Two Heads Are Better Than One: Audio-Visual Speech Error Correction with Dual Hypotheses

Sungnyun Kim[1]*, **Kangwook Jang**[2]*, **Sungwoo Cho**[1], **Joon Son Chung**[2], **Hoirin Kim**[2], **Se-Young Yun**[1]

[1]Kim Jaechul Graduate School of AI, KAIST
[2]School of Electrical Engineering, KAIST

{ksn4397, dnrrkdwkd12, peter8526, joonson, hoirkim, yunseyoung}@kaist.ac.kr

Abstract

This paper introduces a new paradigm for generative error correction (GER) framework in audio-visual speech recognition (AVSR) that reasons over modality-specific evidences directly in the language space. Our framework, **DualHyp**, empowers a large language model (LLM) to compose independent N-best hypotheses from separate automatic speech recognition (ASR) and visual speech recognition (VSR) models. To maximize the effectiveness of DualHyp, we further introduce **RelPrompt**, a noise-aware guidance mechanism that provides modality-grounded prompts to the LLM. RelPrompt offers the temporal reliability of each modality stream, guiding the model to dynamically switch its focus between ASR and VSR hypotheses for an accurate correction. Under various corruption scenarios, our framework attains up to 57.7% error rate gain on the LRS2 benchmark over standard ASR baseline, contrary to single-stream GER approaches that achieve only 10% gain. To facilitate research within our DualHyp framework, we release the code and the dataset comprising ASR and VSR hypotheses at https://github.com/sungnyun/dualhyp.

1 Introduction

Recent advancements have introduced GER frameworks that utilize LLMs to refine ASR outputs. Following the release of N-best ASR hypotheses dataset (Chen et al., 2023a), numerous studies demonstrated the efficacy of LLMs in correcting transcriptions based on the hypotheses list (Hu et al., 2024a,b; Mu et al., 2024, 2025). These powerful correction frameworks, however, presents a fundamental limitation. While the performance of the underlying ASR systems is remarkable in controlled environments (Chiu et al., 2022; Graves, 2012; Peng et al., 2024; Zhang et al., 2020), it degrades significantly in noisy real-world conditions where acoustic distortions are prevalent. To mitigate this challenge, AVSR systems have been developed (Chen et al., 2023b; Han et al., 2024; Kim et al., 2025b, 2024b; Shi et al., 2022a), leveraging complementary visual cues (*e.g.*, lip movements) to enhance robustness against noise.

In the realm of AVSR, integrating visual information into GER frameworks remains a nascent area of research. Existing methods often employ visual adapters (Ghosh et al., 2024) or unified AVSR models (Liu et al., 2025a), both of which process visual data in the feature space. This feature-level fusion struggles when audio and visual streams are corrupted independently, as noise from one modality can easily contaminate the unified representation (Kim et al., 2025a). Moreover, these frameworks heavily rely on a single set of hypotheses generated from one, often error-prone, recognition model.

To address these limitations, we propose **DualHyp**, the first GER framework that explicitly maintains modality-specific pathways from separate ASR and VSR systems (§3). LLM intelligently composes these **dual-stream hypotheses**, leveraging the model's deep contextual understanding in the *language space* rather than forcing the model to interpret complex audio or video embedding subspaces. Building upon this, we introduce **RelPrompt**, a **noise-aware guidance** mechanism that directs the underlying quality of each modality (§4). Since LLMs for GER primarily operate within the language space, they lack modality-level grounding and may incorrectly prioritize unreliable sources. To mitigate this, we incorporate reliability predictors to assess the quality of audio and visual streams, which are fed to the LLM to better elicit the compositional capacity of DualHyp.

Our experiments (§5) show that this DualHyp approach with RelPrompt significantly outperforms prior single-stream GER frameworks across various audio-visual corruption scenarios. We also demonstrate its multilingual capabilities as well as

*Equal contribution

Findings of the Association for Computational Linguistics: ACL 2026, pages 544–564

improved reasoning with larger LLMs. Through qualitative analysis (§6), we investigate the correction mechanism that makes our framework more effective.

2 Related Works

Generative error correction for speech. Recently, there has been growing interest in using LLMs for post-hoc correction of speech recognition outputs. Initial work in GER for ASR demonstrates that LLMs can effectively regenerate transcriptions from N-best hypothesis lists (Chen et al., 2023a). Subsequent research has refined this paradigm by exploring novel prompting strategies like cloze-style completion (Hu et al., 2024a) or by re-injecting acoustic features to better ground the LLM's corrections (Chen et al., 2024; Liu et al., 2025b; Mu et al., 2024, 2025; Radhakrishnan et al., 2023). These foundational works, however, focus exclusively on correcting hypotheses generated from a single, audio-only stream.

Modality fusion in GER for AVSR. Extending GER to the audio-visual domain presents the central challenge of how to effectively fuse multimodal information. Existing approaches perform this fusion in the feature space, before the final language generation step. Ghosh et al. (2024) involved visual adapters (Houlsby et al., 2019; Zhang et al., 2024b) to inject lip-reading features directly into the LLM, while Liu et al. (2025a) used dedicated multimodal encoders to create a unified audio-visual representation. While these methods show promise, their reliance on early, feature-level fusion makes them vulnerable to cross-modal contamination (Hong et al., 2022), where corruption in one modality can degrade the quality of the fused representation.

Motivated by prior works highlighting the benefits of modality-specific processing for robustness (Kim et al., 2025a; Liu et al., 2021; Wang et al., 2024), our approach is designed to isolate corruptions specific to each modality before error correction. In contrast to feature-level fusion methods, we achieve this by deliberately delaying the modality fusion to the generation stage where the LLM operates on independent textual hypotheses from separate ASR and VSR models.

End-to-end LLM-based AVSR. It is important to distinguish our GER framework from an orthogonal line of research that uses LLMs for end-to-end (E2E) ASR (Fathullah et al., 2024; Ma et al., 2024; Yu et al., 2024) and AVSR (Cappellazzo et al., 2025a,b,c; Yeo et al., 2025, 2024). In that paradigm, encoded audio and visual features serve as direct, multimodal prompts for a single generative model. While promising, our decoupled approach offers significant advantages in flexibility.

First, our framework is highly modular and can readily use off-the-shelf ASR systems and LLMs. This contrasts with monolithic E2E models, which require costly pretraining of the entire system for any component update. Second, the system can be easily improved by refining text-based prompts. This avoids the inherent complexity of designing and aligning cross-modal prompts, which is a central challenge in E2E systems.

3 DualHyp Framework

3.1 Uni-modal Generative Error Correction

Recent works have successfully employed LLMs for GER (Chen et al., 2023a; Ghosh et al., 2024; Hu et al., 2024b), where they aim to refine outputs of a uni-modal ASR system. Given an input utterance, the ASR model first generates an N-best list of candidate transcriptions by beam search decoding, denoted as $\mathcal{H}^{\text{asr}} = \{(h_i^a, s_i^a)\}_{i=1}^N$, where h_i^a is the i-th hypothesis and s_i^a is the corresponding log-likelihood score. The LLM takes this hypothesis set as an input and generates a corrected transcription $\hat{y}$ via conditional generation:

$$\hat{y} = \arg\max_y P(y \mid \mathcal{H}^{\text{asr}}; \theta_{\text{LLM}}). \quad (1)$$

This approach has proven effective in clean acoustic conditions; however, its performance is fundamentally capped by the quality of the initial ASR hypotheses. When the source audio is severely corrupted by noise such as negative signal-to-noise (SNR) level, the resulting hypotheses are too erroneous to provide useful signal for correction, creating a performance bottleneck.

In contrast, visual information such as lip movements offers a complementary modality that is invariant to acoustic noise. Visual modality has been shown to be particularly useful in disambiguating homophones or recovering missing segments in noisy environments (Kim et al., 2022, 2024b). Motivated by this, we propose to extend GER beyond a single-stream hypothesis by incorporating both audio and visual modalities in a unified framework.

3.2 Oracle Error Analysis

To ascertain the potential benefits of incorporating a second modality, we conduct an oracle er-

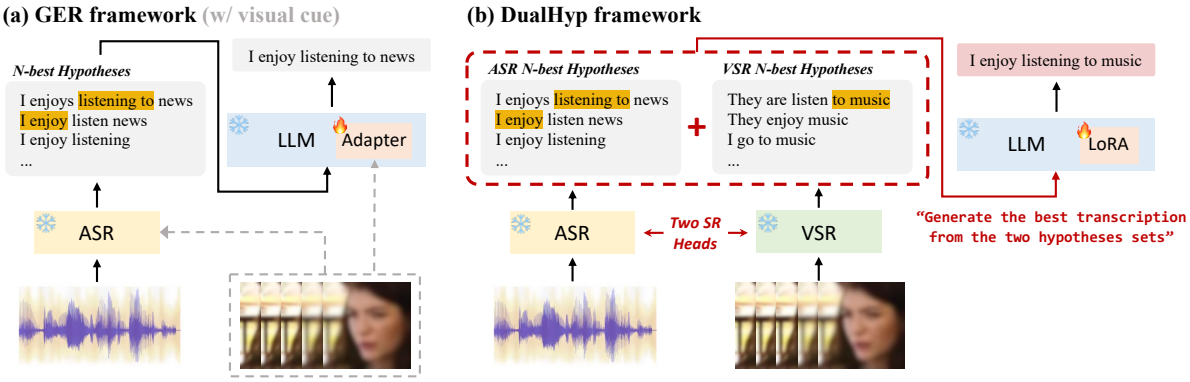

Figure 1: (a) Conventional GER frameworks use a single set of ASR hypotheses and (optionally) injects visual features via an adapter or a multimodal encoder. (b) Our DualHyp framework maintains modality separation, using both ASR and VSR heads to generate two distinct sets of textual hypotheses. The LLM performs compositional reasoning on dual hypotheses in the language space to produce a more robust and accurate transcription.

ror analysis of speech recognition systems, in addition to standard 1-best word error rate (WER). This oracle analysis establishes theoretical lower bounds of ASR and VSR systems in two manners (Chen et al., 2023a): *N-best oracle* (o_{nb}), which selects the single best hypothesis from an N-best list, and *compositional oracle* (o_{cp}), which constructs an optimal transcript by combining correct words from all N-best hypotheses. Table 1 summarizes 1-best WERs of three speech recognition heads: Whisper-large-v3 (Radford et al., 2023) for audio-only, BRAVEn-large (Haliassos et al., 2024) for visual-only, and Auto-AVSR (Ma et al., 2023) for audio-visual. Whisper attains 25.8% WER, while Auto-AVSR is slightly stronger (24.9%), and BRAVEn is markedly weaker (39.7%), confirming that VSR alone lags in overall accuracy.

