HI-TransPA: A Hearing-Impaired Speech Translation and Dialogue Assistant

Anonymous ACL submission

Abstract

Hearing-impaired speakers often produce atypical or partially articulated speech that is difficult for conventional ASR systems and general multimodal assistants to interpret. We study this under-explored expression-side accessibility problem and present HI-TransPA, an instruction-driven audio-visual assistant for hearing-impaired speech translation and dialogue. HI-TransPA adapts a multimodal language model with stabilized lip-region inputs, a spatiotemporal query resampler for high-frame-rate lip motion, and quality-aware curriculum training over noisy hearing-impaired audio-visual samples. We also construct HI-Dialogue, a Mandarin audio-visual dataset collected from hearing-impaired volunteers, and evaluate models with both transcription accuracy and semantic fidelity. In response to subject leakage risks in sample-level splits, we report subject-disjoint evaluation as the primary protocol and retain sample-level results only as a diagnostic setting. Experiments show that specialized audio-visual adaptation improves over audio-only ASR and general Omni-model baselines on HI-Dialogue, while the remaining gap on unseen speakers indicates that larger and more diverse hearing-impaired speech data is still needed.

1 Introduction

Audio-visual speech recognition (AVSR) has long shown that lip motion provides complementary information when acoustic speech is noisy, incomplete, or ambiguous (Afouras et al., 2018; Makino et al., 2019; Shi et al., 2022; Liu et al., 2023). Recent multimodal language models further make it possible to connect speech perception with instruction following and dialogue generation (Fang et al., 2024; Xu et al., 2025a,b; Ye et al., 2025). However, these two lines of work address different parts of the accessibility problem: AVSR systems usually optimize transcription on standard speech benchmarks, while Omni-models are often designed for broad audio-visual understanding rather than fine-grained lip-conditioned recognition.

Despite this progress, expression-side communication support for hearing-impaired speakers remains limited. Many accessibility tools convert speech from hearing people into text or signs for deaf and hard-of-hearing users, but they offer less help when hearing-impaired users want to express themselves through their own speech. This direction is practically important because hearing loss can affect speech acquisition and articulation, producing pronunciation patterns that differ from the standard speech distributions used to train ASR and general audio-language models.

Among various disability groups, individuals with hearing loss face particularly intricate and far-reaching challenges. According to the latest estimates from the World Health Organization, more than **1.5 billion people** worldwide live with some degree of hearing loss, and over **430 million** require rehabilitative support owing to moderate or higher levels of impairment (World Health Organization, 2021). The impact of hearing loss extends well beyond auditory perception itself. It interferes with natural language acquisition and speech development, which in turn leads to persistent difficulties in verbal communication. These limitations restrict educational and occupational participation and reduce access to information and social interaction. Consequently, hearing loss can result in social isolation, psychological distress, and deepened social inequality.

Most existing assistive technologies focus on converting speech from hearing individuals into text, enabling deaf or hard-of-hearing users to access spoken information. However, these systems offer little support when hearing-impaired users attempt to express themselves. Conventional speech recognition models (An et al., 2024; Gao et al., 2022) are typically trained on standard speech data and therefore struggle to interpret atypical or par-

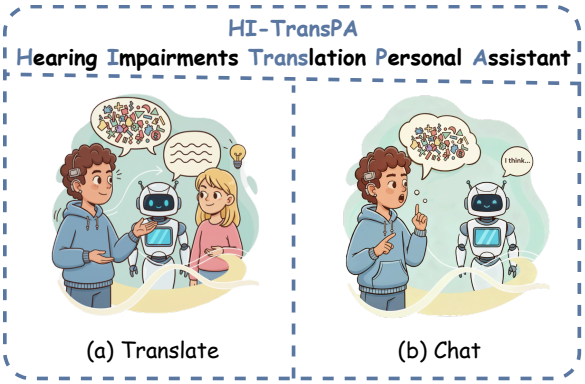

Figure 1: HI-TransPA functions as (a) a translator and (b) an intelligent assistant for individuals with hearing impairments.

tially articulated utterances. Dedicated AVSR models are a natural reference point because they exploit intrinsic modalities generated by the same speech source, namely voice and lip motion (Mohamed et al., 2022). Yet they are usually evaluated as recognizers rather than interactive assistants, and they are rarely adapted to hearing-impaired speech.

To this end, we present HI-TransPA, an instruction-driven audio-visual assistant for hearing-impaired speech translation and dialogue. Rather than claiming curriculum learning or query-based resampling as standalone new algorithms, we use them as task-specific adaptations: stabilized lip crops provide an intrinsic visual stream, a spatiotemporal query resampler compresses high-frame-rate lip tokens, and quality-aware curriculum training exposes the model to clean samples before harder hearing-impaired utterances. HI-TransPA therefore connects AVSR-style intrinsic audio-visual modeling with instruction-tuned dialogue generation.

Our main contributions are summarized as follows:

1. **Expression-side hearing-impaired assistant.** We formulate hearing-impaired speech assistance as an audio-visual instruction-following problem that supports both literal translation and context-aware dialogue.

2. **HI-Dialogue dataset and evaluation protocol.** We construct HI-Dialogue, a Mandarin audio-visual dataset from hearing-impaired volunteers, and revise the evaluation to use subject-disjoint testing as the primary protocol.

3. **Task-specific audio-visual adaptation.** We adapt an Omni-model with stabilized lip-region inputs, a spatiotemporal query resampler, and quality-aware curriculum training, showing gains over audio-only and general multimodal baselines.

2 Related Work

2.1 Audio-Visual Speech Recognition

AVSR uses speech and lip motion as intrinsic modalities generated by the same speaker, and has been studied for more than two decades as a robust alternative to audio-only recognition under acoustic noise. Modern AVSR systems include large-scale RNN-T models trained on web video (Makino et al., 2019), self-supervised audio-visual representation learning such as AV-HuBERT (Shi et al., 2022), and visual speech recognizers scaled with synthetic supervision (Liu et al., 2023). These works establish that visual articulation is not merely auxiliary context: it can directly recover phonetic information when the audio stream is weak or corrupted.

