UF-VSR: A PHONETIC–LEXICAL DECOMPOSITION FOR MANDARIN VISUAL SPEECH RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mandarin visual speech recognition (VSR) must infer characters from lip motion that primarily constrains syllable identity: visually similar articulations can denote different syllables, and one Pinyin syllable can denote many characters. We present UF-VSR, a two-stage system that makes this mismatch explicit. A shared visual encoder with character and toneless-Pinyin decoders first produces complementary lexical and phonetic hypotheses; a 4B language model adapted to model-generated VSR errors then refines the character hypothesis conditioned on both. The probabilistic decomposition is a modeling interpretation rather than a claim of exact marginal inference: the implemented system uses one best hypothesis from each decoder and the language model does not observe video. On CMLR, refinement reduces character error rate (CER) from 11.781 to 10.166 (13.71% relative) for our primary upstream and from 8.021 to 7.603 (5.21% relative) for a stronger upstream. Fixed-track CNVSRC-Multi.Dev CER decreases from 30.87 for a character-only model to 28.39 for the complete pipeline. Mechanism-oriented analyses show where the system succeeds and fails: reliable Pinyin is associated with larger gains, identical-Pinyin errors require context, and refinement ceases to help when either phonetic or character evidence is severely corrupted. These results support a practical Mandarin-specific interface between visual and language models while delimiting, rather than overstating, the evidence for its individual components.

1 INTRODUCTION

Visual speech recognition (VSR) transcribes silent lip-motion video. Better visual encoders and sequence models have substantially improved VSR (Ma et al., 2022b; Prajwal et al., 2022; Park et al., 2025), yet the observation is intrinsically incomplete: multiple phonemes share a viseme, co-articulation blurs their boundaries, and pose and speaker variation perturb appearance (Mattheyses et al., 2013). Mandarin adds a second ambiguity. Lip motion is more directly related to a spoken syllable than to a written Hanzi character, while a toneless Pinyin syllable can correspond to many characters. A direct video-to-character model must therefore solve visual–phonetic and phonetic–lexical ambiguity at once.

This distinction suggests a useful interface rather than a new probability identity. Let a visual model expose both what it believes was articulated (Pinyin) and what it believes was written (characters). A language model can then preserve the visually grounded character hypothesis, reject changes incompatible with the phonetic hypothesis, and use sentence context when pronunciation alone is insufficient. The interface is especially relevant to Mandarin: tones are predominantly conveyed by fundamental-frequency contours and are not directly visible, whereas toneless syllables retain segmental articulation; homophones then require lexical context.

Prior work already establishes important parts of this idea. LipCH-Net combines picture-to-Pinyin and Pinyin-to-Hanzi modules for Mandarin (Zhang et al., 2019); cascaded Mandarin VSR and semantic reconstruction likewise exploit intermediate structure (Zhao et al., 2019; Luo et al., 2025). VALLR predicts phonemes before LLM reconstruction (Thomas et al., 2025), while VSP-LLM injects visual speech representations into an LLM (Yeo et al., 2024). UF-VSR therefore does not claim that marginalizing an intermediate phonetic variable, or phoneme-to-text reconstruction broadly, is new. Its narrower contribution is a Mandarin-specific combination: parallel character/Pinyin pre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

diction retains a lexical anchor alongside a phonetic constraint, and VSR-error adaptation teaches a compact LLM to map their characteristic joint errors to text.

Figure 1 organizes the challenge along two axes. Pinyin may distinguish the target from a visually confusable source; if it does not, only sentence context can resolve the characters. Conversely, incorrect Pinyin can mislead refinement. This view leads us to evaluate not only aggregate CER but also evidence compatibility, Pinyin error rate, homophones, shared-final confusions, and upstream error severity.

