

SCREENASR-BENCH: A TIERED BENCHMARK FOR SPEECH RECOGNITION WITH VISUAL CONTEXT

Mingwei Fu¹, Zhesu Song², Zhikai Zhou², Shen Huang², Shidong Shang², Mengxiao Zhu^{1,*}

¹University of Science and Technology of China

²Tenlai Team, China

ABSTRACT

Current ASR models perform well on general benchmarks but still struggle with acoustically ambiguous words. Visual context can aid recognition in screen-based scenarios, yet existing benchmarks provide limited insight into the benefits of different types of visual evidence. To address this gap, we introduce SCREENASR-BENCH, a tiered benchmark with three progressively challenging levels: Explicit Textual Evidence (ETE), based on direct on-screen text; Explicit Visual Evidence (EVE), using visible content rather than on-screen text; and Implicit Semantic Evidence (ISE), requiring scene-level inference. Extensive experiments on ASR models, Large Audio-Language Models (LALMs) and Omni models show that incorporating visual context substantially improves recall over the audio-only setting across all three model families. The gains are largest for ETE and smallest for ISE, indicating that models leverage direct visual evidence more effectively than indirect semantic cues. To facilitate future research on visually grounded speech recognition, we publicly release SCREENASR-BENCH.¹

Index Terms— ASR benchmark, multimodal speech recognition, visual context, keyword recall

1. INTRODUCTION

Recent advances in speech foundation models [1–7] and large-scale datasets have driven substantial progress in automatic speech recognition (ASR), yielding low word error rate (WER) on standard benchmarks [8–12]. However, some ambiguous keywords remain difficult to recognize from speech alone, even when most of the surrounding utterance is transcribed correctly. Such localized errors may contribute little to aggregate WER while substantially changing the intended meaning. In audio-visual settings, on-screen text, visible entities or actions, and scene semantics can provide contextual visual evidence for recognizing these challenging words. This raises a key question: **how effectively can current speech models exploit different levels of visual information for ambiguous keywords recognition?**

Prior studies have shown that visual context can improve recognition when lexical information is difficult to recover from the acoustic signal. Ghorbani et al. [13] demonstrate improved masked-word recovery with visual context, while Gabeur et al. [14] report larger visual gains under challenging acoustic conditions. Chen et al. [15] use on-screen text to enhance ASR, and Zhao et al. [16] combine lip-reading with textual and graphical information from presentation slides. More recently, Tian et al. [17] extend the scope of visual



Fig. 1. Representative examples of the three visual-evidence levels in SCREENASR-BENCH. From ETE to ISE, the semantic distance between the target keyword and its visual evidence increases.

context to richer and more diverse scenes, including scenarios that require multimodal reasoning to resolve acoustic ambiguities. These studies demonstrate the value of visual information for speech recognition, but they generally treat visual context as a whole, without distinguishing between visual cues that differ in how directly they indicate the target keyword.

A similar limitation persists in existing audio-visual ASR benchmarks. SlideSpeech [18] provides a large-scale slide-enriched corpus, while SlideAVSR [19] introduces a benchmark for recognizing technical terminology with presentation slides. Sinhamahapatra et al. [20] similarly evaluate multimodal speech recognition in conference presentations, and VAPO [21] introduces SlideASR-Bench for entity-rich slide-enhanced recognition. These benchmarks provide valuable resources for evaluating visually assisted ASR, but they treat all cases equally and do not distinguish different levels of visual evidence. As a result, it remains unclear which type of visual evidence a model can actually exploit.

To address this gap, we present SCREENASR-BENCH, a tiered benchmark for evaluating how models exploit different forms of visual evidence. It contains 2,002 cases collected from publicly accessible online videos and each case consists of a speech segment, a corresponding keyframe, a reference transcription, and a designated target keyword. As illustrated in Figure 1, cases are categorized into

*Corresponding author.