The oracle WER results reveal the limitation of single-stream systems and the compelling potential of dual-stream approaches (A + AV or A + V). While strong individual models like Whisper and Auto-AVSR perform o_{cp} WERs of 13.7% and 13.6%, respectively, combining hypotheses from independent audio and visual heads drastically reduces potential errors: **Whisper (ASR) + BRAVEn (VSR)** plummets to 4.5%. This gap indicates that audio and video can provide distinct evidence with highly complementary information. Consequently, an ideal GER model that can compose across ASR and VSR hypotheses could significantly reduce errors relative to single-stream systems.

3.3 DualHyp: Dual-Stream Hypotheses

Existing GER approaches for AVSR either inject visual data into LLMs via adapters (Ghosh et al., 2024) or rely on multimodal encoders that perform early fusion of the modalities (Liu et al., 2025a). Both strategies have notable drawbacks; feature adaptation is insufficient for transferring rich visual cues, whereas early fusion is susceptible to cross-modal interference or modality bias.

SR Head	Input	1-best	o_{nb}	o_{cp}
Whisper-large-v3	A	25.8	16.7	13.7
BRAVEn-large	V	39.7	27.8	24.6
Auto-AVSR	AV	24.9	16.1	13.6
Whisper + Auto-AVSR	A + AV	–	7.0	4.9
Whisper + BRAVEn	A + V	–	**6.4**	**4.5**

Table 1: WER (%) analysis with different speech recognition heads, evaluated on noise-augmented LRS2. The audio stream is augmented with four types of noise, and the video stream is corrupted with four visual degradation types. For details, refer to Appendix B.2. o_{nb}: *N-best oracle*, o_{cp}: *compositional oracle*.

Our approach is guided by a different principle, underscored by perceptual phenomena that the premature fusion of conflicting audio-visual signals can distort recognition outcomes (McGurk and MacDonald, 1976). Inspired by prior work that embeds audio noise into the *language space* (Hu et al., 2024b), we suggest that modality-specific information should be explicitly represented in the language space. This allows the LLM to resolve inconsistencies and compose information from both streams without entangling the signals during the upstream feature processing.

Thus, based on our analysis in Section 3.2, we propose **DualHyp**, a novel GER framework that explicitly leverages separate hypotheses streams from both audio and video modalities. Instead of relying on a single recognizer, we utilize indepen-

Type 1: Multimodal Fragment Composition		*Type 2: Dominant Modality Refinement*	
Utterance (ASR + VSR → DualHyp)	**WER**	**Utterance (ASR + VSR → DualHyp)**	**WER**
ASR 5-best ($\mathcal{H}^{\text{asr}}$)**:**		**ASR 5-best** ($\mathcal{H}^{\text{asr}}$)**:**	
everyone going into the den has a fresh chance to talk it around	35.7	<unk>	100.0
everyone going into the den is given a fresh chance to talk it around	42.9	thank you	100.0
everyone going into the den gives you a fresh chance to talk it around	42.9	all right	100.0
and everyone going into the den has a fresh chance to talk around	35.7	the president	100.0
everyone going into the den has a fresh chance to talk to the ground	42.9	god bless you	100.0
VSR 5-best ($\mathcal{H}^{\text{vsr}}$)**:**		**VSR 5-best** ($\mathcal{H}^{\text{vsr}}$)**:**	
but everyone in today gets a fresh chance to turn things around	35.7	project management is really my special considering	14.3
but everyone as i say gets a fresh chance to turn things around	35.7	project management is really by special considering	28.6
but everyone on its day gets a fresh chance to turn things around	35.7	project management is really my specialist theory	14.3
but everyone it is a saying gets a fresh chance to turn things around	35.7	project management and really my special considering	28.6
but everyone it is the saying gets a fresh chance to turn things around	28.6	project management is really my special discovery	28.6
DualHyp output ($\hat{y}$)**:**		**DualHyp output** ($\hat{y}$)**:**	
everyone going into the den gets a fresh chance to turn things around	**14.3**	**project management is really my specialist area**	**0.0**
Ground-truth:		**Ground-truth:**	
but everyone going into the den gets a fresh chance to turn things round	–	project management is really my specialist area	–

Table 2: Examples of successful correction via DualHyp framework. The upper hypothesis within each 5-best list has a higher log-likelihood score. The colored highlights trace the origin of word fragments in the final DualHyp output, showing how those are sourced from ASR, VSR, or both, with a word being newly generated by the LLM's internal knowledge. ***Type 1*** demonstrates the model combining complementary pieces from both modalities, and ***Type 2*** presents the model identifying and correcting the hypothesis from a more reliable modality.

dent, pretrained ASR and VSR models to process an audio-visual pair. Each recognizer head generates a distinct N-best list:

$$\mathcal{H}^{\text{asr}} = \{(h_i^{\text{a}}, s_i^{\text{a}})\}_{i=1}^{N}, \quad \mathcal{H}^{\text{vsr}} = \{(h_j^{\text{v}}, s_j^{\text{v}})\}_{j=1}^{N}.$$

We then form a combined *dual* hypotheses set, $\mathcal{H}^{\text{dual}} = \mathcal{H}^{\text{asr}} \cup \mathcal{H}^{\text{vsr}}$, which preserves the modality-specific information in each hypothesis set. The LLM is conditioned on this enriched set to generate the DualHyp output:

$$\hat{y} = \arg\max_{y} P(y \mid \mathcal{H}^{\text{dual}}; \theta_{\text{LLM}}). \qquad (2)$$

By maintaining separate modality pathways into the language space, this approach avoids the cross-modal contamination issues seen in early-fusion models. It instead enables the LLM to act as an in-context compositional reasoner (An et al., 2023; Qiu et al., 2022), cross-referencing the audio and visual evidence to resolve ambiguities and reconstruct the intended utterance. Figure 1 illustrates the overview of our DualHyp framework, compared to existing GER approaches.

Analysis. Table 2 provides qualitative analysis of DualHyp to show its effectiveness. We highlight two primary correction mechanisms that explicitly exploit the LLM's capacity for semantic and linguistic reasoning within the language space. *(Type 1) Multimodal Fragment Composition:* The model constructs the output by conducting a word-level semantic alignment across the dual hypotheses. It actively recovers the correct transcription by extracting complementary linguistic fragments, such as preserving the acoustically clear `den` from the ASR stream while adopting the visually distinct phrase `turn things around` from the VSR stream. *(Type 2) Dominant Modality Refinement:* When one modality provides highly degraded information, the model identifies the more coherent stream and exclusively grounds its refinement process on the dominant modality. Furthermore, this refinement process retains the LLM's prior knowledge, as it generates a word not present in any source hypothesis, *i.e.*, `area`. We provide more examples, including failure cases, in Appendix C.

4 Noise-Aware Guidance of DualHyp

The DualHyp framework enables an LLM to compose information from separate ASR and VSR hypotheses. However, since LLM operates purely on these text inputs, the model lacks explicit information about the source signal quality, creating a risk of leveraging unreliable, inaccurate hypotheses (Hong et al., 2023). To bridge this gap from an LLM perspective, we introduce **RelPrompt**, a noise-aware guidance mechanism that explicitly informs the LLM about the temporal reliability of each stream. RelPrompt is achieved by (1) predicting reliability tokens for each modality using external predictors, which are then (2) provided to the LLM's prompt to serve as temporal guidance.

4.1 Reliability Mask Prediction

To generate a compact, time-aligned reliability signal, we segment both the audio and video streams to approximate the duration of a single spoken word. Grounded in the average native English speaking rate (Becker et al., 2022; Yuan et al., 2006), we set the chunk size to 0.4 seconds, *i.e.*, 150 wpm. We process each modality as follows:

- **Audio stream:** The input audio, sampled at 16kHz, is grouped into segments of 6,400 samples (16,000 samples/sec × 0.4 sec).
- **Video stream:** The input video, processed at 25Hz, is grouped into segments of 10 frames (25Hz × 0.4 sec).

We then employ two lightweight predictors consisting of 1D convolutional neural networks (CNN) that operate on the intermediate features extracted from the ASR and VSR encoders, thus avoiding additional feature extraction. For each segment, the predictors produce a discrete token $m_i \in \{\texttt{Clean}, \texttt{Noisy}, \texttt{Mixed}\}$, forming a reliability mask that indicates the quality of the source signal to the LLM. The ground-truth reliability is labeled as `Clean` if <10% of its frames are corrupted, `Noisy` if >60% of its frames are corrupted, and `Mixed` otherwise. Each predictor outputs a sequence of these tokens for its respective modality:

$$\mathbf{m}^{\mathrm{a}} = (m_1^{\mathrm{a}}, \ldots, m_K^{\mathrm{a}}), \quad \mathbf{m}^{\mathrm{v}} = (m_1^{\mathrm{v}}, \ldots, m_K^{\mathrm{v}}).$$

4.2 Reliability Guidance

As illustrated in Figure 2, the reliability token sequences, $\mathbf{m}^{\mathrm{a}}$ and $\mathbf{m}^{\mathrm{v}}$ are appended to the dual hypotheses to directly inform the LLM of each modality's temporal reliability. The entire model is then trained end-to-end, conditioned on both the hypotheses and reliability masks to generate the final transcript:

$$\hat{y} = \arg\max_{y} P\big(y \mid \mathcal{H}^{\mathrm{dual}}, \mathbf{m}^{\mathrm{a}}, \mathbf{m}^{\mathrm{v}}; \theta_{\mathrm{LLM}}\big). \quad (3)$$

This format allows the LLM to learn the correlation between the reliability tokens and hypotheses quality. Crucially, this approach avoids the need for explicit word-level alignment, which is infeasible for N-best lists with variable lengths and erroneous words (Gekhman et al., 2022; Qiu et al., 2021). Additionally, the RelPrompt token sequence enhances the interpretability of LLM's reasoning, revealing when the model switches its focus between the ASR and VSR hypotheses.

```
Below are the best-hypothesis transcribed
from ASR and VSR. Revise it using the
words which are only included into
other-hypotheses, and write the response
for the true transcription.  Refer to the
audio and video masks for reliability.

### ASR Best-hypothesis: {h_1^a}
### ASR Other-hypotheses: {h_2^a || ··· || h_N^a}
### Audio Mask: [C] [N] [N] [M] [C] ···

### VSR Best-hypothesis: {h_1^v}
### VSR Other-hypotheses: {h_2^v || ··· || h_N^v}
### Video Mask: [C] [C] [C] [N] [N] ···

### Response:
```

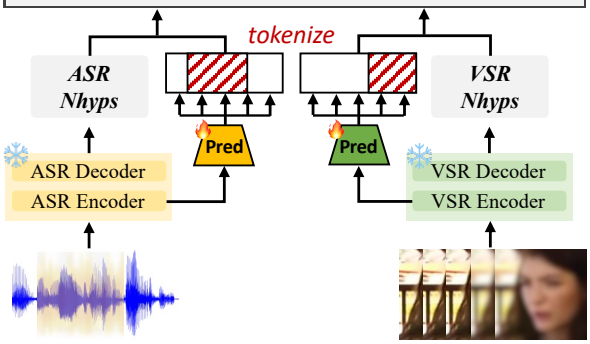

Figure 2: An overview of our DualHyp with RelPrompt. Each predictor uses ASR/VSR encoder features to generate a noise-aware token sequence. These masks accurately guide the LLM to dynamically switch the model's focus between the ASR and VSR hypotheses.