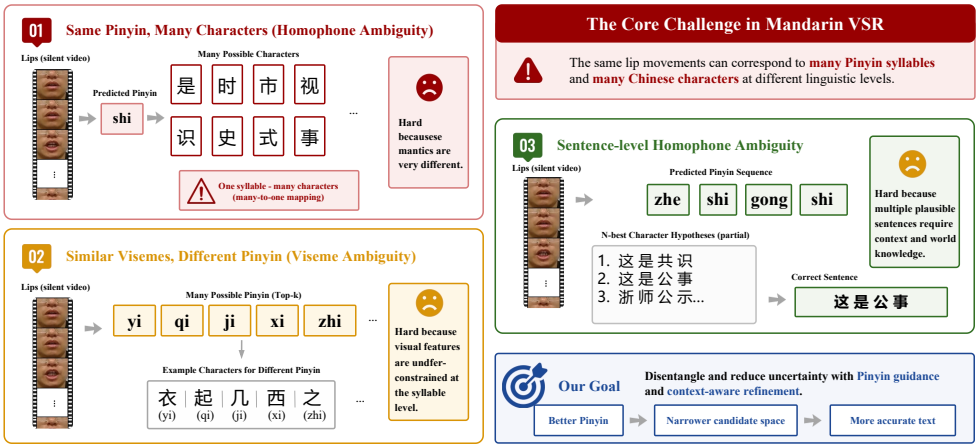

Figure 1: Mandarin VSR ambiguity at character, syllable, and sentence levels. The phonetic interface helps only when predicted Pinyin is informative and reliable; identical-Pinyin alternatives require lexical context. The arrow from Pinyin to characters is one-to-many.

Our contributions are:

- a dual-decoder Mandarin VSR model that predicts characters and toneless Pinyin from a shared visual representation, exposing both lexical and phonetic evidence;
- a single-best, Pinyin-and-character-conditioned refinement stage adapted using errors generated by VSR checkpoints of different quality; and
- evaluation on CMLR and fixed-track CNVSRC-Multi.Dev, plus mechanism-oriented analyses that identify the operating regime and failure modes of refinement.

Crucially, the present experiments establish the behavior of the complete interface, not the isolated causal contribution of every prompt field. We state this boundary throughout and identify the missing character-only intervention as the highest-priority future control.

2 Related Work

Visual speech recognition. Modern VSR uses convolutional or Transformer frontends with CTC, attention, or sequence-to-sequence objectives (Fenghour et al., 2021; Ma et al., 2022b; Prajwal et al., 2022). Mandarin systems have explored cascaded modeling, speech-recognizer distillation, hard-sample semantic reconstruction, and large-vocabulary benchmarks (Zhao et al., 2019; 2020; Luo et al., 2025; Chen et al., 2024; Liu et al., 2025a). These methods show that language structure matters, but stronger visual upstreams can produce larger gains than post-correction; we therefore report absolute and relative changes for two upstream qualities.

Phonetic intermediate representations. LipCH-Net's picture-to-Pinyin/Pinyin-to-Hanzi design is a direct Mandarin antecedent (Zhang et al., 2019). More recent systems use phonemes or visual speech units before reconstruction (Teng et al., 2025; Thomas et al., 2025). Cascade-free methods instead integrate intermediate representations to avoid committing to an erroneous phonetic sequence. That concern applies directly here: UF-VSR commits to one Pinyin hypothesis, so it is modular and

easy to adapt but can propagate errors. Our dual output partly mitigates this failure by retaining the character hypothesis; it does not eliminate it. Section 5 quantifies when the cascade becomes harmful.

Language-model refinement. LLMs have been used for ASR correction, rescoring, and VSR integration (Tur et al., 2024; Ma et al., 2025; Yeo et al., 2024; Liu et al., 2025b). Pinyin-aware correction is particularly relevant to Chinese homophone errors (Li et al., 2025). Unlike VALLR, which centers reconstruction on a phoneme sequence, UF-VSR supplies a parallel character hypothesis as an anchor. Unlike VSP-LLM, it is a modular text interface rather than an end-to-end visual-token interface. The trade-off is loss of visual uncertainty at refinement time.

Table 1: Positioning against the closest architectural precedents. "LM access" denotes what reaches the language-level reconstruction stage, not what is used by the visual encoder.

System	Target	LM access	Primary distinction
LipCH-Net	Mandarin	Pinyin	Pinyin-to-Hanzi cascade
VALLR	English	Phonemes	LLM phoneme reconstruction
VSP-LLM	English	Visual features	Visual–LLM integration
UF-VSR	Mandarin	Pinyin + characters	Parallel anchor and constraint

Table 1 makes the novelty boundary explicit. UF-VSR neither originates phonetic intermediates nor claims an end-to-end visual foundation model. Its distinguishing object is the *interface*: a Mandarin LLM receives a phonetic hypothesis without discarding the independently decoded character hypothesis. This favors modularity and inspectability over preserving continuous visual evidence.