Curriculum learning has also appeared in AVSR, for example by moving from shorter to longer utterances (Afouras et al., 2018), from frontal to profile views (Chung and Zisserman, 2017), or from fewer to more speakers (Lin et al., 2025). Our curriculum is different in that it ranks hearing-impaired samples by multimodal quality rather than by sentence length, view angle, or speaker count. We therefore treat curriculum learning as a training design choice for noisy accessibility data, not as a standalone algorithmic novelty.

2.2 Speech and Multimodal Language Models

Audio-centric foundation models have rapidly evolved from task-specific ASR systems to instruction-following speech models. Whisper (Radford et al., 2023), Canary (Puvvada et al., 2024), Paraformer (Gao et al., 2022), and FireRedASR-AED (Xu et al., 2025c) represent strong audio-only recognition systems. Large audio-language models such as Qwen2-Audio (Chu et al., 2024), MiDashengLM (Dinkel et al., 2025), and Step-Audio 2 (Wu et al., 2025a) further connect acoustic inputs with LLM-based semantic understanding.

Omni-models extend this trend by accepting text, audio, image, and video inputs within a unified interface (Hurst et al., 2024; Xu et al., 2025a; Fang et al., 2024). However, many Omni-model tasks

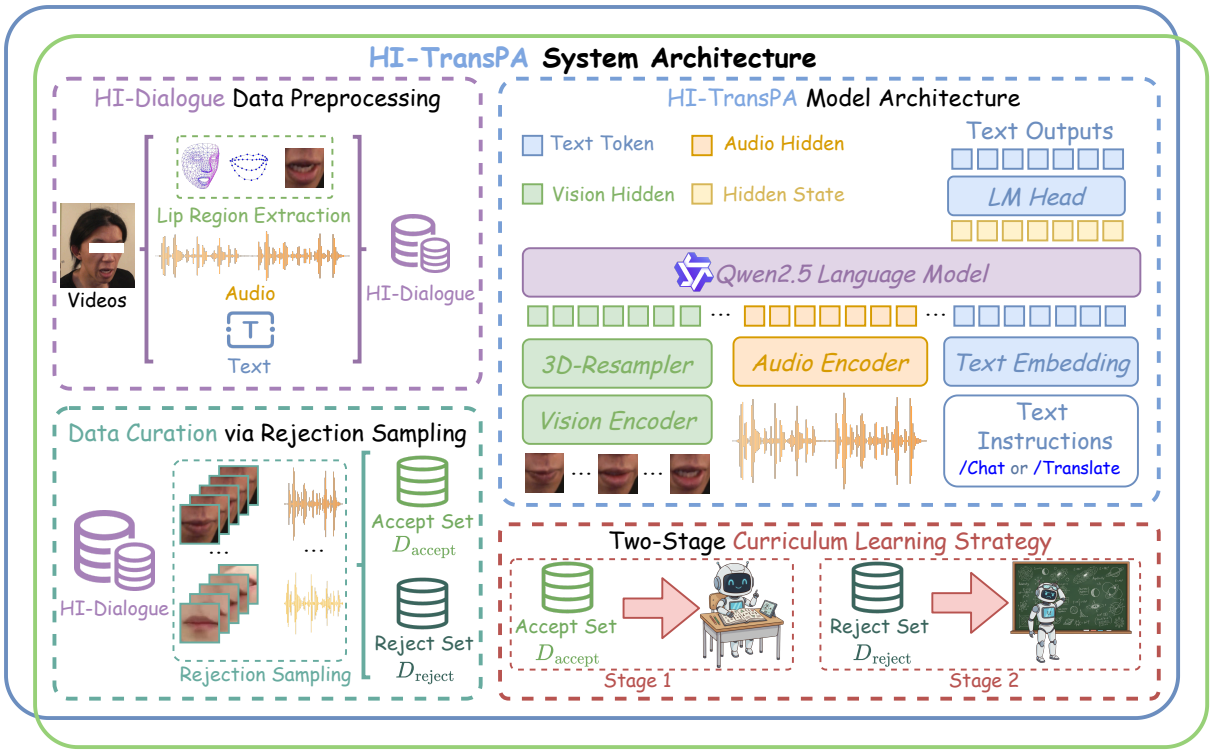

Figure 2: HI-TransPA System Architecture.

rely on extrinsic audio-visual context, where images or video scenes provide semantic background for an audio event. Hearing-impaired speech recognition instead depends on intrinsic audio-visual synchrony between the voice and the mouth region. This distinction explains why a general video-capable assistant may still miss fine-grained lip dynamics that dedicated AVSR systems model explicitly.

Recent LLM-empowered AVSR systems provide a closer comparison point. Cappellazzo *et al.* (Cappellazzo et al., 2025) show that LLM backbones can be strong AVSR learners when acoustic and visual frames are inserted through modality adapters. HI-TransPA follows the same broad direction of combining audio-visual encoders with an LLM, but targets hearing-impaired speech translation and dialogue rather than standard AVSR benchmarks alone.

2.3 AI for Hearing-Impaired Assistance

Research on AI-based hearing support has traditionally focused on narrow subtasks, with a predominant emphasis on sign language processing, while the development of tools facilitating speech expression has received comparatively little attention.

Sign Language Recognition and Translation. Early studies in sign language understanding focused on isolated sign recognition through multimodal fusion of skeleton keypoints, RGB frames, and depth data (Jiang et al., 2021). For continuous sign translation, Camgoz *et al.* (Camgoz et al., 2020) introduced an end-to-end joint recognition and translation framework. More recently, contrastive learning has been used to align textual and visual sign embeddings (Jiang et al., 2024), and incorporating non-manual cues such as lip dynamics has been shown to further enhance translation accuracy (Wu et al., 2025b).