¹<https://huggingface.co/datasets/MingweiFu/ScreenASR-Bench>

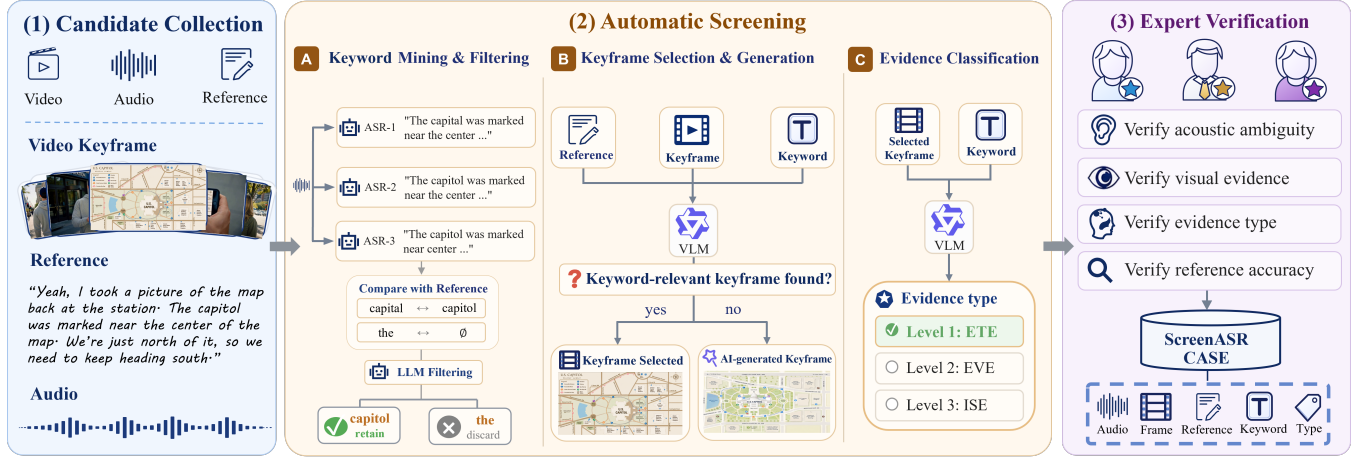


Fig. 2. Overview of the SCREENASR-BENCH construction pipeline, consisting of three stages: (1) candidate collection, (2) automatic screening, and (3) expert verification. Candidate keywords are mined from multi-ASR results and the VLM refers to Qwen3.5-35B-A3B.

three levels according to the relationship between the target keyword and the visual evidence: Explicit Textual Evidence (ETE), Explicit Visual Evidence (EVE), and Implicit Semantic Evidence (ISE), corresponding respectively to readable text, directly visible referents that are not shown as text, and indirect scene-level semantics. From ETE to ISE, the relationship between the target keyword and its visual evidence becomes progressively less direct, increasing the difficulty of effectively leveraging the visual context. In summary, the contributions of this paper are as follows:

- We present SCREENASR-BENCH, a tiered benchmark for evaluating how effectively speech models exploit different levels of visual evidence for challenging lexical recognition.
- ASR models, LALMs, and Omni models are systematically evaluated on SCREENASR-BENCH, providing a comprehensive analysis of their ability to leverage visual context across the three evidence categories.
- We publicly release SCREENASR-BENCH to facilitate reproducible research and promote further advances in multimodal ASR.

2. SCREENASR-BENCH

In this section, we describe the construction pipeline of SCREENASR-BENCH, define the three levels of visual evidence and analyze the statistics of the benchmark.

2.1. Dataset Construction Pipeline

Constructing SCREENASR-BENCH requires both sufficient scale for reliable evaluation and sufficient annotation accuracy for use as ground truth; manual annotation alone does not scale, while fully automatic construction does not ensure label reliability. Therefore, we adopt a three-stage pipeline combining automatic screening with expert verification, as shown in Figure 2.

Stage 1: Candidate Collection. SCREENASR-BENCH is built from publicly available online videos. Each candidate instance originates from a video segment and consists of an audio segment, its reference transcription, and a set of candidate keyframes. The keyframes are sampled from the video according to the timestamps of the audio

segment. Since some of the reference transcriptions are provided in inverse text normalization (ITN) form, we apply an LLM-based text normalization (TN) step to convert all references into a unified textual format.