3 METHOD

3.1 INTERFACE AND MODELING VIEW

For video $\mathbf{X} = \{x_t\}_{t=1}^{T}$, the target is a character sequence $\mathbf{Y} = \{y_l\}_{l=1}^{L}$. We introduce a toneless-Pinyin sequence $\mathbf{P} = \{p_m\}_{m=1}^{M}$. After punctuation normalization, each Mandarin character is paired with one syllable; the reference reading of a polyphonic character is determined in sentence context when Pinyin targets are constructed. We omit tones because lip video does not contain the acoustic pitch contour that principally distinguishes them. This choice reduces the visual target space but deliberately leaves tone-induced homophones for contextual resolution.

The law of total probability provides motivation:

$$P(\mathbf{Y} \mid \mathbf{X}) = \sum_{\mathbf{P}} P(\mathbf{Y} \mid \mathbf{P}, \mathbf{X}) P(\mathbf{P} \mid \mathbf{X}). \tag{1}$$

Equation 1 is not our inference algorithm or a theoretical novelty claim. The implemented LLM neither observes $\mathbf{X}$ nor marginalizes over $\mathbf{P}$. Instead, the visual model compresses its evidence into one character sequence and one Pinyin sequence:

$$\mathbf{X} \xrightarrow{f_{\mathrm{VSR}}} (\hat{\mathbf{Y}}_{\mathrm{vsr}}, \hat{\mathbf{P}}) \xrightarrow{f_{\mathrm{LLM}}} \hat{\mathbf{Y}}. \tag{2}$$

We call this a phonetic–lexical decomposition: a modeling interpretation that makes the two uncertainties inspectable, not exact uncertainty propagation.

3.2 PHONETIC-AWARE VISUAL INFERENCE

A 3D-convolutional frontend and ResNet-50 (He et al., 2016) extract frame features. A SANM encoder (Gao et al., 2020) in a Paraformer-based architecture (Gao et al., 2022; Shi et al., 2023) produces $\mathbf{H} \in \mathbb{R}^{T' \times d}$. A SANM character decoder and a lightweight Transformer Pinyin decoder (Vaswani et al., 2017) model $P(\mathbf{Y} \mid \mathbf{H})$ and $P(\mathbf{P} \mid \mathbf{H})$, respectively.

Each branch combines CTC and cross-entropy (CE):

$$\mathcal{L}_k = \lambda_{\mathrm{ctc}} \mathcal{L}_{\mathrm{ctc}}^{k} + (1 - \lambda_{\mathrm{ctc}}) \mathcal{L}_{\mathrm{ce}}^{k}, \quad k \in \{\mathrm{char}, \mathrm{py}\}. \tag{3}$$

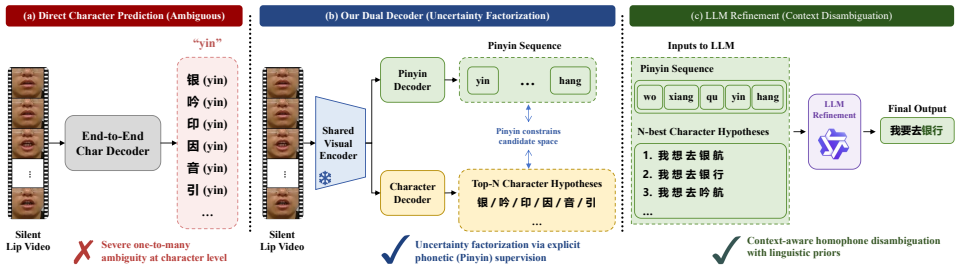

Figure 2: UF-VSR. A shared visual encoder produces one-best character and toneless-Pinyin hypotheses through parallel decoders. The adapted LLM receives exactly these two text fields and returns one transcription. Any Top-N alternatives shown inside the illustration are conceptual candidate spaces; all reported experiments use Top-1 inputs.