5 Experiments

5.1 Experimental Setup

We conduct our experiments on the LRS2 AVSR benchmark (Son Chung et al., 2017) with the WER metric. All models are trained and tested under diverse, synthetically corrupted audio-visual conditions, following the protocol of CAV2vec (Kim et al., 2025a). Unless specified otherwise, our DualHyp framework is composed of a Whisper-large-v3 (Radford et al., 2023) ASR head, a BRAVEn-large (Haliassos et al., 2024) VSR head, and a TinyLlama (Zhang et al., 2024a) LLM, which we fine-tune using LoRA (Hu et al., 2022). Full details of implementation, corruption protocol, and baseline methods are provided in Appendix B. We also offer additional results on another AVSR benchmark, LRS3 (Afouras et al., 2018), in Appendix D to support solid performance.

5.2 LRS2 Benchmark Results

Table 3 presents the benchmark results, where we isolate modality-specific robustness by either varying audio noise against fixed visual corruption

Method	Input	Babble (B)	Speech (S)	Music (M)	Natural (N)	Overall (O)
ASR oracle o_{nb}/o_{cp}	A	30.9 / 26.9	19.4 / 14.1	8.0 / 6.6	8.5 / 7.4	16.7 / 13.7
ASR + VSR oracle o_{nb}/o_{cp}	A + V	11.7 / 8.8	6.6 / 4.4	3.5 / 2.2	3.6 / 2.7	6.4 / 4.5
Whisper-large-v3 (Radford et al., 2023)	A	40.0	36.5	12.7	14.2	25.8
BRAVEn-large (Haliassos et al., 2024)	V	-	-	-	-	$39.7_{(+53.9\%)}$
GER (Chen et al., 2023a)	A	$39.3_{(-1.8\%)}$	$34.4_{(-5.8\%)}$	$11.5_{(-9.4\%)}$	$13.2_{(-7.0\%)}$	$24.6_{(-4.7\%)}$
RobustGER (Hu et al., 2024b)	A	$39.3_{(-1.8\%)}$	$33.8_{(-7.4\%)}$	$11.7_{(-7.9\%)}$	$13.1_{(-7.7\%)}$	$24.5_{(-5.0\%)}$
LipGER (Ghosh et al., 2024)	AV	$39.3_{(-1.8\%)}$	$34.2_{(-6.3\%)}$	$12.0_{(-5.5\%)}$	$13.4_{(-5.6\%)}$	$24.7_{(-4.3\%)}$
GER w/ Auto-AVSR†	AV	**18.9**$_{(-52.8\%)}$	$39.0_{(+6.8\%)}$	$17.4_{(+37.0\%)}$	$18.1_{(+27.5\%)}$	$23.3_{(-9.7\%)}$
RelPrompt (ours, w/o DualHyp)	A	$39.7_{(-0.8\%)}$	$33.6_{(-7.9\%)}$	$11.8_{(-7.1\%)}$	$13.1_{(-7.7\%)}$	$24.5_{(-5.0\%)}$
RelPrompt (ours, w/o DualHyp)	V	-	-	-	-	$39.3_{(+52.3\%)}$
DualHyp w/o LLM (ROVER) (Fiscus, 1997)	A + V	$29.4_{(-26.5\%)}$	$30.1_{(-17.5\%)}$	$13.8_{(+8.7\%)}$	$14.7_{(+3.5\%)}$	$22.0_{(-14.7\%)}$
DualHyp (ours)	A + V	$21.6_{(-46.0\%)}$	$17.9_{(-51.0\%)}$	$8.1_{(-36.2\%)}$	$9.3_{(-34.5\%)}$	$14.2_{(-45.0\%)}$
+ RelPrompt (ours)	A + V	$20.4_{(-49.0\%)}$	**16.0**$_{(-56.2\%)}$	**8.0**$_{(-37.0\%)}$	**8.2**$_{(-42.3\%)}$	**13.2**$_{(-48.8\%)}$

(a) Audio: random noise [-10, 10] dB, Video: 50% segment occluded with object

Method	Input	Object	Hands	Pixelate	Blur	Overall
ASR oracle o_{nb}/o_{cp}	A	-	-	-	-	10.9 / 6.6
ASR + VSR oracle o_{nb}/o_{cp}	A + V	4.7 / 2.8	4.5 / 2.6	4.8 / 2.7	4.1 / 2.4	4.5 / 2.6
Whisper-large-v3 (Radford et al., 2023)	A	26.7	26.7	26.7	26.7	26.7
BRAVEn-large (Haliassos et al., 2024)	V	$39.7_{(+48.7\%)}$	$35.1_{(+31.5\%)}$	$39.4_{(+47.6\%)}$	$31.7_{(+18.7\%)}$	$36.5_{(+36.7\%)}$
GER (Chen et al., 2023a)	A	-	-	-	-	$23.9_{(-10.5\%)}$
RobustGER (Hu et al., 2024b)	A	-	-	-	-	$24.9_{(-6.7\%)}$
LipGER (Ghosh et al., 2024)	AV	$24.2_{(-9.4\%)}$	$24.3_{(-9.0\%)}$	$24.3_{(-9.0\%)}$	$24.1_{(-9.7\%)}$	$24.3_{(-9.0\%)}$
GER w/ Auto-AVSR†	AV	$29.5_{(+10.5\%)}$	$26.6_{(-0.4\%)}$	$29.1_{(+9.0\%)}$	$23.5_{(-12.0\%)}$	$27.2_{(+1.9\%)}$
DualHyp w/o LLM (ROVER) (Fiscus, 1997)	A + V	$22.9_{(-14.2\%)}$	$21.9_{(-18.0\%)}$	$23.6_{(-11.6\%)}$	$20.7_{(-22.5\%)}$	$22.3_{(-16.5\%)}$
DualHyp (ours)	A + V	$12.0_{(-55.1\%)}$	$11.8_{(-55.7\%)}$	$12.7_{(-52.6\%)}$	$11.1_{(-58.4\%)}$	$11.9_{(-55.4\%)}$
+ RelPrompt (ours)	A + V	**11.9**$_{(-55.4\%)}$	**11.0**$_{(-58.8\%)}$	**11.9**$_{(-55.4\%)}$	**10.2**$_{(-61.8\%)}$	**11.3**$_{(-57.7\%)}$

(b) Audio: speech noise 0 dB, Video: random segment corrupted

Table 3: WER% (↓) results on the LRS2 test set under joint audio-visual corruption. (a) Performance across varying audio noise types, with a fixed visual corruption (50% segment occluded by an object). (b) Performance across varying visual corruption types, with a fixed audio corruption (0 dB speech noise). We also show the relative WER reduction in parentheses compared to the Whisper-large-v3 ASR baseline. All the ASR and VSR heads are Whisper-large-v3 and BRAVEn-large, respectively. †: We implement a GER model using hypotheses generated from an early-fusion approach, Auto-AVSR (Ma et al., 2023), which has been trained on LRS2 with babble noise.

(Table 3a) or varying visual corruption against fixed audio noise (Table 3b). Our proposed **DualHyp + RelPrompt** achieves the lowest **overall WER of 13.2%** under audio variability and **11.3%** under visual variability, representing a relative improvement of 48.8% and 57.7% compared to the ASR baseline, Whisper-large-v3. This confirms our core hypothesis that LLMs can perform robust compositional reasoning when provided with separate ASR and VSR hypotheses.

In contrast, all baseline methods show clear limitations. ASR-only models like GER (Chen et al., 2023a) or RobustGER (Hu et al., 2024b) are fundamentally capped by the input audio quality and struggle under low SNRs (also refer to §6.4). Audio-visual approach like LipGER (Ghosh et al., 2024) fails to improve over the standard GER framework, which shows that injecting video via additional adapter is insufficient for LLM to fully exploit the visual modality while harming its stability due to cross-modal gap (Gao et al., 2023; Li et al., 2023; Zhang et al., 2024c). Similarly, GER w/ Auto-AVSR (Ma et al., 2023) exhibits strong but narrow performance, excelling only on the babble noise that the model's AVSR head is specifically trained on, failing to generalize to other conditions (see §6.3 for further analysis). Furthermore, while DualHyp without LLM reasoning (ROVER, Fiscus (1997)) outperforms feature-level early fusion through text-level late fusion, it remains inferior to our full framework. This indicates that simple voting mechanisms are insufficient to capture the semantic complementarity across multimodal hypotheses, underscoring the necessity of explicit LLM reasoning.