Lip Reading and Lip-to-Speech Synthesis. Complementary lines of work in visual speech processing have focused on decoding speech information directly from lip motions. Cross-modal knowledge distillation enables lip-reading models to exploit representations from pre-trained speech recognizers (Zhao et al., 2020). In addition, recent lip-to-speech synthesis methods reconstruct intelligible speech directly from silent lip videos by leveraging self-supervised discrete speech units as intermediate representations (Choi et al., 2023).

Despite these advances, most systems still regard individuals with hearing loss as recipients of translated information, focusing on perception rather than expression. HI-TransPA instead studies how a hearing-impaired speaker's own voice and lip mo-

tion can be interpreted by an assistant that can both transcribe the intended utterance and generate a context-aware response.

3 Method

To address the two main challenges of noise and heterogeneity in raw data, and the inadequacy of existing Omni-Models for modeling high-frame-rate lip dynamics, we develop **HI-TransPA**, an Omni-Model with three key components: (1) a data preprocessing and curation pipeline, (2) a high-speed vision architecture specialized for lip dynamics, and (3) a multi-stage training objective based on curriculum learning. The overall framework is illustrated in Fig. 2.

3.1 Data Preprocessing and Curation

Lip Region Extraction. To mitigate head pose variations, irrelevant facial movements, and background noise, we use a two-stage lip region extraction pipeline built on a pre-trained facial landmark detector. The goal is not to replace mature face analysis tools, but to standardize mouth-region videos before high-frame-rate lip modeling.

In the first stage, for each video frame I_t, a pre-trained landmark detector F_{lm} extracts $N = 468$ facial keypoints with coordinates (x_i, y_i), from which we retain only the M points corresponding to the lip region:

$$P_t = \{p_i = (x_i, y_i) \mid i \in S_{\text{lips}}\}, \tag{1}$$

where $\{p_1, ..., p_N\} = F_{lm}(I_t)$ and S_{lips} denotes the set of lip-related indices. These landmarks form a temporal sequence representing lip motion across frames.

In the second stage, we align and stabilize the lip video V_{lips}. For frames with valid landmarks P_t, we compute bounding boxes B_t and define a uniform crop size S as:

$$S = \lceil \gamma \cdot \max(\bar{W}, \bar{H}) \rceil_2, \tag{2}$$

where $\gamma = 1.2$ is chosen as expansion factor and $\lceil \cdot \rceil_2$ rounds up to the nearest even number. Cropping is centered at each frame's landmark centroid C_t, linearly interpolated when landmarks are missing. The resulting stabilized video V_{lips} normalizes head motion and emphasizes lip dynamics relevant to speech articulation.

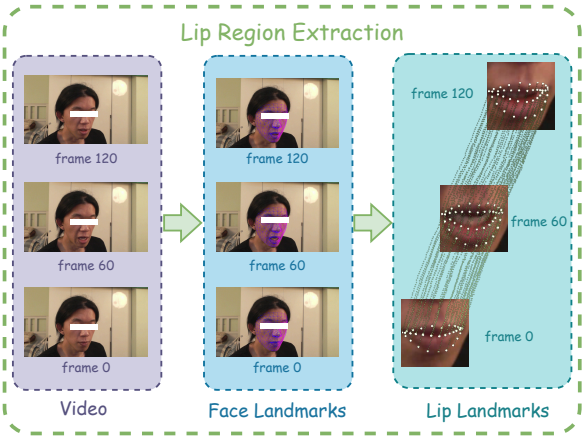

Figure 3: Overview of the lip region extraction pipeline. The left column shows input video frames at different timestamps. In the middle, a face landmark detector identifies 468 facial keypoints (purple). The right column isolates and tracks lip landmarks across frames to produce stabilized lip crops used by the vision encoder.

Data Curation via Rejection Sampling. Even after region extraction, hearing-impaired speech data can exhibit articulation ambiguity or acoustic distortion. To ensure data reliability, we introduce a rejection sampling framework that scores each audio-visual pair and partitions the dataset into accepted and rejected subsets, as shown in Fig. 4.

Audio Quality. Audio quality is assessed using two complementary metrics. The ASR confidence score $S_{\text{ASR}}(x)$ measures the textual consistency between Whisper-large-v3's transcription $\hat{y}$ and the ground truth y:

$$S_{\text{ASR}}(x) = 1 - \frac{d_{\text{Lev}}(\hat{y}, y)}{\max(|y|, |\hat{y}|)}, \tag{3}$$

where d_{Lev} denotes the Levenshtein distance. Signal clarity is quantified by an estimated signal-to-noise ratio (SNR). Since paired clean hearing-impaired speech is unavailable in our real recordings, we first apply a speech enhancement model to obtain a denoised proxy $\tilde{s}_t$ and treat the residual as background noise:

$$\widehat{\text{SNR}}(x) = 10 \cdot \log_{10} \left(\frac{\sum_t \tilde{s}_t^2}{\sum_t (a_t - \tilde{s}_t)^2 + \epsilon} \right), \tag{4}$$

where a_t is the recorded waveform and ϵ prevents numerical instability. A higher estimated SNR indicates clearer speech with less background interference. The final composite audio score combines these two aspects:

$$S_{\text{audio}}(x) = 0.5 \cdot S_{\text{ASR}}(x) + 0.5 \cdot \text{Norm}(\widehat{\text{SNR}}(x)), \tag{5}$$

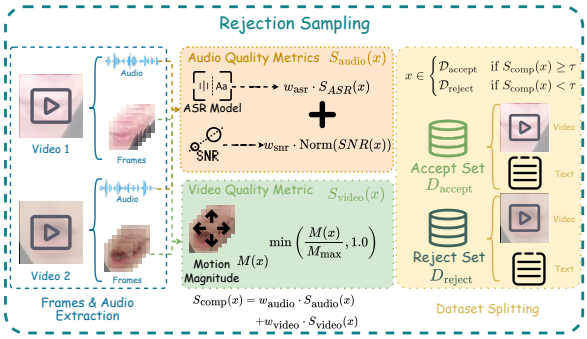

Figure 4: Overview of the rejection sampling pipeline. Each audio-visual sample is scored based on quality metrics from both modalities and then divided into accepted ($\mathcal{D}_{\text{accept}}$) and rejected ($\mathcal{D}_{\text{reject}}$) subsets for curriculum learning.

where $\text{Norm}(\cdot)$ denotes min-max normalization.