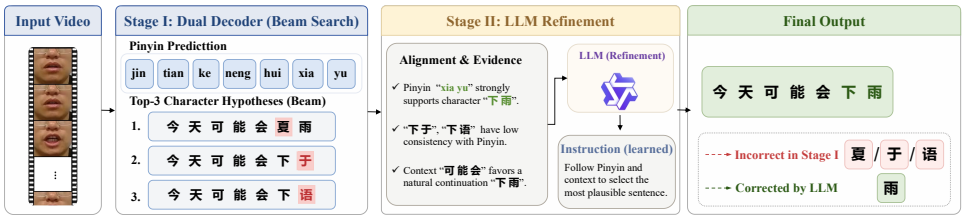

Figure 3: A representative refinement example. Stage I produces one-best Pinyin and character hypotheses; Stage II aligns their complementary evidence with sentence context and returns the corrected transcription. This is the actual Top-1 interface used in the experiments, rather than an N-best input.

The joint loss is

$$\mathcal{L}_{\text{VSR}} = \alpha \mathcal{L}_{\text{char}} + (1-\alpha)\mathcal{L}_{\text{py}}, \qquad \alpha = 0.5. \tag{4}$$

The same CTC/CE interpolation and decoding configuration is used for both branches and held fixed across the character-only and dual-decoder comparisons. Joint training encourages the shared encoder to retain syllable-level structure; at inference, the Pinyin output also becomes an explicit refinement input.

3.3 Constrained Language Refinement

The decoders return

$$\hat{\mathbf{Y}}_{\text{vsr}} = \arg\max_{\mathbf{Y}} P(\mathbf{Y} \mid \mathbf{H}), \qquad \hat{\mathbf{P}} = \arg\max_{\mathbf{P}} P(\mathbf{P} \mid \mathbf{H}). \tag{5}$$

The structured instruction contains (i) predicted toneless Pinyin, (ii) the one-best character hypothesis, and (iii) a directive to return only the corrected Mandarin sentence. The output is

$$\hat{\mathbf{Y}} = f_{\text{LLM}}(\hat{\mathbf{P}}, \hat{\mathbf{Y}}_{\text{vsr}}). \tag{6}$$

The character field anchors lexical content; Pinyin can veto phonetically incompatible changes; sentence context resolves alternatives compatible with both. Because both fields can be wrong, this is a soft learned constraint, not guaranteed constrained decoding.

Error-aware adaptation. We decode the VSR training data with early, middle, and late checkpoints. Each instruction triple $(\hat{\mathbf{P}}, \hat{\mathbf{Y}}_{\text{vsr}}, \mathbf{Y})$ exposes the LLM to severe, moderate, and residual VSR errors while also including cases that should be preserved. We adapt the LLM with LoRA (Hu et al., 2022): $W = W_0 + AB$, with rank $r \ll d$, under the autoregressive loss

$$\mathcal{L}_{\text{LLM}} = -\sum_{t} \log P_\theta(y_t \mid y_{<t}, \hat{\mathbf{P}}, \hat{\mathbf{Y}}_{\text{vsr}}). \tag{7}$$

This construction is motivated by coverage of VSR-specific errors. Since we do not compare against zero-shot, final-checkpoint-only, or matched-size generic correction data, results should be attributed to the adapted pipeline as a whole, not specifically to multi-checkpoint sampling.

4 EXPERIMENTS

4.1 DATA, METRICS, AND PROTOCOL

CMLR. CMLR contains 102,072 sentence utterances from 11 speakers (Zhao et al., 2019). We use 70,874 training utterances and the standard 20,418-utterance test set. Decoding three training checkpoints yields 212,622 instruction instances for LLM adaptation. No CMLR test reference is used for adaptation.

CNVSRC-Multi.Dev. The evaluation set contains 43 speakers: 23 recorded in controlled studios and 20 drawn from Internet video (Chen et al., 2024; Liu et al., 2025a). We use the fixed CNVSRC 2025 S3 resources: CN-CVS, CNVSRC.Dev, CN-CVS2-P1, and CN-CVS3 (Chen et al., 2023; Wang et al., 2025). We omit the earlier manuscript's "open" row because its additional resources were not specified well enough for a reproducible comparison.

Internal stress test. We additionally report an internal set with pose, illumination, speaker, and environment shift. The available experiment record does not contain publication-grade speaker, duration, demographic, collection, annotation, or split documentation. We therefore label these numbers as a secondary stress test, make no dataset contribution claim, and do not use them to support generalization to unseen speakers. Release and complete documentation are required before this set can serve as a reproducible benchmark.