The success of our DualHyp with RelPrompt approach stems from two aspects: (1) a text-level late fusion strategy and (2) the ability to dynamically leverage the more reliable modality. Our late fusion provides the LLM with rich, modality-specific evidence in a unified text format that is readily processed by the LLM. The isolation of modalities also

Method	Input	A^cV^c	A^cV^n	A^nV^c
Whisper-large-v3	A	3.8	3.8	25.8
BRAVEn-large	V	26.9	36.5	26.9
GER	A	2.6	2.6	24.6
DualHyp	A + V	**1.9**	2.1	11.5
+ RelPrompt	A + V	**1.9**	**2.0**	**9.9**

Table 4: Performance under different modality conditions on LRS2, with clean audio or video (X^c) and noisy audio or video (X^n), $X \in \{A, C\}$.

ensures that corruption in one stream does not contaminate the other. Then, RelPrompt dynamically leverages the more reliable stream, utilizing visual hypotheses when audio quality is low and falling back on audio hypotheses when the visual stream is degraded. An ablation study in Table 3a confirms this, showing that RelPrompt yields only marginal gains in single-modality inference and intinsically relies on the dual-stream framework to facilitate cross-modal arbitration. Notably, the overall success of this dual-stream approach is achieved even though our VSR model is substantially weaker than ASR, suggesting that our framework's potential is scalable as more powerful VSR models emerge.

Clean audio or video inputs. Even in the clean audio settings (Table 4), our DualHyp methods achieve the lowest WER, showing they effectively capitalize on the high-quality audio stream. In the noisy-audio/clean-video setting, while GER is severely hampered by corrupted audio (24.6%), RelPrompt leverages clean visual hypotheses to dramatically improve to 9.9%. The gap between DualHyp (11.5%) and its reliability-guided version demonstrates that dynamically detecting clean signal (in this case video) and giving the LLM explicit hints about which to trust is effective.

5.3 Larger LLMs

We investigate the impact of LLM scale by evaluating our methods with three different models: TinyLlama (Zhang et al., 2024a), Phi-2 (Javaheripi et al., 2023), and Llama-3.2-3B (Meta AI, 2024). The results in Table 5 show that the benefits of a larger LLM are most pronounced within our proposed framework. For the GER baseline, scaling the LLM yields only marginal gains, indicating that the performance is limited by the quality of the single-stream input hypotheses. In contrast, our models benefit more significantly from a larger LLM's capacity. The effect is greatest for DualHyp + RelPrompt, which achieves the best overall WER

Method	LLM (Params.)	B	S	M	N	O
GER	TinyLlama (1.1B)	39.3	34.4	11.5	13.2	24.6
	Phi-2 (2.7B)	39.0	33.7	11.9	13.0	24.4
	Llama-3.2 (3.2B)	38.9	34.1	11.6	12.9	24.4
DualHyp	TinyLlama (1.1B)	21.6	17.9	8.1	9.3	14.2
	Phi-2 (2.7B)	21.6	19.0	7.8	8.7	14.3
	Llama-3.2 (3.2B)	20.4	16.0	**7.2**	**8.1**	12.9
DualHyp + RelPrompt	TinyLlama (1.1B)	20.4	16.0	8.0	8.2	13.2
	Phi-2 (2.7B)	21.1	18.2	8.0	8.5	14.0
	Llama-3.2 (3.2B)	**19.6**	**14.1**	7.4	8.2	**12.3**

Table 5: WER (%) comparison using different LLMs on the LRS2 benchmark. The corruption strategy follows Table 3a, where **B**, **S**, **M**, and **N** represent each noise type with the overall result (**O**).

Method	Es	Fr	It	Pt	Avg
Whisper-large-v3	49.6	**46.8**	52.3	52.7	50.4
mAV-HuBERT	70.5	81.7	73.7	74.1	75.0
GER	50.6	47.8	58.5	52.3	52.3
DualHyp	**47.3**	47.9	**47.2**	**49.0**	**47.9**

Table 6: WER (%) comparison with multilingual babble noise (SNR = 0 dB) on the MuAViC dataset.

of 12.3% with Llama-3.2. This suggests that by providing a richer and more comprehensive input, our framework creates a more sophisticated reasoning task that can effectively leverage the capabilities of LLMs.

5.4 Multilingual AVSR

To evaluate our framework in a multilingual context, we conduct experiments on the MuAViC dataset (Anwar et al., 2023) with adding multilingual babble noise at SNR 0 dB (Kim et al., 2025b). While the Whisper ASR head remains the same as in prior experiments, a VSR head is fine-tuned from mAV-HuBERT (Kim et al., 2024a) for each language, due to the absence of strong multilingual VSR system. Llama-3.2-3B is employed for the multilingual reasoning. In Table 6, our framework outperforms both Whisper and GER in three of the four languages. However, this performance gain can be limited when VSR performance is severely degraded, as observed in the French case. We thus anticipate that the performance gains of our methodology will become even more significant as more powerful multilingual VSR models emerge.

6 Analysis

6.1 Qualitative Analysis

Our qualitative analysis in Figure 3 illustrates how RelPrompt corrects failures of the baseline DualHyp framework by providing explicit reliability signals. (*Left*): RelPrompt uses clean video tokens

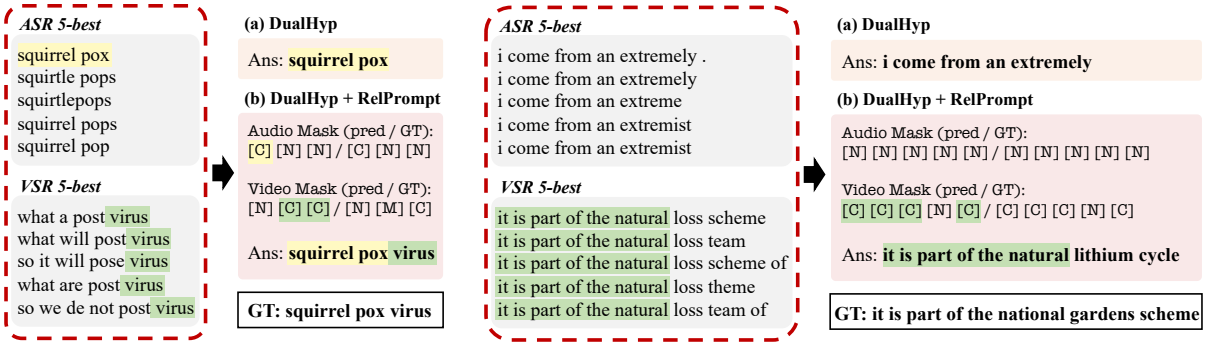

Figure 3: Qualitative analysis comparing RelPrompt to the DualHyp baseline. RelPrompt uses reliability tokens (*i.e.*, masks) to explicitly inform the input signal quality, correctly guiding the use of ASR and VSR hypotheses.

SNR	Acc.	Precision	Recall	F1	WER
−10 dB	84.7	95.3	87.8	91.4	25.8
−5 dB	83.9	95.0	87.1	90.9	17.8
0 dB	82.2	94.4	85.5	89.7	7.2
5 dB	79.6	93.0	82.8	87.6	3.4
10 dB	76.2	90.9	78.2	84.1	2.5

Table 7: Performance (%) of the reliability mask predictors with randomly corrupted audio and video segments. The metrics evaluate the classification of segments as `noisy`, which includes the `mixed` category.

Method	Input	# hyps	B	S	M	N	O
GER	A	5	39.3	34.4	11.5	13.2	24.6
	AV	5	18.9	39.0	17.4	18.1	23.3
	AV	10	18.2	38.1	16.7	17.6	22.6
DualHyp	A + AV	10	17.1	26.7	**7.2**	8.4	14.8
	A + V	10	21.6	17.9	8.1	9.3	14.2
DualHyp + RelPrompt	A + AV	10	**15.4**	25.9	7.3	8.8	14.3
	A + V	10	20.4	**16.0**	8.0	**8.2**	**13.2**

Table 8: WER (%) comparison of different hypotheses from single-stream (GER) and dual-stream (DualHyp) generation heads. Note that the AVSR head is trained on LRS2 with babble noise (Ma et al., 2023), unlike the ASR and VSR heads.

`[C]` as a cue to trust the last part of the VSR hypotheses, allowing it to recover the word (`virus`) which the baseline has missed. (*Right*): The ASR system is presented with fluent but entirely incorrect hypotheses. By referencing the consistently noisy audio tokens `[N]`, the LLM correctly identifies the ASR stream as unreliable and pivots to the more accurate VSR candidates. In contrast, without the RelPrompt mechanism, the model lacks any modality-level grounding and produces a completely incorrect output. These cases demonstrate that by providing explicit reliability tokens, RelPrompt empowers the LLM to act as an intelligent controller, grounding its compositional reasoning in the predicted quality of the source signals.

6.2 Reliability Mask Prediction

In Table 7, our evaluation of the reliability predictors reveals two key strengths. First, the predictor shows consistently high precision (>90%), which ensures that its `noisy` flags are highly trustworthy and prevents the main model from incorrectly discarding clean data. Second, the recall naturally decreases as the SNR increases. This is a desirable behavior, as the predictor conservatively labels mildly corrupted audio segments as `clean`, allowing the model to continue exploiting the useful signal.

6.3 Comparison with an AVSR Head

Our analysis in Table 8 highlights two key findings regarding hypothesis generation. First, modality diversity of hypotheses is more crucial than sheer quantity. Simply increasing the number of hypotheses for the single-stream GER (5 → 10 AV hypotheses) yields only a marginal gain for overall performance (23.3% → 22.6%), compared to DualHyp using 5-best hypotheses from each distinct modality (23.3% → 14.2%).

Second, while AVSR hypotheses might seem viable alternatives to VSR, they remain overly dependent on the audio modality. This is particularly evident under the speech noise condition, where the visual stream is crucial for disambiguating target utterance from interfering speech. In this scenario, DualHyp (A + AV) struggles (26.7% WER), as the early fusion of AVSR embeddings makes visual information rely on the corrupted audio. Instead, DualHyp (A + V) leverages the audio-independent VSR stream to achieve 17.9%, demonstrating the superiority of using disentangled hypotheses. These findings are further supported by our LRS3 experiments (Table 15).

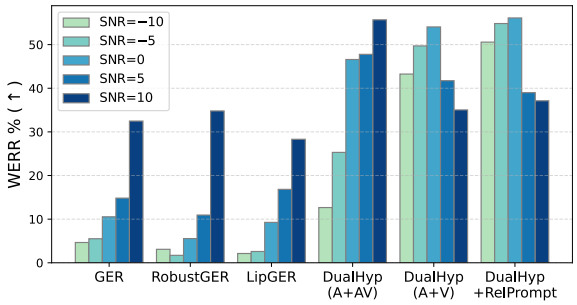

Figure 4: WERR at different audio SNRs, under speech noise. Higher WERR indicates greater improvement over the Whisper ASR baseline.