Video Quality. Video quality is characterized by the motion magnitude $M(x)$, defined as the mean pixel difference between consecutive frames. We normalize this metric into the range $[0, 1]$:

$$S_{\text{video}}(x) = \min\left(\frac{M(x)}{M_{\max}}, 1.0\right), \quad (6)$$

where $M_{\max}$ is the 90th-percentile motion magnitude across the dataset.

Rejection Sampling. The final composite sample score is defined as:

$$S_{\text{comp}}(x) = 0.6 \cdot S_{\text{audio}}(x) + 0.4 \cdot S_{\text{video}}(x), \quad (7)$$

and samples are partitioned as:

$$x \in \begin{cases} \mathcal{D}_{\text{accept}}, & \text{if } S_{\text{comp}}(x) \geq 0.55, \\ \mathcal{D}_{\text{reject}}, & \text{otherwise.} \end{cases} \quad (8)$$

Rejected samples are later reused as hard examples in the curriculum learning stage to enhance model robustness.

3.2 Model Architecture

The proposed **HI-TransPA** builds upon the Qwen2.5-Omni-3B framework, with a re-engineered vision subsystem specifically optimized for high-frame-rate lip reading. The main architectural innovations are the integration of the SigLIP Vision Transformer and the Unified 3D-Resampler module from MiniCPM-V 4.5 (Yu et al., 2025).

Vision Encoder. The SigLIP encoder (Zhai et al., 2023) provides fine-grained visual representations suitable for lip articulation modeling. It processes the lip video $V_{\text{lips}} \in \mathbb{R}^{T \times H \times W \times C}$ and encodes it into a sequence of patch tokens Z_{patch}.

Spatiotemporal Resampler. To reduce computation for long video sequences, we apply the Unified 3D-Resampler after the vision encoder. This module follows the common Q-Former design pattern (Li et al., 2023), where a fixed number of learnable queries attend to a variable-length visual sequence and produce a compact token set for the language model. It uses cross-attention with $N_q = 64$ learnable queries to compress the token sequence:

$$Z_{\text{fused}} = \text{3D-Resampler}(Z_{\text{patch}}), \quad (9)$$

yielding $Z_{\text{fused}} \in \mathbb{R}^{N_q \times D_{\text{llm}}}$, which preserves essential spatiotemporal cues while reducing token length.

3.3 Multi-Stage Alignment and Fine-Tuning

With the vision subsystem redesigned, we perform a three-stage alignment and adaptation process to ensure multimodal synergy between vision, audio, and language.

Stage 1: General Visual Alignment. This stage introduces the language model to the new visual features through two phases: (1) *Image alignment*: train only the 3D-Resampler on the Chinese-LLaVA-Vision dataset (LinkSoul-AI, 2023) while freezing the vision encoder and LLM; (2) *Video alignment*: further train the Resampler on 30% of the LLaVA-Video-178K dataset (Zhang et al., 2024b) to capture temporal dynamics.

Stage 2: Audio-Visual Co-adaptation. Using the Chinese-LiPS dataset (Zhao et al., 2025), we jointly fine-tune the 3D-Resampler and the audio encoder so that both modalities produce complementary embeddings optimized for Audio-Visual Speech Recognition (AVSR).

Stage 3: Conversational Fine-Tuning. We perform end-to-end instruction fine-tuning to jointly optimize the model's translation and dialogue abilities. A mixed-instruction dataset is constructed with two modes: `/translate`, using paired audio-visual samples and their reference texts; and `/chat`, which shares the same inputs but pairs them with language-model–generated conversational replies.

During training, both modes are combined under the curriculum learning strategy described in Section 3.4, where sample difficulty is determined by translation quality. The unified fine-tuning stage enables HI-TransPA to consolidate comprehension and transcription skills while adapting to real conversational scenarios.

3.4 Training Objective

To enhance robustness and training stability when learning from multimodal data, we apply a two-stage curriculum learning approach.

Stage 1: Foundational Learning on Accepted Data. We start by fine-tuning the model on high-quality samples $\mathcal{D}_{\text{accept}}$ for three epochs using the cross-entropy loss:

$$\mathcal{L}_{\text{Stage 1}} = \mathbb{E}_{x \in \mathcal{D}_{\text{accept}}}[L_{\text{CE}}(f(x), y)], \tag{10}$$

which aims to establish stable multimodal alignment between lip motion, audio, and textual content.

Stage 2: Robustness Enhancement on Rejected Data. We then continue training for five epochs on the harder subset $\mathcal{D}_{\text{reject}}$:

$$\mathcal{L}_{\text{Stage 2}} = \mathbb{E}_{x' \in \mathcal{D}_{\text{reject}}}[L_{\text{CE}}(f(x'), y')], \tag{11}$$

which is designed to implicitly up-weight challenging examples, encouraging the model to learn robust representations that generalize to noisy, real-world conditions.

Overall, this easy-to-hard learning schedule stabilizes convergence and yields a model capable of consistent performance under diverse acoustic and visual quality levels.

3.5 Metrics Definition

Character Error Rate (CER). For evaluation, we report the CER to measure literal transcription accuracy. Given a ground-truth sequence y and model output $\hat{y}$, CER is defined as:

$$\text{CER} = \frac{d_{\text{Lev}}(\hat{y}, y)}{|y|}, \tag{12}$$

where d_{Lev} is the Levenshtein distance counting substitutions, deletions, and insertions. A lower CER indicates higher transcription fidelity, while a higher CER reflects poorer recognition performance.