Stage 2: Automatic Screening. This stage converts raw candidates into keyword-keyframe pairs with preliminary evidence labels, and proceeds in three steps.

(i) *Keyword mining and filtering.* We transcribe each audio segment with multiple ASR systems and align their hypotheses with the normalized reference; the words or phrases where the hypotheses disagree with the reference are taken as candidate keywords. Using multiple recognizers instead of a single one prevents the benchmark from being biased toward the error patterns of any specific model, so that the mined keywords are those that are difficult for ASR systems in general rather than for one particular system. Since many mismatches involve filler words or common function words, we further use an LLM to filter the candidates and keep only content keywords, such as named entities and domain-specific terms.

(ii) *Keyframe selection and generation.* Given the reference text, the mined keyword, and the candidate keyframes, a vision language model (VLM) selects the frame that provides the strongest visual evidence for the keyword. In some cases, none of the sampled frames contains information relevant to the keyword; for these cases, we synthesize a substitute keyframe using GPT Image 2.0 [22]. Synthesized frames account for about 10% of the final benchmark.

(iii) *Evidence classification.* Given the keyword and its selected keyframe, the VLM further determines in what way the keyframe supports the keyword, and assigns one of the three evidence levels (ETE, EVE, and ISE). Qwen3.5-35B-A3B [23] is employed as the VLM in both (ii) and (iii).

Stage 3: Expert Verification. As the annotations produced by the automatic stages are not always reliable, all remaining cases are manually verified by human annotators. Each case is independently reviewed along four aspects: the accuracy of the reference transcription, the difficulty of recognizing the keyword from audio alone, the visual relevance of the selected keyframe, and the correctness of the assigned evidence level. A case is retained only if the keyword is relatively difficult to recognize from the audio alone, and the keyframe provides clear visual evidence for it. Finally, each retained case consists of an audio segment, a keyframe, a keyword, an evidence level, and a verified reference transcription.

Table 1. Model-level WER and Recall. Vision denotes the visual input: – for audio-only, Caption for a VLM-generated keyframe caption, and Image for direct keyframe input. ETE, EVE, and ISE are the three evidence levels. Best and second-best results are bold and underlined.

Model	Vision	ETE		EVE		ISE		Overall	
		WER	Recall	WER	Recall	WER	Recall	WER	Recall
ASR Models									
FireRedASR2-LLM	-	7.02	35.11	5.34	43.07	6.15	34.62	6.22	37.87
MiMo-V2.5-ASR	-	6.97	47.07	4.80	59.28	5.22	54.99	5.77	53.43
Qwen3-ASR-1.7B	-	7.88	34.10	5.99	42.52	6.47	39.10	6.87	38.37
	Caption	6.08	77.86	5.34	52.77	6.03	44.40	5.81	60.58
Qwen-Audio-3.0-ASR	-	7.43	45.55	5.37	58.45	6.00	53.16	6.36	52.08
	Caption	6.80	47.07	4.81	60.53	5.57	54.18	5.80	53.68
Seed-ASR-2.0	-	5.68	48.28	4.01	62.22	4.57	51.43	4.82	54.09
	Caption	4.19	82.57	<u>3.50</u>	70.74	4.20	58.28	3.96	72.34
Large Audio-Language Models									
MOSS-Audio-8B-Instruct	-	7.19	40.71	5.07	51.11	9.32	46.44	7.01	45.87
	Caption	5.20	81.42	4.75	58.59	5.59	50.10	5.15	65.48
MiMo-Audio-7B-Instruct	-	9.04	43.00	6.84	51.66	14.28	45.42	9.63	46.72
	Caption	6.90	82.06	6.04	60.25	6.56	51.12	6.52	66.58
StepAudio-2.5-Chat	-	9.39	45.93	6.57	52.77	7.51	48.88	7.94	49.12
	Caption	9.46	71.25	6.95	56.09	8.25	51.53	8.29	60.93
Omni Models									
Qwen3-Omni	-	6.81	56.23	4.69	66.62	5.32	59.47	5.70	60.78
	Image	5.85	80.41	4.26	75.90	6.57	64.56	5.49	74.89
	Caption	4.47	79.13	<u>3.50</u>	73.68	<u>3.98</u>	66.60	4.01	74.09
MiMo-V2.5	-	11.39	62.98	5.69	68.42	10.68	63.95	9.25	65.18
	Image	6.72	<u>88.68</u>	8.32	77.01	6.10	71.28	7.11	80.19
	Caption	5.35	90.20	4.27	74.93	4.79	69.86	4.83	<u>79.69</u>
Qwen3.5-Omni-Plus	-	5.97	62.09	4.24	69.81	4.98	65.99	5.12	65.83
	Image	5.10	81.81	3.51	78.09	4.50	73.12	4.40	78.33
	Caption	3.68	84.86	3.00	<u>77.95</u>	3.76	<u>72.51</u>	3.47	79.33
Doubao-Seed-2.0-lite	-	6.85	53.56	4.81	58.31	5.44	57.03	5.79	56.13
	Image	5.23	84.27	4.09	71.33	4.65	66.80	4.69	75.29
	Caption	<u>4.17</u>	85.68	<u>3.50</u>	72.58	4.19	66.19	<u>3.94</u>	76.14