We report character error rate $\mathrm{CER} = (S + D + I)/N$. For Pinyin analysis, PER is defined analogously after the same normalization. Character-level ambiguity analyses use only positions in examples for which input, output, and reference lengths agree, preventing insertion/deletion alignment from being counted as a substitution correction.

Implementation. The visual frontend uses 3D convolution and ResNet-50; the encoder is SANM. The character-only baseline has a SANM decoder, and UF-VSR adds a Transformer Pinyin decoder. We adapt Qwen3-4B-Instruct-2507 and Qwen3.5-4B (Qwen Team, 2025; 2026) with ms-swift (Zhao et al., 2024) and LoRA using learning rate 10^{-4}, rank 8, and scale 32. Inference removes fixed empty reasoning wrappers and applies no manual correction or lexical normalization. The two Qwen models test sensitivity within one model family and scale; they do not establish family- or scale-independent behavior.

4.2 MAIN RESULTS

Table 2: CER (%) on fixed-track CNVSRC-Multi.Dev and the secondary internal stress test. Lower is better. S1 = CN-CVS + CNVSRC.Dev; S2 adds CN-CVS2-P1; S3 adds CN-CVS3.

Method	Setting	Multi.Dev	Internal
2023 baseline (Chen et al., 2024)	S1	58.77	–
2024 baseline (Liu et al., 2025a)	S2	52.42	–
2025 baseline (Wang et al., 2025)	S3	31.91	40.21
Character-only baseline	S3	30.87	38.60
+ Pinyin decoder	S3	–	37.23
UF-VSR, complete	S3	**28.39**	**35.49**

Table 2 keeps training resources explicit. The character-only reimplementation improves over the 2025 baseline, and the complete system further lowers Multi.Dev CER by 2.48 absolute points. On the internal stress test, adding the Pinyin decoder lowers CER from 38.60 to 37.23 before LLM refinement, while the complete system reaches 35.49. The latter numbers are descriptive only because the set is not yet adequately documented.

Table 3: CMLR CER (%) for two character upstreams and adapted LLMs. Relative reduction is computed against the corresponding unrefined row.

Character upstream	Refinement	CER	Rel. ↓
Primary VSR	None	11.781	–
	Qwen3-4B	11.138	5.46
	Qwen3.5-4B	**10.166**	**13.71**
Ma et al. VSR	None	8.021	–
	Qwen3-4B	7.812	2.61
	Qwen3.5-4B	**7.603**	**5.21**

Table 3 separates upstream strength from refinement. The 13.71% headline reduction applies to the weaker primary upstream. With the stronger independently trained character upstream based on Ma et al. (2022a; 2023), the margin is 5.21%, and its unrefined 8.021 CER is already below the refined primary model. For this cross-upstream test, the character hypothesis is replaced while the Pinyin hypothesis remains supplied by our Pinyin branch. Thus this experiment measures portability of the refinement interface, not a fully independent second dual-decoder system. Upstream quality is the larger factor, while refinement provides a smaller consistent gain.

Table 4: Refinement behavior on the CMLR primary upstream. Sentence accuracy (S.Acc.), erroneous-sentence improvement (Imp.), and preservation of already-correct sentences (Pres.) are percentages.

Refinement	CER	S.Acc.	Imp.	Pres.
None	11.781	47.20	–	–
Qwen3-4B	11.138	56.06	37.24	–
Qwen3.5-4B	**10.166**	**58.08**	**44.45**	98.87

Qwen3.5 fully corrects 21.60% of erroneous first-stage sentences while preserving 98.87% of correct inputs. It also produces fewer regressions on erroneous inputs than Qwen3 (1,967 versus 2,434). This comparison indicates sensitivity to backbone quality, but both models are 4B Qwen variants; testing Llama, DeepSeek, and larger models remains necessary.

5 What Evidence Drives Refinement?

5.1 Pinyin Reliability and Compatibility

When Pinyin is entirely correct but the character hypothesis is wrong, Qwen3 and Qwen3.5 improve 59.54% and 72.51% of utterances, reducing CER from 10.180 to 5.441 and 3.738. Figure 4(a) shows that, among erroneous character hypotheses, correct rather than erroneous Pinyin coincides with improvement-rate increases of 25.5 and 32.1 percentage points. Figure 4(b) shows positive average gains at PER $\leq 20\%$ but degradation above 50%. Pinyin is therefore useful only within a reliability regime.