6.4 SNR-wise WER Improvement

Figure 4 reveals opposing trends in WER reduction (WERR, Liu et al. (2025a)) between single-stream and dual-stream methods. For single-stream methods, WERR increases with better audio quality, as their effectiveness is limited to refining an already decent ASR output. In contrast, our dual-stream framework maintains a high WERR even at very low SNRs by leveraging VSR hypotheses. Furthermore, the addition of RelPrompt consistently boosts performance, with the most significant gains observed in low-SNR scenarios. This confirms that by effectively utilizing the reliability information about corruption provided by RelPrompt, our framework can substantially reduce errors precisely when the audio is most challenging.

7 Conclusion

In this study, we introduced DualHyp, a novel GER framework for AVSR that deliberately delays modality fusion to the language space, where an LLM performs compositional reasoning on independent hypotheses from ASR and VSR models. We further enhanced this with RelPrompt, a noise-aware guidance mechanism that guides the LLM with explicit, time-aligned reliability signals for each modality. The experiments showed that our new framework significantly outperforms single-stream GER approaches, highlighting a flexible paradigm that leverages modular integration.

Limitations

While our framework demonstrates significant robustness and scalability in AVSR, it still holds two primary limitations that are common to most GER systems.

First, the performance of our framework is fundamentally dependent on the quality of its consisting components, especially the upstream SR heads. If the initial hypotheses from the SR head are of poor quality, as seen in our results of the MuAViC French case, the LLM's ability to perform corrections is limited. This dependency currently restricts the framework's applicability beyond English, because there is no publicly available, high-quality multilingual VSR model, making adaptation to the low-resource speech recognition and translation challenging.

Second, multiple modules in our structure introduces computational latency, posing a challenge for real-time applications. Although the ASR and VSR streams can be processed in parallel, the final LLM correction step is sequential, creating an unavoidable bottleneck. While modern efficiency techniques like flash attention can mitigate this to an extent, the approach remains inherently slower than a single end-to-end model, making deployment on resource-constrained edge devices a significant hurdle.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No.2022-0-00641, XVoice: Multi-Modal Voice Meta Learning].

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052.

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. In *Proc. Interspeech 2023*, pages 4064–4068.

Brett A Becker, Daniel Gallagher, Paul Denny, James Prather, Colleen Gostomski, Kelli Norris, and Garrett Powell. 2022. From the horse's mouth: The words we use to teach diverse student groups across three

continents. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*, pages 71–77.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2025a. Large language models are strong audio-visual speech recognition learners. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Umberto Cappellazzo, Minsu Kim, and Stavros Petridis. 2025b. Adaptive audio-visual speech recognition via matryoshka-based multimodal llms. *arXiv preprint arXiv:2503.06362*.

Umberto Cappellazzo, Minsu Kim, Stavros Petridis, Daniele Falavigna, and Alessio Brutti. 2025c. Scaling and enhancing llm-based avsr: A sparse mixture of projectors approach. *arXiv preprint arXiv:2505.14336*.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. 2023a. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688.

Chen Chen, Yuchen Hu, Qiang Zhang, Heqing Zou, Beier Zhu, and Eng Siong Chng. 2023b. Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12607–12615.

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, EngSiong Chng, and Chao-Han Huck Yang. 2024. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. In *International Conference on Learning Representations*.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, and 1 others. 2024. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proc. ASRU(Automatic Speech Recognition and Understanding Proceedings)*, pages 347–354.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, and 1 others. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Zorik Gekhman, Dina Zverinski, Jonathan Mallinson, and Genady Beryozkin. 2022. Red-ace: Robust error detection for asr using confidence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2800–2808.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Purva Chiniya, Utkarsh Tyagi, Ramani Duraiswami, and Dinesh Manocha. 2024. Lipger: Visually-conditioned generative error correction for robust automatic speech recognition. In *Proc. Interspeech 2024*, pages 1920–1924.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alexandros Haliassos, Andreas Zinonos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2024. Braven: Improving self-supervised pre-training for visual and auditory speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11431–11435. IEEE.

HyoJung Han, Mohamed Anwar, Juan Pino, Wei-Ning Hsu, Marine Carpuat, Bowen Shi, and Changhan Wang. 2024. Xlavs-r: Cross-lingual audio-visual speech representation learning for noise-robust speech perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12896–12911.

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

Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. 2022. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. In *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*, pages 2838–2842. International Speech Communication Association.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large

language models. In *International Conference on Learning Representations*.

Yuchen Hu, Chen Chen, Chengwei Qin, Qiushi Zhu, EngSiong Chng, and Ruizhe Li. 2024a. Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 666–679.

Yuchen Hu, Chen Chen, Chao-han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Ensiong Chng. 2024b. Large language models are efficient learners of noise-robust speech recognition. In *International Conference on Learning Representations*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, and 1 others. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3.

Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. 2022. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1174–1182.

Minsu Kim, Jeonghun Yeo, Se Jin Park, Hyeongseop Rha, and Yong Man Ro. 2024a. Efficient training for multilingual visual speech recognition: Pre-training with discretized visual speech representation. In *ACM Multimedia 2024*.

Sungnyun Kim, Sungwoo Cho, Sangmin Bae, Kangwook Jang, and Se-Young Yun. 2025a. Multi-task corrupted prediction for learning robust audio-visual speech representation. In *The Thirteenth International Conference on Learning Representations*.

Sungnyun Kim, Kangwook Jang, Sangmin Bae, Sungwoo Cho, and Se-Young Yun. 2025b. Mohave: Mixture of hierarchical audio-visual experts for robust speech recognition. In *Forty-second International Conference on Machine Learning*.

Sungnyun Kim, Kangwook Jang, Sangmin Bae, Hoirin Kim, and Se-Young Yun. 2024b. Learning video temporal dynamics with cross-modal attention for robust audio-visual speech recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 447–454. IEEE.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023. Prompting large language models for zero-shot domain adaptation in speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Hong Liu, Wanlu Xu, and Bing Yang. 2021. Audio-visual speech recognition using a two-step feature fusion strategy. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1896–1903. IEEE.

Rui Liu, Hongyu Yuan, Guanglai Gao, and Haizhou Li. 2025a. Listening and seeing again: Generative error correction for audio-visual speech recognition. *Information Fusion*, 120:103077.

Yanyan Liu, Minqiang Xu, Yihao Chen, Liang He, Lei Fang, Sian Fang, and Lin Liu. 2025b. Denoising ger: A noise-robust generative error correction with llm for speech recognition. *arXiv preprint arXiv:2509.04392*.

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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and 1 others. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Bingshen Mu, Xucheng Wan, Naijun Zheng, Huan Zhou, and Lei Xie. 2024. Mmger: Multi-modal and multi-granularity generative error correction with llm for joint accent and speech recognition. *IEEE Signal Processing Letters*.

Bingshen Mu, Kun Wei, Pengcheng Guo, and Lei Xie. 2025. Mixture of lora experts with multi-modal and multi-granularity llm generative error correction for accented speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, and 1 others. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. In *Proc. Interspeech 2024*, pages 352–356.

David Qiu, Qiujia Li, Yanzhang He, Yu Zhang, Bo Li, Liangliang Cao, Rohit Prabhavalkar, Deepti Bhatia, Wei Li, Ke Hu, and 1 others. 2021. Learning word-level confidence for subword end-to-end asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6393–6397. IEEE.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022. Evaluating the impact of model scale for compositional generalization in semantic parsing. In *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pages 9157–9179.

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*, pages 28492–28518. PMLR.

Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér. 2023. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022a. Learning audio-visual speech representation by masked multimodal cluster prediction. *International Conference on Learning Representations*.

Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022b. Robust self-supervised audio-visual speech recognition. In *Proc. Interspeech 2022*, pages 2118–2122.

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

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456.

Kenny TR Voo, Liming Jiang, and Chen Change Loy. 2022. Delving into high-quality synthetic face occlusion segmentation datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4711–4720.

He Wang, Pengcheng Guo, Pan Zhou, and Lei Xie. 2024. Mlca-avsr: Multi-layer cross attention fusion based audio-visual speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8150–8154. IEEE.

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. 2025. Mms-llama: Efficient llm-based audio-visual speech recognition with minimal multimodal speech tokens. *arXiv preprint arXiv:2503.11315*.

Jeonghun Yeo, Seunghee Han, Minsu Kim, and Yong Man Ro. 2024. Where visual speech meets language: Vsp-llm framework for efficient and context-aware visual speech processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11391–11406.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12637–12641. IEEE.

Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an integrated understanding of speaking rate in conversation. In *Interspeech*. Pittsburgh, PA.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024a. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. IEEE.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024b. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024c. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Appendix

A Release of DualHyp Dataset

To facilitate future research within our DualHyp framework, we publicly release the hypotheses dataset[1]. The primary motivation of this dataset construction is to decouple the computationally expensive hypothesis generation step from the LLM fine-tuning process. By providing pre-generated ASR and VSR hypotheses, this dataset will allow researchers to focus directly on developing novel language-space fusion and correction strategies, significantly lowering the barrier to entry.

Our DualHyp hypotheses are mainly created from LRS2 (Son Chung et al., 2017) and LRS3 (Afouras et al., 2018) datasets. LRS2 is a benchmark of British English speech from BBC that covers diverse speakers and topics, while LRS3 consists of spoken utterances from TED and TEDx recordings. For LRS2, our hypotheses dataset covers the standard splits (45,830 training, 1,082 validation, and 1,243 test utterances). For high-resource training (dealt in Appendix D.1), we use additional 95,642 utterances, and for the LRS3 experiments (dealt in Appendix D.4), we use 30,775 training utterances.

For both the ASR head[2] and VSR head[3], we generate hypotheses using a beam size of 50. We select the 5-best unique hypotheses from each stream. If fewer than five unique hypotheses are generated, we randomly sample from the existing ones to reach the size 5. This results in a total of 10 hypotheses (5 from ASR, 5 from VSR) that are fed into the LLM for error correction.

Each entry also includes the ground-truth transcription and metadata detailing the specific audio or visual corruption applied (see Appendix B.2). Because different corruption types result in different output hypotheses, we separately save the dataset for each corruption condition. For training, a complete dataset is formed by merging these individual sets and randomly sampling hypotheses.