2.2. Visual Evidence Levels

We categorize the cases in SCREENASR-BENCH into three levels according to the relationship between the keyword and the visual evidence, with examples shown in Figure 1. From Level 1 to Level 3, the connection between the keyword and its visual evidence becomes progressively less direct, which we call increasing semantic distance. This division allows us to examine how model performance changes as the visual evidence becomes harder to exploit.

Level 1 – Explicit Textual Evidence (ETE). The keyword appears directly as readable text in the frame, so it can be obtained by reading the text. In Figure 1(a), the word “capitol” is printed on the map, which distinguishes it from the acoustically similar “capital.”

Level 2 – Explicit Visual Evidence (EVE). The keyword is no longer written in the frame, but the entity, object or action it refers to is directly shown. Identifying the keyword therefore requires recognizing the visual content and linking it to the spoken context, rather than simply reading text. In Figure 1(b), the open landscape seen from the rooftop supports “plain” rather than “plane”.

Level 3 – Implicit Semantic Evidence (ISE). The frame contains neither the keyword nor anything that directly corresponds to it, and the keyword can only be inferred from the scene as a whole, the domain it belongs to, or the related semantic context. This level therefore requires scene understanding and cross-modal reasoning. In Figure 1(c), the frame shows a League of Legends match, and

recognizing this domain favors the in-game ability name over the acoustically similar alternative.

2.3. Dataset analysis

Figure 3 summarizes the statistics of SCREENASR-BENCH, which contains 2,002 cases. Most audio segments are short, with 91.82% shorter than 20 seconds, a range comparable to that commonly used in ASR evaluation. For the evidence levels, ETE, EVE, and ISE account for 39.31%, 36.11%, and 24.58% of the cases, respectively. ISE is the least frequent, as implicit semantic evidence is harder to collect than evidence explicitly presented in the keyframe, but all three levels still receive substantial coverage.

3. EXPERIMENTS

3.1. Experimental Setup

Experiments cover three model families: ASR systems, large audio language models (LALMs), and Omni models. The ASR systems are FireRedASR2-LLM [24], MiMo-V2.5-ASR [25], Qwen3-ASR-1.7B [26], Qwen-Audio-3.0-ASR [27], and Seed-ASR-2.0 [28]. The LALMs are MOSS-Audio-8B-Instruct [29], MiMo-Audio-7B-Instruct [30], and StepAudio-2.5-Chat [31]. The Omni models are Qwen3-Omni [32], MiMo-V2.5 [25], Qwen3.5-Omni-Plus [33], and