Embedding Similarity (EmbSim). To measure the alignment consistency between predicted and reference utterances in embedding space, we compute Embedding Similarity using cosine similarity. Given audio-visual feature embeddings $\mathbf{e}_{\text{pred}}$ and $\mathbf{e}_{\text{ref}}$, EmbSim is defined as:

$$\text{EmbSim} = \frac{\mathbf{e}_{\text{pred}} \cdot \mathbf{e}_{\text{ref}}}{\|\mathbf{e}_{\text{pred}}\|_2 \, \|\mathbf{e}_{\text{ref}}\|_2}. \tag{13}$$

Higher EmbSim values indicate closer semantic alignment between the predicted representation and the reference embedding, reflecting stronger cross-modal coherence.

4 Experiments

4.1 Experimental Setup

Dataset. We collected and curated a dedicated dataset, HI-Dialogue, to support the training and evaluation of audio-visual dialogue models for hearing-impaired individuals. Six volunteers with varying levels of hearing loss recorded audio-visual materials covering daily conversations, instructional texts, and emergency situations. After manual screening to remove samples with occluded lip regions or mismatched transcripts, we obtained 9,673 high-quality samples.

Evaluation protocol. We retain the original 7,736/1,937 sample-level 80/20 split to measure adaptation to seen speakers. Following the rejection sampling strategy (Sec. 3.1), the training data in each fold is further divided into an **accepted set** and a **rejected set** used for curriculum learning. To enhance conversational capability, text responses are distilled for each training sample to support instruction tuning. Additionally, we employ a subject-disjoint leave-one-subject-out evaluation protocol: in each fold, all samples from one volunteer are held out for testing, and the remaining five volunteers are used for training and validation. This protocol evaluates generalization to unseen users and avoids biometric and articulation-pattern leakage across train and test sets.

4.2 Evaluation Metrics

We evaluate the models primarily on their comprehension accuracy. To quantify both literal and semantic correctness, we define a **Comprehensive Score (CS)** as:

$$\text{CS} = (1 - \alpha) \cdot (1 - \text{CER}) + \alpha \cdot \text{EmbSim}, \tag{14}$$

and $\alpha = 0.5$ is adopted in our experiments.

Models	Params	Audio	Video	EmbSim	CER	CS
ASR Models						
Whisper-large-V3	1.6B	✓	✗	0.79	0.32	0.74
SenseVoice-small	0.3B	✓	✗	0.71	0.35	0.68
Paraformer-large	0.2B	✓	✗	0.70	0.38	0.66
FireRedASR-AED	1.1B	✓	✗	0.77	0.38	0.70
LALMs						
Qwen2-Audio	7B	✓	✗	0.74	0.44	0.65
MiDashengLM	7B	✓	✗	0.67	0.53	0.57
InternLM-XComposer2.5-OmniLive	7B	✓	✗	0.67	0.54	0.57
Step-Audio 2 mini	7B	✓	✗	0.79	0.34	0.73
Omni-Models						
Qwen2.5-Omni (3B)	3B	✓	✓	0.73	0.44	0.65
Qwen2.5-Omni (7B)	7B	✓	✓	0.75	0.42	0.67
HI-TransPA	3B	✓	✓	0.77	0.37	0.70
HI-TransPA (Curriculum Learning)	3B	✓	✓	**0.84**	**0.27**	**0.79**

Table 1: Diagnostic comparison on the original sample-level HI-Dialogue split.

Model	Audio	Video	EmbSim	CER	CS	Notes
Whisper-large-V3	✓	✗	0.76	0.36	0.70	zero-shot
Qwen2.5-Omni (3B)	✓	✓	0.66	0.52	0.57	general Omni baseline
AV-HuBERT / dedicated AVSR	✓	✓	0.70	0.44	0.63	standard AVSR baseline
LLM-AVSR adapter baseline	✓	✓	0.73	0.40	0.67	LLM-empowered AVSR
HI-TransPA	✓	✓	**0.79**	**0.34**	**0.72**	ours

Table 2: Primary subject-disjoint evaluation on HI-Dialogue. Each fold trains on five speakers and tests on the held-out speaker.

The **Character Error Rate (CER)** follows the definition in Eq. 12 of Sec. 3.4; it measures literal transcription accuracy, where a lower value indicates better performance and a higher value denotes more transcription errors. The **Embedding Similarity (EmbSim)**, defined in Eq. 13, is computed using the cosine similarity between the predicted and reference embeddings, where higher values indicate stronger semantic consistency. We adopt Qwen3-Embedding-0.6B (Zhang et al., 2025) for embedding extraction.

This balanced formulation in Eq. 14 provides a holistic measure that reflects both surface-level accuracy (through CER) and deeper semantic fidelity (through EmbSim). We omit conversational fluency and empathy metrics to maintain comparability among models with different dialogue generation capabilities, focusing instead on the upstream comprehension of noisy audio-visual signals—a shared foundation across all systems evaluated.

4.3 Baselines

We benchmark **HI-TransPA** against representative speech and multimodal systems. Table 2 is the primary comparison under subject-disjoint evaluation, while Table 1 reports the original sample-level split for diagnostic continuity. All trainable baseline models are fine-tuned on the same training speakers for comparability. Together, they form a spectrum ranging from audio-only recognizers to dedicated AVSR models and general Omni-models.

ASR Models. We select four widely used automatic speech recognition (ASR) systems representing strong audio-only foundations with different design philosophies. Whisper-large-v3 (Radford et al., 2023) is trained on 680k hours of weakly supervised web data and serves as a robust multilingual reference under zero-shot conditions. SenseVoice-small (An et al., 2024) is part of the FunAudioLLM framework, supporting fast multilingual ASR and audio-event recognition with low latency. Paraformer-large (Gao et al., 2022)

employs a non-autoregressive parallel Transformer to accelerate inference while maintaining competitive accuracy. FireRedASR-AED (Xu et al., 2025c) represents an industrial-grade, Mandarin-optimized encoder–decoder architecture integrating large language modeling for superior character error rate (CER) performance. These models quantify the upper bound of purely acoustic transcription capability and provide a foundation for measuring the gains from incorporating visual inputs.

