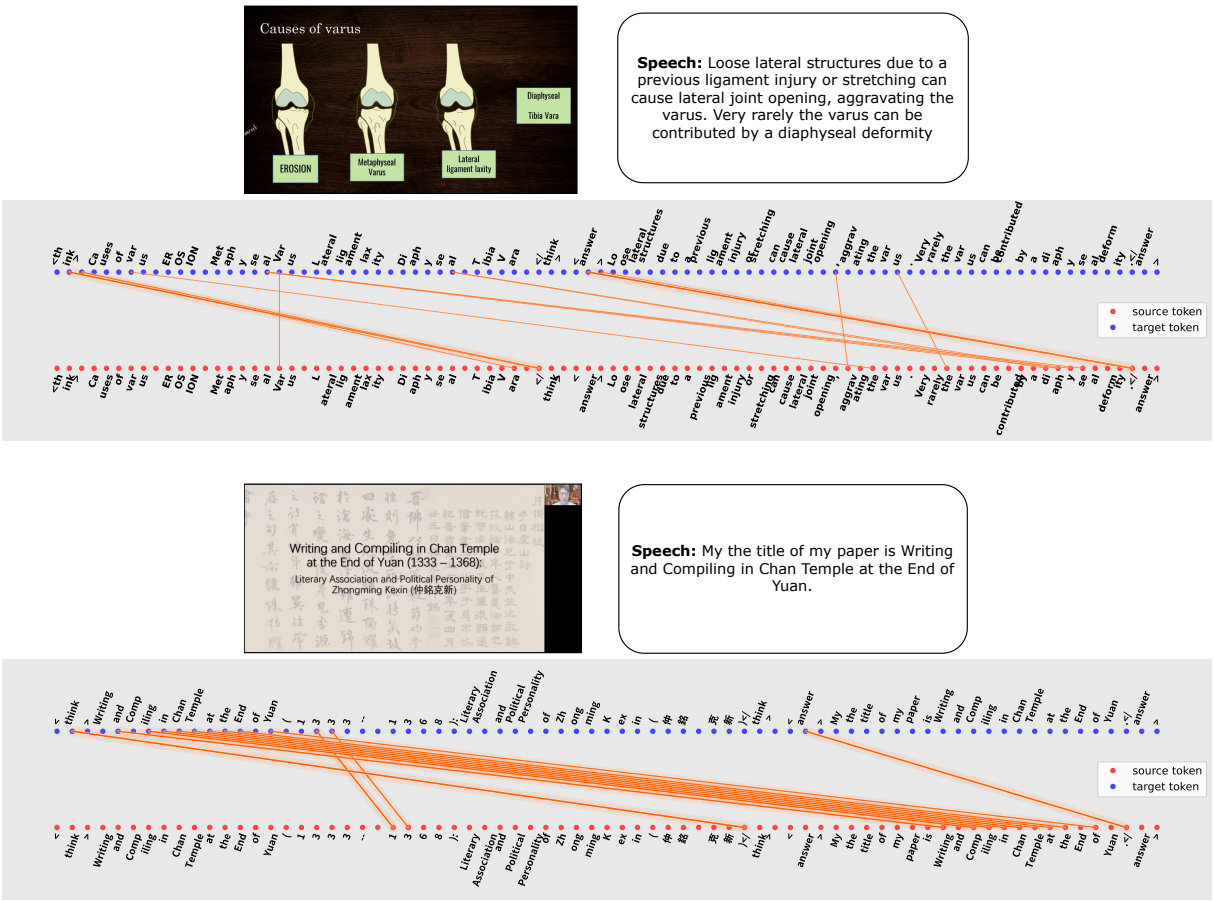


Figure 9: More cases of attention visualization.