B Experimental Details

B.1 Implementation of DualHyp

We fine-tune the LLM using LoRA (Hu et al., 2022) with a rank of $r = 16$. The number of trainable parameters is 4.5M for TinyLlama[4], 23.6M for Phi-2[5], and 24.3M for Llama-3.2[6]. For TinyLlama, we apply LoRA to the attention layers (key, value, query, and projection) only. For the larger Phi-2 and Llama-3.2 models, we apply LoRA to both the attention module and the feed-forward network (FFN) layers to ensure better convergence. All models are trained for 5 epochs with a batch size of 32 and a learning rate of 1e-4.

For the implementation of RelPrompt, our reliability predictors are designed lightweight, with only 1.1M parameters each. The architecture consists of two 1D-convolutional layers followed by average pooling and a final linear classifier to match the segment size. For training these predictors, we create ground-truth labels for each 0.4-second segment based on its constituent frames: a segment is labeled `[C]` (Clean) if less than 10% of its frames are corrupted, `[N]` (Noisy) if more than 60% of its frames are corrupted, and `[M]` (Mixed) otherwise.

The full DualHyp + RelPrompt model is trained for 8 hours on a single NVIDIA A6000 GPU, using a learning rate of 2e-4 for the main LLM (with LoRA) and 1e-4 for the reliability predictors. For all data pre-processing and evaluation, we use the publicly available packages following the LipGER codebase[7].

B.2 Corruption Protocol

In our study, all models are trained and evaluated under challenging noisy conditions to assess their robustness in real-world scenarios. To ensure a robust evaluation, we introduce a diverse set of synthetic corruptions into the LRS2 dataset, following the protocol established by Kim et al. (2025a)[8].

- Audio corruptions: We augment the audio streams with four types of noise, similar to Shi et al. (2022b). We use speech noise from the LRS3 dataset (Afouras et al., 2018) and babble, music, and natural sounds from the MUSAN corpus (Snyder et al., 2015).
- Visual corruptions: We apply four common visual degradation types: object occlusion (Voo et al., 2022), hands occlusion, pixelation, and blur (Kim et al., 2025a).

[1] https://github.com/sungnyun/dualhyp
[2] https://huggingface.co/openai/whisper-large-v3
[3] https://github.com/ahaliassos/raven
[4] https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0
[5] https://huggingface.co/microsoft/phi-2
[6] https://huggingface.co/meta-llama/Llama-3.2-3B
[7] https://github.com/Sreyan88/LipGER
[8] https://github.com/sungnyun/cav2vec

During training, we randomly apply one of these corruption types to each sample. The duration of the applied corruption is also randomized, with its portion sampled from a Beta distribution ($\alpha, \beta = 2.0$) to simulate varying levels of interference.

For evaluation, we apply background noise to the entire audio sample and corrupt partial video segments to better reflect real-world scenarios. For Table 3a, audio noise is applied to the entire time duration with SNR randomly sampled from [-10, 10] dB, while half of the video segments is occluded with object. For Table 3b, 0 dB SNR of speech noise is augmented to the whole audio, while video corruption length is sampled from Beta distribution.

To assess overall performance, we report the average WER from a single comprehensive evaluation run. This run covers all test samples and incorporates a diverse range of noisy conditions to ensure the statistical credibility of our methods.

B.3 Baseline Methods

For a fair comparison, we train all baseline methods from scratch on the same set of corrupted audio-visual data. We have found that this diverse noise training significantly boosts the performance of all GER-based methods, which establishes a strong set of baselines for our evaluation. Our primary baselines are:

- GER (Chen et al., 2023a): The foundational LLM-based error correction framework that operates on N-best hypotheses from an ASR model.
- RobustGER (Hu et al., 2024b): An extension of GER designed to improve robustness against noisy audio conditions.
- LipGER (Ghosh et al., 2024): An audio-visual GER method that incorporates visual features via an adapter but still relies on a single stream of ASR hypotheses for correction.
- GER w/ Auto-AVSR (Ma et al., 2023): A strong baseline we implement by feeding the N-best hypotheses from the early-fusion Auto-AVSR model into a standard GER framework.
- ROVER (Fiscus, 1997): A post-processing ensemble framework that constructs a word transition network through the iterative alignment of multiple hypotheses.

We note that training these single-stream GER baselines presents a significant stability issue when using highly corrupted data. As detailed in our analysis (§6.4), the performance of these models is capped by the quality of the initial ASR hypotheses. During training, low-SNR audio produces poor learning signal, which causes the model to learn to over-correct already accurate transcriptions while failing to fix genuinely erroneous ones. We have observed their performance degradation (up to +5% WER) when trained with the same data as DualHyp. To ensure stable convergence for our baseline comparisons, we therefore construct their training dataset exclusively from audio samples with SNR ≥ 0 dB, just as Ghosh et al. (2024); Hu et al. (2024b) have constructed their training datasets.

Our ROVER implementation conducts iterative alignment of hypotheses based on the 1-best hypothesis to construct a word transition network. We employ majority voting for each column with `<NULL>` tokens to prevent excessive insertion errors. Tie-breaking is resolved based on the rank order of the input hypothesis lists. No confidence scores or external LM weights are employed, as our setup relies solely on text hypotheses.

C DualHyp Analysis

C.1 Success Case

As a supplement to the cases presented in Table 2, Table 9 provides further qualitative examples that illustrate the successful mechanisms of our DualHyp framework. These successes can be categorized into two main patterns.

The first pattern, *Multimodal Fragment Composition*, involves the model's ability to recover correct transcriptions by leveraging complementary fragments from both ASR and VSR hypotheses. This can be seen when the framework fuses the beginning of an ASR hypothesis and the end of a VSR hypothesis as in the first case (*i.e.*, combining `which upset some...` from ASR and `...female residents` from VSR), or vice versa as in the second case (*i.e.*, `so rather than...` and `relying on...`). The compositional fusion of correct sub-sequences from noisy ASR and VSR inputs highlights the DualHyp's robustness by leveraging a generative correction ability of LLM.

The second pattern is *Dominant Modality Refinement*, where the model identifies and grounds the prediction in the more reliable modality, even when that modality's best hypothesis is not perfect. This is evident in `the armed forces` and `already in the states` cases, where the model primarily

Method	Utterance	WER (%)
	Type 1: Multimodal Fragment Composition	
ASR 5-best	which upset someone in the next day	71.4
	which i am saying some of you may or may not understand	128.6
	which upset someone who knew not what to do	100.0
	which i am saying is a lot easier than it is today	157.1
	which upset someone you know what i mean	85.7
VSR 5-best	we jumped at some of our female residents	42.9
	we jumped at some of our female races	57.1
	we jumps at some of our female residents	42.9
	we jump set some of our female residents	42.9
	we jumped at some of our female reasons	57.1
DualHyp output	**which upset some of our female residents**	**0.0**
Ground-truth	which upset some of our female residents	–
ASR 5-best	so what are the dangers of relying on this information	62.5
	so what are the dangers of relying on disinformation	87.5
	so my other thing is just relying on this information	50.0
	so what are the dangers relying on this information	50.0
	so my other thing is relying on this information	50.0
VSR 5-best	so rather than just regarding all this information	25.0
	so rather than regarding all this information	37.5
	so rather than to regard all this information	37.5
	so rather than just regarding this information	25.0
	so rather than argue this information	37.5
DualHyp output	**so rather than just relying on this information**	**0.0**
Ground-truth	so rather than just relying on this information	–
	Type 2: Dominant Modality Refinement	
ASR 5-best	the armed forces were	33.3
	the armed forces go	33.3
	and the armed forces were	66.7
	and the armed forces go	66.7
	in the armed forces but	66.7
VSR 5-best	i feel disco	100.0
	helpful disco	100.0
	i fell this	100.0
	time for disco	100.0
	helpful to this	100.0
DualHyp output	**the armed forces**	**0.0**
Ground-truth	the armed forces	–
ASR 5-best	is already in the states	25.0
	in the united states	50.0
	of the united states	75.0
	from the rest of the united states	100.0
	in one of the other states	75.0
VSR 5-best	already understand	75.0
	already understanding	75.0
	already understands	75.0
	already understanding that	100.0
	we are writing these things	125.0
DualHyp output	**already in the states**	**0.0**
Ground-truth	already in the states	–

Table 9: Successful examples of the GER process using DualHyp. Highlights illustrate how the final output is assembled from partial information scattered across the ASR (audio) and VSR (video) 5-best hypothesis lists. These cases illustrate two primary successful patterns: multi-modal fragment composition and dominant modality refinement.

Method	Utterance	WER (%)
	Type 1: Over-reliance on Plausible but Inaccurate Hypotheses	
ASR 5-best	what could be in the world	75.0
	to be in the world	50.0
	i can not believe the world	100.0
	it should be in the world	75.0
	good to be in the world	75.0
VSR 5-best	what we do is	100.0
	when we do this	125.0
	what we do here is	100.0
	what we do with this	125.0
	what we are doing is	100.0
DualHyp output	**what we do here is**	**125.0**
Ground-truth	**probably in the world**	–
ASR 5-best	that is what i am talking about	100.0
	i can not believe it	100.0
	i just can not believe it	42.9
	i am just going to come with you	85.7
	that is what i am saying	114.3
VSR 5-best	i just asked them to win	42.9
	i just asked him to win	42.9
	i just asked them to wait	57.1
	i just asked him to wait	57.1
	i just ask them to win	42.9
DualHyp output	**i just asked them to wait**	**57.1**
Ground-truth	**i just can not seem to win**	–
	Type 2: Hallucination and Semantic Association Errors	
ASR 5-best	no no no	100.0
	no no very good	166.7
	love love love	100.0
	no no	100.0
	i love november	66.7
VSR 5-best	it goes november	66.7
	november	66.7
	it is on november	66.7
	it is not november	66.7
	it is called november	66.7
DualHyp output	**november and december**	**100.0**
Ground-truth	**end of november**	–
ASR 5-best	this is the best bathroom downtown	50.0
	this is the best bathroom town in town	25.0
	this is the best bathroom in town	25.0
	this is the best bath hotel in town	12.5
	this is the best basketball town in town	50.0
VSR 5-best	this is the best bad hotel in town	12.5
	this is the best band hotel in town	12.5
	this is the best bat hotel in town	0.0
	this is the best pat hotel in town	12.5
	this is the best baton hotel in town	12.5
DualHyp output	**this is the best bistro in town**	**25.0**
Ground-truth	**this is the best bat hotel in town**	–

Table 10: Failure examples of the GER process using DualHyp. Green highlights illustrate the correct words from ground-truth, whereas red highlights illustrate wrong words from inference. These cases illustrate two primary error patterns: over-reliance on plausible but inaccurate hypotheses and hallucination based on semantic association.