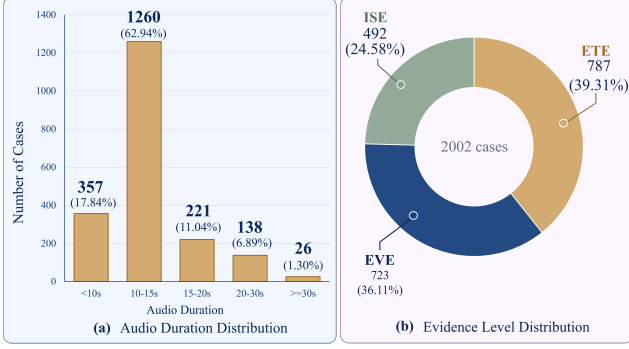


Fig. 3. Statistics of the SCREENASR-BENCH.

Doubao-Seed-2.0-lite [34].

All models are evaluated with speech alone, and models that support context are further evaluated with visual context in two forms: a textual caption of the keyframe generated by Qwen3.5-35B-A3B [23], and the keyframe image itself. ASR systems and LALMs take only the caption, whereas omni models take both.

3.2. Evaluation Metrics

System performance is evaluated using WER, Recall, and normalized visual gain (NVG). WER is used to measure overall transcription accuracy, whereas Recall is used to evaluate the recognition of annotated keywords:

$$\text{Recall} = \frac{N_{\text{hit}}}{N_{\text{total}}}, \quad (1)$$

where N_{total} is the total number of annotated keywords and N_{hit} denotes the number of keywords correctly recognized by the model.

To quantify the improvement brought by visual context, we further introduce NVG:

$$\text{NVG} = \frac{\text{Recall}_{A+V} - \text{Recall}_A}{1 - \text{Recall}_A}, \quad (2)$$

where Recall_A and Recall_{A+V} denote recall under the audio-only and audio-visual conditions, respectively. An NVG of 0 indicates no improvement over the audio-only baseline, whereas an NVG of 1 indicates that perfect recall is achieved under the audio-visual condition. Negative values indicate that keyword recognition is degraded by the incorporation of visual context.

3.3. Main Results

Table 1 reports experimental results on three model families and Table 2 summarizes visual-context utilization across three evidence levels. Several general observations can be drawn from these results.

Beyond WER: Keyword Recall. Under the audio-only condition, Seed-ASR-2.0 achieves the lowest WER of 4.82% but only 54.09% recall. In contrast, omni models generally achieve higher recall, with Qwen3.5-Omni-Plus reaching 65.83%. These results show that lower WER does not necessarily indicate better recognition of target keywords. Since this benchmark focuses on target keywords, recall directly measures whether they are correctly recognized.

Benefits of visual context. Across all three model families, incorporating visual context consistently improves keyword recognition. Specifically, average recall increases from 48.18% to 62.20% for ASR models, from 47.24% to 64.33% for LALMs, and from 61.98% to 77.24% for omni models. At the same time, average

Table 2. Level-wise visual-context utilization. Recall, ΔRecall , and NVG are averaged within each family model; ΔRecall denotes the absolute recall gain.

Family	Level	Recall (%)		ΔRecall	NVG(%)
		Audio-only	Audio + Vision		
ASR	ETE	42.64	69.17	26.53	45.16
	EVE	54.40	61.35	6.95	15.13
	ISE	47.90	52.29	4.39	8.33
LALM	ETE	43.21	78.24	35.03	61.34
	EVE	51.85	58.31	6.46	13.37
	ISE	46.91	50.92	4.01	7.49
Omni	ETE	58.72	84.38	25.66	62.24
	EVE	65.79	75.18	9.39	27.08
	ISE	61.61	68.87	7.26	18.88

WER decreases from 6.02% to 5.19%, from 8.19% to 6.65%, and from 6.47% to 4.74%, respectively. These results demonstrate that visual context provides effective cues for recognizing challenging keywords while generally improving overall transcription accuracy.