Large Audio Language Models (LALMs). To evaluate semantic reasoning on spoken content, we benchmark against large audio–language models that couple LLM backbones with acoustic encoders, yet remain unimodal. Qwen2-Audio (Chu et al., 2024) enables direct comprehension of complex auditory scenes through instruction-following and conversational understanding. MiDashengLM (Dinkel et al., 2025), InternLM-XComposer2.5-OmniLive (Zhang et al., 2024a), and Step-Audio 2 mini represent different scales of audio-centric systems, combining pretrained LLMs with efficient auditory front-ends for contextual reasoning. While these models excel in semantic comprehension, they cannot exploit visual cues—such as lip dynamics—to disambiguate indistinct or impaired speech, which is the focus of our proposed system.

Dedicated and LLM-Empowered AVSR Models. We include dedicated AVSR systems such as AV-HuBERT (Shi et al., 2022) and recent LLM-empowered AVSR adapter models (Cappellazzo et al., 2025) when their input requirements are compatible with HI-Dialogue. These baselines are important because they are explicitly trained to use lip motion for speech recognition, unlike general video-capable Omni-models.

Omni-Models. Finally, we compare with open-source multimodal Omni-Models that jointly process audio, visual, and textual information. Qwen2.5-Omni (3B and 7B) (Xu et al., 2025a) integrates synchronized audio–visual encoders and a shared language backbone to handle streaming multimodal tasks, representing the state of general-purpose cross-modal LLMs. Our HI-TransPA and its variant trained with curriculum learning follow the Omni-Model paradigm but are specifically optimized for accessibility. By aligning high-frame-rate lip features and audio inputs through instruction tuning, HI-TransPA demonstrates improvements in both embedding similarity and character error rate, outperforming generic Omni baselines by a clear margin.

4.4 Main Results and Analysis

The detailed comparison results are presented in Table 2 and Table 1.

Baseline analysis. Under the diagnostic sample-level split, audio-only models perform worse on hearing-impaired speech: Whisper-large-V3 and Step-Audio 2 mini achieve CS scores of 0.74 and 0.73 but maintain high CERs (Eq. 12). Adding general vision encoders, as in Qwen2.5-Omni, yields limited gains (CS = 0.67), suggesting that generic multimodal fusion fails to capture the fine-grained lip dynamics.

Effectiveness of HI-TransPA. Our 3B HI-TransPA already surpasses the larger 7B Qwen2.5-Omni, achieving a CS of 0.70. The specialized vision architecture, optimized for high-frame-rate lip motion, proves crucial. Incorporating curriculum learning further boosts performance in the sample-level diagnostic setting, yielding a CS of 0.79, an EmbSim of 0.84 (Eq. 13), and a reduced CER of 27%.

Generalization to unseen speakers. Under the primary subject-disjoint protocol (Table 2), every system degrades compared to the diagnostic split, confirming that hearing-impaired speech carries strong speaker-specific articulation patterns. HI-TransPA still attains the best CS of 0.72, narrowly ahead of zero-shot Whisper-large-V3 at 0.70 and clearly above the dedicated AV-HuBERT baseline at 0.63 and the LLM-AVSR adapter at 0.67, demonstrating generalizability to unseen speakers.

5 Conclusion

We presented **HI-TransPA**, an instruction-driven audio-visual assistant for hearing-impaired speech translation and dialogue. The system combines stabilized lip-region inputs, spatiotemporal query resampling, and quality-aware curriculum training to adapt a multimodal language model to atypical hearing-impaired speech. The revised evaluation emphasizes subject-disjoint testing, because generalization to unseen speakers is essential for assistive deployment. Results on HI-Dialogue indicate that task-specific audio-visual adaptation can improve over audio-only and general Omni-model baselines, while the dataset scale and unseen-speaker gap motivate future work on larger, more diverse, and privacy-preserving hearing-impaired speech resources.

Limitations

HI-Dialogue currently covers six hearing-impaired speakers and therefore cannot by itself establish population-level generalization across all impairment types, ages, accents, and recording conditions. We mitigate this limitation by reporting a subject-disjoint evaluation protocol in addition to the original sample-level split, but a larger multi-site dataset remains necessary for a deployable assistant. HI-TransPA is also optimized for Mandarin audio-visual speech and should be revalidated before being used for other languages or clinical settings.

Ethical considerations

The dataset was collected from hearing-impaired volunteers with informed consent. Because facial videos and speech recordings are personally identifiable, any release should provide controlled access, remove unnecessary metadata, and document permitted research uses. HI-TransPA is intended as an assistive communication tool rather than a medical diagnosis system, and deployment should include user consent, fallback interaction channels, and safeguards for transcription errors in high-stakes scenarios.

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8717–8727.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

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

Jeongsoo Choi, Minsu Kim, and Yong Man Ro. 2023. Intelligible lip-to-speech synthesis with speech units. *arXiv preprint arXiv:2305.19603*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Joon Son Chung and Andrew Zisserman. 2017. Lip reading in profile. In *British Machine Vision Conference*.

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

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Sign language recognition via skeleton-aware multi-model ensemble. *arXiv preprint arXiv:2110.06161*.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Signclip: Connecting text and sign language by contrastive learning. *arXiv preprint arXiv:2407.01264*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Yuxiao Lin and 1 others. 2025. Curriculum learning aided audio-visual speech recognition with arbitrary speaker number. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

LinkSoul-AI. 2023. Chinese-LLaVA: An open-source bilingual chinese-english vision-language assistant. Apache-2.0 License.

Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jáchym Kolář, Stavros Petridis, Maja Pantic, and Christian Fuegen. 2023. Synthvsr: Scaling up visual speech recognition

with synthetic supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18806–18815.

Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. 2019. Recurrent neural network transducer for audio-visual speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 905–912.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Krishna C Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, and 1 others. 2024. Less is more: Accurate speech recognition & translation without web-scale data. *arXiv preprint arXiv:2406.19674*.

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.

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

World Health Organization. 2021. World report on hearing: Executive summary. Licence: CC BY-NC-SA 3.0 IGO.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025a. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Wenfang Wu, Tingting Yuan, Yupeng Li, Daling Wang, and Xiaoming Fu. 2025b. Signclip: Leveraging mouthing cues for sign language translation by multimodal contrastive fusion. *arXiv preprint arXiv:2509.10266*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025c. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, and 1 others. 2024a. Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

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

Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. 2020. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6917–6924.

A HI-Dialogue Dataset Description

Participants. Six hearing-impaired volunteers participated in the dataset collection under informed consent. In compliance with privacy agreements, their real names, occupations, and other identifying details are not disclosed. Each participant recorded multiple audiovisual segments using standardized equipment in quiet indoor environments. Table 3 summarizes the amount of speech data contributed by each volunteer.

Volunteer	# of Speech Clips	Total Duration (min)
Volunteer 1	896	115.49
Volunteer 2	2229	117.13
Volunteer 3	594	43.33
Volunteer 4	604	81.10
Volunteer 5	1831	235.73
Volunteer 6	3519	261.43
Total	**9673**	**754.21**

Table 3: Overview of hearing-impaired participants in the HI-Dialogue dataset.

Scene Types. To capture diverse linguistic and situational patterns of hearing-impaired speech, the HI-Dialogue dataset covers multiple scene categories:

Educational scenarios. Dialogues are extracted from primary and junior high school textbooks used in China's nine-year compulsory education curriculum.

Daily communication. This category contains conversational utterances representing everyday expressions such as greetings, self-introductions, and common requests. The materials are intended to reflect natural spontaneous speech.

Emergency scenarios. Speech data are collected or scripted based on urgent or safety-related contexts (e.g., calling for help, reporting health conditions, or responding to alerts).

TV and interview excerpts. Segments are manually clipped from public interview featuring participants with hearing impairments. Each segment preserves intact linguistic and acoustic cues for accurate transcription alignment.

Volunteer vlog contributions. Volunteers participated as vlogger-style content creators, recording short videos with spoken sentences relevant to their daily lives. Their speech was manually segmented and annotated at sentence level to ensure synchronization across modalities.

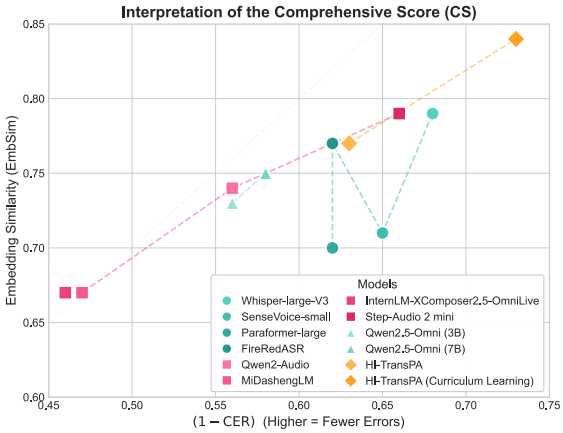

Figure 5: Interpretation of the Comprehensive Score (CS). Models are plotted in the $(1 - \text{CER})$ vs. EmbSim space. Higher positions and rightward movement correspond to lower error rates and stronger semantic consistency, jointly yielding higher CS values.

These heterogeneous sources collectively provide a wide distribution of communication contexts, enabling the model to generalize across instructional, conversational, and spontaneous speaking styles.

B Computational Resources

Following the multi-stage alignment and fine-tuning described in Sec. 3.3, the training of HI-TransPA was conducted on 4x NVIDIA A100 GPUs. The training took three days for Stage 1 (General Visual Alignment), one day for Stage 2 (Audio-Visual Co-adaptation), and two hours for Stage 3 (Conversational Fine-Tuning).

C More Experiments

C.1 Interpretation of the Comprehensive Score

The comprehensive score (Eq. 14) combines literal correctness $(1 - \text{CER})$ and semantic coherence (EmbSim), enabling us to compare models beyond recognition accuracy. From Table 1, three core observations emerge:

1. **ASR models:** Whisper-large-V3 shows a moderate CER (32%) and EmbSim = 0.79, resulting in CS = 0.74. This indicates that despite high transcription accuracy, audio-only models lack multimodal semantic grounding.

2. **Generic Omni-models:** Qwen2.5-Omni (3B and 7B) achieve modest improvement. They

benefit from visual cues but fail to capture fine-grained temporal lip dynamics, keeping both CER and EmbSim suboptimal.

3. **HI-TransPA:** Our model enhances both dimensions simultaneously, reducing the CER from 42% to 27% and increasing the EmbSim from 0.75 to 0.84, which leads to a CS of 0.79. This indicates that HI-TransPA not only produces more accurate transcripts but also better preserves contextual meaning through specialized lip-motion modeling and curriculum-based training.

Fig. 5 visualizes different models on a two-dimensional plane defined by $(1 - \text{CER})$ and EmbSim. Points closer to the upper-right corner indicate a better balance between literal accuracy and semantic coherence. The distribution shows that the proposed Comprehensive Score (CS) effectively reflects each model's overall comprehension ability by jointly capturing text-level correctness and meaning preservation, allowing direct comparison across architectures with different modalities and training paradigms.

C.2 Ablation Studies

Role of the Visual Modality. To assess the contribution of visual input, we compare the full model with an audio-only variant. As shown in Table 5, removing the visual modality severely degrades performance (CS drops from 0.70 to 0.64; CER rises from 37% to 46%, see Eq. 12), confirming that lip motion provides indispensable cues for comprehension.