Method	Utterance	WER (%)
ASR 5-best	to your baby of this year when she asked	100.0
	to your baby of this year when she asks	100.0
	there was your baby of this year when she asked	100.0
	there is your baby of this year when she asked	88.9
	to your baby of this year when she asked	100.0
VSR 5-best	there was no air so there was no sound	22.2
	there was no hit so there was no sound	33.3
	there was no heat so there was no sound	33.3
	there was no heart there was no sound	44.4
	there was no it so there was no sound	33.3
DualHyp output	**there was your baby of this year when she asked**	**100.0**

RelPrompt output	`Audio Mask (pred / GT):`
	`[N][N][N][N][N][N] / [N][N][N][N][N][N]`
	`Video Mask (pred / GT):`
	`[C][M][N][N][M][C] / [C][M][N][N][M][C]`
	there was no heat so there was no sound
		33.3

Ground-truth	there is no air so there is no sound	–
ASR 5-best	it is the same	80.0
	at the same time .	100.0
	at the same time	100.0
	which again opens the elements	60.0
	it is the same .	100.0
VSR 5-best	and it opens your eyes	100.0
	it opens to the enemies	60.0
	it opens to the animation	60.0
	and it opens to the enemies	80.0
	and it opens to the animation	80.0
DualHyp output	**it opens to the east**	**60.0**

RelPrompt output	`Audio Mask (pred / GT):`
	`[N][N][N][C][C] / [M][N][N][C][C]`
	`Video Mask (pred / GT):`
	`[C][N][N][N][C] / [C][M][N][N][C]`
	it opens to the elements
		40.0

Ground-truth	again open to the elements	–
ASR 5-best	like one hundreds of one thousands of people do every year	0.0
	like one hundreds or one thousands of people do every year	9.1
	one hundreds of one thousands of people do every year	9.1
	like one hundreds of one thousands of people do every year .	9.1
	like one hundreds and one thousands of people do every year	9.1
VSR 5-best	like one hundreds of one thousands of people or so every	27.3
	like one hundreds of one thousands of people or so every year	18.2
	like one hundreds of one thousands of people or so often	27.3
	like one hundreds of one thousands of people do every	9.1
	like one hundreds of one thousands of people or so whoever	27.3
DualHyp output	**like one hundreds or one thousands of people do every year**	**9.1**

RelPrompt output	`Audio Mask (pred / GT):`
	`[C][N][N][N][N][C][C] / [C][N][N][N][N][C][C]`
	`Video Mask (pred / GT):`
	`[C][C][C][C][N][N][C] / [C][C][C][N][N][N][N]`
	like one hundreds of one thousands of people do every year
		0.0

Ground-truth	like one hundreds of one thousands of people do every year	–

Table 11: Qualitative examples of successful corrections by DualHyp with RelPrompt. These cases show how RelPrompt improves upon the baseline DualHyp by leveraging the predicted reliability masks (`pred`) to trust or discard certain parts of hypotheses from the ASR and VSR streams. Ground-truth masks (`GT`) are also shown for comparison.

refines ASR's strong-but-flawed hypotheses while disregarding the less plausible VSR candidates. These cases highlight that providing the LLM with separate, modality-specific hypotheses is a more effective correction strategy than relying on a single or early-fused representation, as it allows the model to reason over distinct evidences.

C.2 Failure Case

Following the successful cases, we also present and analyze several typical failures, which often occur when both modalities provide highly ambiguous information. As illustrated in Table 10, these failures can be categorized into two primary patterns.

The first failure pattern is *Over-reliance on Plausible but Inaccurate Hypotheses*, where LLMs are misled by a semantically incorrect candidate from one modality. In the `probably in the world` example, the LLM disregards the partially correct ASR hypotheses and instead adopts the coherent but entirely wrong VSR hypothesis. Second example shows that the model favors the plausible but incorrect verb `to wait` from VSR candidates, although the correct verb `to win` is also present in the other VSR hypotheses. These cases show that the ambiguity between hypotheses can make the LLM confuse and incorrectly prioritize a plausible but wrong candidate. The over-reliance issue is a common drawback in all GER frameworks but can be mitigated to some extent by leveraging our RelPrompt, as shown in Figure 3.

The second failure pattern involves *Hallucination and Semantic Association Errors*, where the LLM generates words that are not present in any of the provided hypotheses. This often occurs when the model is biased towards a specific keyword and generates a semantically related but incorrect term, as seen in the `end of november` example where it generates `december` out of nowhere. In the last case, the model's strong prior knowledge can override direct evidence, misinterpreting `bat hotel` as `bistro`. This reveals the fundamental duality of leveraging the LLM's internal knowledge for GER, where context-aware corrections produce not only useful generative revisions but also factually incorrect hallucinations, suggesting a potential direction for future research on controlling this mechanism.

C.3 RelPrompt

Table 11 provides the qualitative examples that demonstrate how RelPrompt successfully corrects errors for the cases where baseline DualHyp framework fails. In the first example (`there is no air...`), DualHyp is misled by entirely incorrect ASR hypotheses (`your baby of...asked`). RelPrompt, in contrast, uses its predicted audio reliability tokens (all `[N]`) and rather clean video reliability tokens (`[C]`) to correctly identify the audio stream as unreliable, allowing it to pivot to the more accurate VSR hypotheses for a much better result.

In the second and third examples, the reliability masks guide the model to capitalize on the structure from the cleaner VSR stream at the beginning of the utterance, while correctly extracting a more accurate key phrase (*i.e.*, `the elements` and `do every year`) from the ASR stream to form the ending. The baseline DualHyp method, lacking this guidance, is confused by the conflicting signals and produces errors by incorporating some flawed hypotheses. These cases demonstrate how the explicit reliability signals empower the model to intelligently arbitrate between hypotheses at a sub-sentence level, composing the final output from the most reliable fragments of each modality.

D Additional Results

D.1 High-Resource Training

We investigate how our framework scales with additional training data by augmenting the main LRS2 training set (29 hours) with either the larger LRS3 dataset (59 hours) or a high-resource LRS2 pretraining set (HR, 195 hours). The results in Table 12 show that GER fails to benefit from more data, showing no to adverse impact on the performance. In contrast, our DualHyp frameworks consistently improve with larger training sets. The best performance is achieved by DualHyp + RelPrompt when trained with the high-resource data, reaching 12.8% overall WER on audio corruptions and 10.1% overall WER on visual corruptions. This indicates that while the bottleneck of single-stream GER is not readily resolved by scaling data, our compositional framework has the capacity to effectively leverage more data to enhance its robustness.

D.2 SNR-wise WER Improvement

Figure 5 provides the entire result of Figure 4 with WERR across all four audio noise types. Across all conditions, the single-stream baselines show that WERR is proportional to the audio quality, providing significant gains only at high SNRs. The DualHyp (A + AV) variant also illustrates this principle;

Method	+LRS3	+HR	Object occlusion (50%) Babble	Speech	Music	Natural	Overall	Speech noise (SNR = 0 dB) Object	Hands	Pixelate	Blur	Overall
GER	✗	✗	39.3	34.4	11.5	13.2	24.6	-	-	-	-	23.9
	✓	✗	39.1	34.3	11.7	12.8	24.5	-	-	-	-	23.9
	✗	✓	39.2	34.1	11.7	13.1	24.5	-	-	-	-	25.6
DualHyp	✗	✗	21.6	17.9	8.1	9.3	14.2	12.0	11.8	12.7	11.1	11.9
	✓	✗	21.0	17.9	7.8	8.7	13.9	12.7	11.4	12.3	11.2	11.9
	✗	✓	21.9	17.7	**7.7**	**8.0**	13.8	12.1	11.0	11.5	10.5	11.3
DualHyp +RelPrompt	✗	✗	20.4	16.0	8.0	8.2	13.2	11.9	11.0	11.9	10.2	11.3
	✓	✗	**19.5**	15.4	7.8	8.5	**12.8**	11.7	11.0	11.9	**9.6**	11.1
	✗	✓	20.1	**15.1**	**7.7**	8.3	**12.8**	**10.5**	**9.4**	**10.9**	9.7	**10.1**

Table 12: The effect of dataset integration (+LRS3) and high-resource (+HR) training.

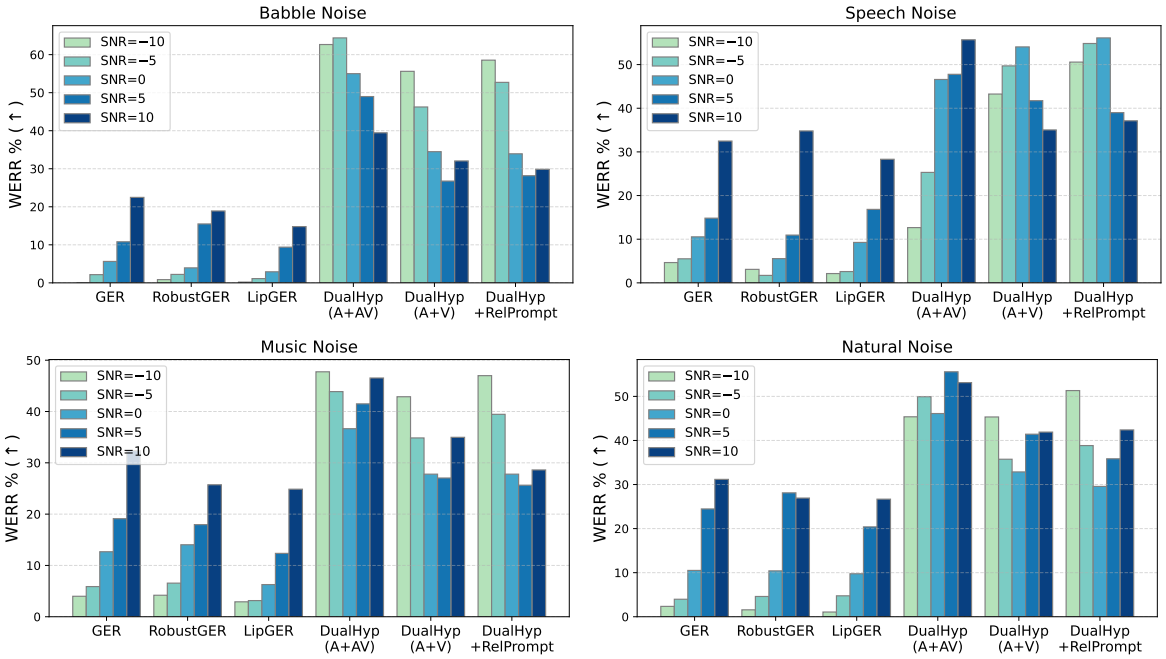

Figure 5: Word error rate reduction (WERR) at different audio SNRs, under diverse types of noise. Higher WERR indicates greater improvement over the Whisper ASR baseline. The experimental setup is identical to Table 3a.