Effects of visual input representation. The effectiveness of visual context may depend on how the visual information is represented. For Qwen3-Omni, caption and image inputs achieve comparable overall recall (74.09% vs. 74.89%), while caption input yields a substantially lower WER (4.01% vs. 5.49%). The other omni models exhibit a similar pattern. This suggests that textualized visual information can be more effectively used for transcription. A possible explanation is that captions provide compact information closely aligned with the speech content, whereas raw images may introduce irrelevant visual information. More effective use of raw visual context therefore remains an important direction for omni models.

Visual gains decrease with semantic distance. As shown in Table 2, visual context improves recall across all three model families, but the magnitude of the improvement varies substantially across evidence levels. The absolute recall gain is consistently highest for ETE, followed by EVE and ISE. For ASR models, the average recall gain decreases from 26.53% for ETE to 6.95% for EVE and 4.39% for ISE, with the same pattern observed for LALMs and omni models. Mean NVG follows a similar level-wise trend across all three families, confirming that both absolute and normalized gains diminish as the semantic distance between the visual evidence and the target term increases. These results indicate that visual context is more effective when the evidence is explicit and directly related to the keyword, while current models struggle to exploit implicit visual cues, highlighting the need to improve their ability to reason over and integrate such evidence.

4. CONCLUSION

We introduce SCREENASR-BENCH, a human-verified, tiered benchmark for evaluating how effectively speech models leverage visual context to recognize challenging keywords. It covers three levels of visual evidence: ETE, EVE, and ISE. Across ASR models, LALMs, and omni models, visual context consistently improves keyword recall. However, the larger gains on ETE than on EVE and ISE show that current models exploit on-screen text more effectively than indirect visual evidence, highlighting the need for stronger visual grounding and reasoning. We hope SCREENASR-BENCH will provide a useful foundation for future research on multimodal ASR.