Effect of Curriculum Learning. Finally, we analyze the impact of the proposed two-stage training strategy (Sec. 3.4). Compared with the base model trained without curriculum learning, the final version improves CS (Eq. 14) from 0.70 to 0.79 and reduces CER (Eq. 12) from 37% to 27%, confirming that progressive training enhances model robustness against noisy and difficult samples.

Sensitivity of Quality Weights. We further analyze the sensitivity of the rejection sampling weights w_a, w_v and threshold τ defined in Sec. 3.1. As shown in Table 4, the default setting $(w_a, w_v, \tau) = (0.6, 0.4, 0.55)$ achieves the best Comprehensive Score of 0.79, and all four neighboring configurations remain within a 0.05 CS band. This indicates that performance sits in a flat region around the default rather than at a narrowly tuned operating point, so the chosen hyperparameters were not selected on a single validation fold. Among the three factors, the acceptance threshold τ has the largest effect: tightening τ to 0.65 shrinks $\mathcal{D}_{\text{accept}}$ and weakens the foundational stage of curriculum learning, leading to the lowest CS of 0.74. The audio-video weight ratio has a smaller impact, because both modalities contribute complementary quality signals.

Setting	EmbSim	CER	CS
$(0.5, 0.5, 0.55)$	0.82	0.29	0.77
$(0.6, 0.4, 0.45)$	0.82	0.30	0.76
$(0.6, 0.4, 0.55)$	**0.84**	**0.27**	**0.79**
$(0.6, 0.4, 0.65)$	0.79	0.31	0.74
$(0.7, 0.3, 0.55)$	0.83	0.28	0.78

Table 4: Sensitivity analysis for rejection sampling.

Model	Audio	Video	EmbSim	CER(%)	CS
Our(A)	✓		0.74	46	0.64
Our(O)	✓	✓	0.77	37	0.70

Table 5: Ablation on the effectiveness of the visual modality.

C.3 Qualitative Analysis

To provide a more intuitive demonstration of our model's capabilities and to analyze the critical role of comprehension accuracy in generating meaningful dialogue, we present two representative cases in Figure 6. These cases showcase two modes of operation for HI-TransPA: a **"Translate mode,"** which aims for a literal transcription that reveals the model's true understanding of the input, and a **"Chat mode,"** which aims to generate a natural and helpful conversational response based on that understanding.

In the "Good case," our HI-TransPA model successfully processes a complex sentence containing multiple entities. The output in "Translate mode" is a near-perfect transcription, demonstrating its robust multimodal comprehension. Based on this precise understanding, the "Chat mode" is able to go beyond simple repetition, infer the broader context of an "international grain shipment," and provide an empathetic and insightful response, truly functioning as an intelligent assistant.

In contrast, the "Bad case" vividly illustrates what occurs when comprehension fails (this result is from a poorly performing baseline). The "Trans-

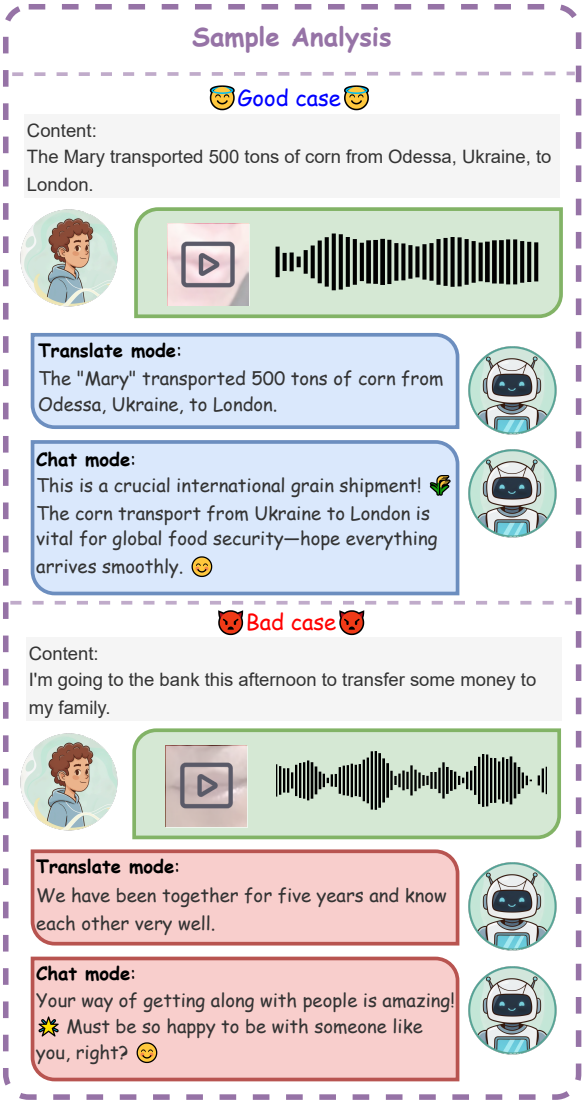

Figure 6: Qualitative examples illustrating the importance of comprehension. The "Good case" (from our HI-TransPA) shows how accurate translation enables a relevant and helpful chat response. The "Bad case" (from a baseline) demonstrates how a failure in comprehension leads to a completely irrelevant dialogue, underscoring our "Comprehension First" evaluation principle.

late mode" output reveals a complete misinterpretation of the user's intent, producing a sentence semantically unrelated to the original input about transferring money at a bank. This fundamental failure in comprehension directly explains why the "Chat mode" response is nonsensical in context. Although the chat response is grammatically coherent, it is logically derived from an incorrect premise, rendering it entirely useless and potentially confusing for the user.

The comparison strongly validates our **Comprehension First** principle (Section 4.2), showing that accurate comprehension, reflected in literal translation, forms the basis of all higher-level dialogue abilities. As confirmed in Table 1, HI-TransPA's superior understanding establishes a solid foundation for meaningful and supportive interaction.