Method	Input	Arabic	German	Greek	Spanish	French	Italian	Portuguese	Russian
Whisper-large-v3	A	**91.7**	**55.7**	**54.4**	49.6	**46.8**	52.3	52.7	**50.9**
mAV-HuBERT	V	$102.0_{(+11\%)}$	$96.6_{(+73\%)}$	$87.1_{(+60\%)}$	$70.5_{(+42\%)}$	$81.7_{(+75\%)}$	$73.7_{(+41\%)}$	$74.1_{(+41\%)}$	$80.9_{(+59\%)}$
GER	A	$96.9_{(+6\%)}$	$56.2_{(+1\%)}$	$57.7_{(+6\%)}$	$50.6_{(+2\%)}$	$47.8_{(+2\%)}$	$58.5_{(+12\%)}$	$52.3_{(-1\%)}$	$54.1_{(+6\%)}$
DualHyp (ours)	A+V	$106.8_{(+16\%)}$	$100.4_{(+80\%)}$	$77.3_{(+42\%)}$	**47.3**$_{(-5\%)}$	$47.9_{(+2\%)}$	**47.2**$_{(-10\%)}$	**49.0**$_{(-7\%)}$	$58.9_{(+16\%)}$

Table 13: WER (%) comparison with multilingual babble noise (SNR = 0 dB) on the MuAViC dataset. Subscript values of mAV-HuBERT indicate the relative WER increase compared to Whisper-large-v3.

it achieves a high WERR on familiar babble noise but does not show such strong correction capabilities on speech noise, especially at low SNRs. This demonstrates a key limitation of the early-fusion AVSR head: since it is affected by the audio corruption, it may fail to provide a truly independent and useful signal for error correction. In contrast, our DualHyp frameworks demonstrate superior robustness by maintaining high WERR even at very low SNRs, effectively leveraging the visual stream when the audio is most corrupted.

D.3 MuAViC Results

Table 13 presents our full multilingual results on the MuAViC dataset (Anwar et al., 2023). We observe that standard GER shows limited effectiveness across all languages, suggesting inherent difficulty of error correction in non-English contexts,

Method	Input	Babble (B)	Speech (S)	Music (M)	Natural (N)	Overall (O)
ASR oracle o_{nb} / o_{cp}	A	26.9 / 23.2	19.6 / 13.5	4.5 / 3.7	5.0 / 4.6	14.0 / 11.2
ASR + VSR oracle o_{nb} / o_{cp}	A + V	7.9 / 5.8	5.6 / 3.5	1.6 / 1.0	1.9 / 1.3	4.2 / 2.9
Whisper-large-v3	A	32.6	34.5	7.3	7.8	20.6
BRAVEn-large	V	-	-	-	-	$31.9_{(+54.9\%)}$
GER (Chen et al., 2023a)	A	$32.4_{(-0.6\%)}$	$35.4_{(+2.6\%)}$	$7.6_{(+4.1\%)}$	$8.0_{(+2.6\%)}$	$20.9_{(+1.5\%)}$
RobustGER (Hu et al., 2024b)	A	$32.5_{(-0.3\%)}$	$36.0_{(+4.3\%)}$	$7.7_{(+5.5\%)}$	$8.1_{(+3.8\%)}$	$21.1_{(+2.4\%)}$
LipGER (Ghosh et al., 2024)	AV	$32.4_{(-0.6\%)}$	$34.4_{(-0.3\%)}$	$7.6_{(+4.1\%)}$	$8.1_{(+3.8\%)}$	$20.6_{(-0.0\%)}$
GER w/ Auto-AVSR†	AV	$17.9_{(-45.1\%)}$	$45.6_{(+32.2\%)}$	$14.2_{(+94.5\%)}$	$11.0_{(+41.0\%)}$	$22.2_{(+7.8\%)}$
DualHyp (ours)	A + V	$16.3_{(-50.0\%)}$	$18.2_{(-47.2\%)}$	$\mathbf{5.6}_{(-23.3\%)}$	$5.5_{(-29.5\%)}$	$11.4_{(-44.7\%)}$
+ RelPrompt (ours)	A + V	$\mathbf{14.9}_{(-54.3\%)}$	$\mathbf{16.2}_{(-53.0\%)}$	$5.7_{(-21.9\%)}$	$\mathbf{5.1}_{(-34.6\%)}$	$\mathbf{10.5}_{(-49.0\%)}$

(a) Audio: random noise [-10, 10] dB, Video: 50% segment occluded with object

Method	Input	Object	Hands	Pixelate	Blur	Overall
ASR oracle o_{nb} / o_{cp}	A	-	-	-	-	8.3 / 5.3
ASR + VSR oracle o_{nb} / o_{cp}	A + V	3.2 / 1.8	3.0 / 1.6	3.0 / 1.5	2.5 / 1.4	2.9 / 1.6
Whisper-large-v3	A	23.8	23.8	23.8	23.8	23.8
BRAVEn-large	V	$31.9_{(+34.0\%)}$	$30.8_{(+29.4\%)}$	$29.5_{(+23.9\%)}$	$23.8_{(-0.0\%)}$	$29.0_{(+21.8\%)}$
GER (Chen et al., 2023a)	A	-	-	-	-	$26.0_{(+9.2\%)}$
RobustGER (Hu et al., 2024b)	A	-	-	-	-	$27.1_{(+13.9\%)}$
LipGER (Ghosh et al., 2024)	AV	$26.2_{(+10.1\%)}$	$26.0_{(+9.2\%)}$	$25.9_{(+8.8\%)}$	$26.0_{(+9.2\%)}$	$26.0_{(+9.2\%)}$
GER w/ Auto-AVSR†	AV	$47.5_{(+99.6\%)}$	$44.2_{(+85.7\%)}$	$42.0_{(+76.5\%)}$	$38.2_{(+60.5\%)}$	$43.0_{(+80.7\%)}$
DualHyp (ours)	A + V	$12.2_{(-48.7\%)}$	$10.9_{(-54.2\%)}$	$10.8_{(-54.6\%)}$	$9.6_{(-59.7\%)}$	$10.9_{(-54.2\%)}$
+ RelPrompt (ours)	A + V	$\mathbf{11.0}_{(-53.8\%)}$	$\mathbf{10.5}_{(-55.9\%)}$	$\mathbf{10.1}_{(-57.6\%)}$	$\mathbf{8.8}_{(-63.0\%)}$	$\mathbf{10.1}_{(-57.6\%)}$

(b) Audio: speech noise 0 dB, Video: random segment corrupted

Table 14: WER% (↓) results on the LRS3 test set under joint audio-visual corruption. (a) Performance across varying audio noise types, with a fixed visual corruption (50% segment occluded by an object). (b) Performance across varying visual corruption types, with a fixed audio corruption (0 dB speech noise). We also show the relative WER reduction in parentheses compared to the Whisper-large-v3 ASR baseline. All the ASR and VSR heads are Whisper-large-v3 and BRAVEn-large, respectively. †: We implement a GER model using hypotheses generated from an early-fusion approach, Auto-AVSR (Ma et al., 2023), which has been trained on LRS3 with babble noise.

even when the LLM itself is multilingual. Our DualHyp framework is designed to aid this reasoning by providing the LLM with more comprehensive evidence from both ASR and VSR streams, achieving performance improvements in three languages.

However, our results reveal that a large disparity between the ASR and VSR quality can exacerbate the LLM's inherent weakness in multilingual reasoning. While for the languages where DualHyp succeeds, the VSR head maintains a relatively consistent performance gap around 40% higher than the ASR baseline, for the languages where DualHyp underperforms (*e.g.*, Greek), this gap widens significantly to over 60%. Meanwhile, for Arabic, the hypotheses from both modalities are of exceptionally poor quality (>90% WER), leaving the LLM with no useful source to compose.

D.4 LRS3 Results

Similar to Table 3 and 8 in the main paper, Table 14 and 15 respectively present additional results on the LRS3 dataset (Afouras et al., 2018) to demonstrate

Method	Input	# hyps	B	S	M	N	O
GER	A	5	32.4	35.4	7.6	8.0	20.9
	AV	5	17.9	45.6	14.2	11.0	22.2
	AV	10	18.3	44.8	14.1	10.6	21.9
DualHyp	A + AV	10	19.3	34.8	6.0	5.4	16.4
	A + V	10	16.3	18.2	**5.6**	5.5	11.4
DualHyp + RelPrompt	A + AV	10	19.0	32.9	6.2	6.9	16.3
	A + V	10	**14.9**	**16.2**	5.7	**5.1**	**10.5**

Table 15: WER (%) comparison of different hypotheses from single-stream (GER) and dual-stream (DualHyp) generation heads, on the LRS3 dataset. The corruption strategy follows Table 14a.

the generalizability of our findings. A key difference from the LRS2 experiments is that our VSR head, BRAVEn-large, has also been fine-tuned on LRS3, making it a much stronger, in-domain model supporting the ASR stream. This serves to amplify the benefits of our dual-stream approach.

As shown in Table 14, our DualHyp + RelPrompt framework achieves an overall WER of 10.5% on audio corruptions and 10.1% on visual corruptions. The performance gap between our method and

GER w/ Auto-AVSR is even larger than on LRS2 (also refer to Table 15), confirming that as the quality of the independent VSR head improves, the advantage of our language-space fusion becomes more pronounced. We also observe that on LRS3, the ASR hypotheses, while coherent, are often homogeneous and contain similar errors across the N-best list. Our DualHyp approach is particularly effective in this case, as the independent VSR hypotheses provide the diversity to break out of the ASR's error patterns.

E LLM Usage

We acknowledge the use of LLM in this paper as a writing assistant, with its role strictly limited to supporting tasks such as paraphrasing, grammar check, and improving clarity. All conceptual contributions, method development, experimental design, and analysis are done without using LLMs or AI Agents.