5. REFERENCES

- [1] Jinyu Li, “Recent advances in end-to-end automatic speech recognition,” *arXiv preprint arXiv:2111.01690*, 2021.
- [2] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, et al., “SALMONN: Towards generic hearing abilities for large language models,” in *Proc. ICLR*, 2024, pp. 16607–16629.
- [3] Ziyang Ma, Guanrou Yang, Wenxi Chen, Zhifu Gao, Yexing Du, Xiquan Li, et al., “SLAM-LLM: A modular, open-source multimodal large language model framework and best practice for speech, language, audio and music processing,” *IEEE Journal of Selected Topics in Signal Processing*, 2026.
- [4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. ICML*, 2023, pp. 28492–28518.
- [5] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, et al., “Qwen2-Audio technical report,” *arXiv preprint arXiv:2407.10759*, 2024.
- [6] Zhesu Song, Lu Wang, Wei Deng, Zhuo Yang, Yong Wu, and Bin Xia, “Index-ASR technical report,” *arXiv preprint arXiv:2601.00890*, 2025.
- [7] Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, et al., “FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs,” *arXiv preprint arXiv:2407.04051*, 2024.
- [8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [9] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, “AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *Proc. Oriental COCOSA*, 2017, pp. 1–5.
- [10] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, “AISHELL-2: Transforming mandarin ASR research into industrial scale,” *arXiv preprint arXiv:1808.10583*, 2018.
- [11] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, et al., “MLS: A large-scale multilingual dataset for speech research,” in *Proc. Interspeech*, 2020, pp. 2757–2761.
- [12] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, et al., “FLEURS: Few-shot learning evaluation of universal representations of speech,” in *Proc. SLT*, 2023, pp. 798–805.
- [13] Shahram Ghorbani, Yashesh Gaur, Yu Shi, and Jinyu Li, “Listen, look and deliberate: Visual context-aware speech recognition using pre-trained text-video representations,” in *Proc. SLT*, 2021, pp. 621–628.
- [14] Valentin Gabeur, Paul Hongsuck Seo, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid, “AVATAR: Unconstrained audiovisual speech recognition,” in *Proc. Interspeech*, 2022, pp. 2818–2822.
- [15] Junli Chen, Changli Tang, Yixuan Li, Guangzhi Sun, and Chao Zhang, “OCR-enhanced multimodal ASR can read while listening,” in *Proc. ICASSP*, 2026, pp. 18482–18486.
- [16] Jinghua Zhao, Yuhang Jia, Shiyao Wang, Jiaming Zhou, Hui Wang, and Yong Qin, “Chinese-LiPS: A chinese audio-visual speech recognition dataset with lip-reading and presentation slides,” in *Proc. ICME*, 2025, pp. 1–6.
- [17] Wenjie Tian, Mingchen Shao, Bingshen Mu, Xuelong Geng, Chengyou Wang, Yujie Liao, et al., “Seeing the context: Rich visual context-aware speech recognition via multimodal reasoning,” *arXiv preprint arXiv:2603.07263*, 2026.
- [18] Haoxu Wang, Fan Yu, Xian Shi, Yue Zhang Wang, Shiliang Zhang, and Ming Li, “SlideSpeech: A large-scale slide-enriched audio-visual corpus,” in *Proc. ICASSP*, 2024, pp. 11076–11080.
- [19] Hao Wang, Shuhei Kurita, Shuichiro Shimizu, and Daisuke Kawahara, “SlideAVSR: A dataset of paper explanation videos for audio-visual speech recognition,” in *Proc. ALVR*, 2024, pp. 129–137.
- [20] Supriti Sinhamahapatra and Jan Niehues, “Do slides help? multi-modal context for automatic transcription of conference talks,” in *Proc. EMNLP*, 2025, pp. 16111–16121.
- [21] Rui Hu, Delai Qiu, Yining Wang, Shengping Liu, and Jitao Sang, “VAPO: End-to-end slide-enhanced speech recognition with omni-modal large language models,” in *Proc. ACL*, 2026, pp. 9417–9432.
- [22] OpenAI, “System Card: ChatGPT Images 2.0 and Thinking Mode,” Apr. 2026.
- [23] Qwen Team, “Qwen3.5: Towards native multimodal agents,” Official model documentation, 2026.
- [24] Kaituo Xu, Yan Jia, Kai Huang, Junjie Chen, Wenpeng Li, Kun Liu, et al., “FireRedASR2S: A state-of-the-art industrial-grade all-in-one automatic speech recognition system,” *arXiv preprint arXiv:2603.10420*, 2026.
- [25] Xiaomi MiMo Team, “MiMo-V2.5,” Official model card/documentation, 2026.
- [26] Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, et al., “Qwen3-ASR technical report,” *arXiv preprint arXiv:2601.21337*, 2026.
- [27] Chuanmeng Bian, Daren Chen, Peixin Chen, Zhigao Chen, Zhiyun Fan, Zhifu Gao, et al., “Qwen-Audio-3.0-ASR technical report,” *arXiv preprint arXiv:2609.07549*, 2026.
- [28] ByteDance Seed / Volcano Engine, “Seed-ASR 2.0,” Official model documentation, 2026.
- [29] Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, et al., “MOSS-Audio technical report,” *arXiv preprint arXiv:2606.01802*, 2026.
- [30] Xiaomi LLM-Core Team, “MiMo-Audio: Audio language models are few-shot learners,” *arXiv preprint arXiv:2512.23808*, 2025.
- [31] Bin Lin, Bo Zhao, Boyong Wu, Chao Yan, Chen Wu, Cheng Yi, et al., “Stepaudio 2.5 technical report,” *arXiv preprint arXiv:2605.23463*, 2026.
- [32] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, et al., “Qwen3-Omni technical report,” *arXiv preprint arXiv:2509.17765*, 2025.
- [33] Qwen Team, “Qwen3.5-Omni technical report,” *arXiv preprint arXiv:2604.15804*, 2026.
- [34] ByteDance Seed, “Seed2.0 model card: Towards intelligence frontier for real-world complexity,” *arXiv preprint arXiv:2607.00248*, 2026.