At aligned substitution positions, correction rates are 59.6%/69.9% when Pinyin supports the target and 20.2%/23.9% when it instead supports the erroneous source (Qwen3/Qwen3.5). This threefold association is consistent across the two backbones. It is not a controlled estimate of Pinyin's incremental effect: correct Pinyin may co-occur with easier video and a better character hypothesis. A matched character-only, Pinyin-only, masked/shuffled-Pinyin, and oracle-Pinyin intervention is required for causal attribution.

5.2 Two Axes of Mandarin Ambiguity

We distinguish whether phonetics separates source from target and whether sentence context can select the target. For identical-Pinyin character substitutions, phonetics is non-discriminative; corrections must come from lexical and sentential context. For different syllables sharing a final, partial articulatory similarity makes the visual problem hard, but predicted initials can help. Different-final

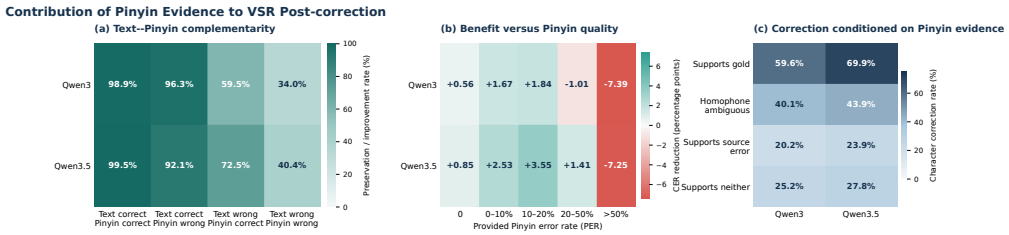

Figure 4: Pinyin evidence on CMLR. (a) Character–Pinyin complementarity. (b) CER change by Pinyin error rate; positive values denote improvement. (c) Correction conditioned on whether Pinyin supports the target or source. These are conditional associations, not prompt-field ablations.

errors serve as a less phonetically compatible control. Sentence-level homophones hold the entire Pinyin sequence fixed and test contextual reconstruction alone.

Table 5: CMLR ambiguity analysis (%). Character rows report aligned-position correction; the sentence row reports utterance improvement.

Ambiguity	Qwen3	Qwen3.5
Identical Pinyin, different character	40.19	**44.03**
Shared final, different Pinyin	30.29	**36.46**
Different-final control	25.87	**29.14**
Sentence-level homophone	50.45	**57.12**

Table 5 shows the same ordering for both LLMs. Identical-Pinyin substitutions reach 40.19% and 44.03% correction, demonstrating what context can recover when Pinyin cannot discriminate. Shared-final errors are corrected more often than the different-final control. On sentence-level homophones, refinement reduces CER from 9.002 to 6.047 with Qwen3 and 4.870 with Qwen3.5, with utterance improvement rates of 50.45% and 57.12%. These results support complementary roles: phonetic compatibility constrains some edits, while context resolves homophones.

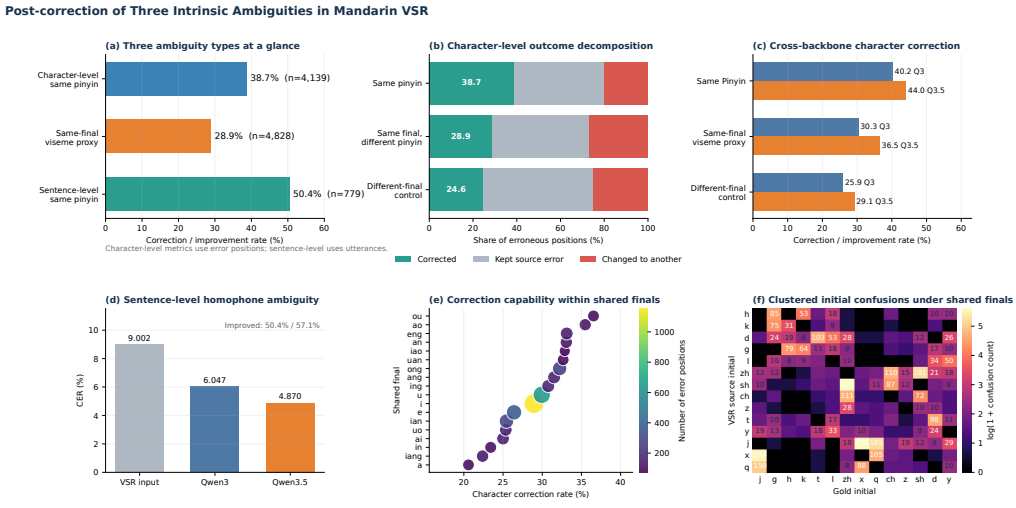

Figure 5: Mandarin-specific CMLR analysis. (a–c) correction outcomes for identical-Pinyin, shared-final, and control substitutions; (d) sentence-level homophones; (e–f) shared-final frequency, correction, and source–target initial confusions. Shared finals are a linguistic proxy, not a distance measured in visual feature space.

5.3 Failure Regime

For Qwen3.5, relative CER reductions are 29.52%, 25.05%, and 11.67% for input-CER intervals $(0, 10]$, $(10, 20]$, and $(20, 50]$. Above 50% input CER, the relative change is -0.30%. Together with degradation above 50% PER, this result bounds the method's useful regime: the LLM repairs local or moderate errors but cannot reconstruct a sentence when both anchors are badly corrupted. Fluency is not evidence of visual grounding.

The cascade offers modularity—the visual model and LLM can be trained and updated separately—but commits early to Top-1 hypotheses. A cascade-free system can preserve continuous visual evidence, and an N-best or confidence-weighted interface could propagate uncertainty more faithfully. Runtime and memory also increase because a 4B model is invoked after visual decoding; we do not yet report standardized latency or CER–cost curves.

6 Discussion and Limitations

The results justify three deliberately bounded claims. First, joint character/Pinyin training improves the secondary internal stress-test upstream (38.60 to 37.23 CER), although the dataset must be documented before that result can carry broad weight. Second, the complete adapted interface improves two CMLR character upstreams and fixed-track CNVSRC-Multi.Dev. Third, conditional analyses explain when predicted Pinyin agrees with successful correction and when it becomes harmful.

They do not establish that providing Pinyin causes the entire refinement gain. The decisive experiment—the same LLM, adaptation data, optimization, and character hypotheses with versus without Pinyin—is absent, as are Pinyin-only, shuffled-Pinyin, oracle-Pinyin, and no-LoRA controls. Likewise, the current data do not isolate multi-checkpoint error sampling. We consequently avoid attributing the 13.71% reduction to Pinyin alone or presenting error-aware construction as independently validated.

Generality is also limited. Qwen3 and Qwen3.5 share a family and parameter scale, so they are a within-family robustness check, not evidence across model families or scaling regimes. Toneless Pinyin matches visible segmental cues but discards tone; adding predicted tones without acoustic evidence could inject noise, while a confidence-aware lexical model might infer tone only when context supports it. Finally, the internal set requires a datasheet and release protocol. These limitations motivate concrete next steps rather than weakening the narrower conclusion supported here.

Interpreting the two inputs. The character and Pinyin fields should be understood as fallible views of the same video, not as a hierarchy in which Pinyin is always authoritative. Four regimes follow. When both agree and are correct, the safest action is preservation. When characters are locally wrong but Pinyin supports the target, phonetic compatibility can guide a repair. When Pinyin is non-discriminative—as for identical-Pinyin characters or entire homophonic sentences—the character hypothesis and language context must carry the decision. When both fields are badly corrupted, the posterior implied by the prompt is weakly grounded and fluent rewriting is unsafe. This taxonomy connects the aggregate results to a practical policy: invoke or trust refinement only when upstream confidence and cross-view agreement indicate that at least one anchor is reliable.

Qualitative behavior. The observed examples reflect these regimes. In a character-level homophone, the upstream selects a character meaning "person" where sentence context requires its identically pronounced alternative meaning "allow"; refinement can repair the character although Pinyin alone cannot choose between them. In a shared-final error, an upstream *bu* is corrected to *pu* in the phrase "representatives generally stated"; the shared final explains the visual confusion, while the predicted initial and context favor the target. By contrast, when both Pinyin and characters are heavily corrupted, the LLM may return a grammatical sentence unsupported by the video. These cases motivate selective refinement, explicit abstention, and calibration rather than unconditional post-editing.

Reproducibility scope. We report the retained training counts, fixed-track resources, LoRA settings, exact prompt fields, Top-1 protocol, normalization rule, and analysis filters available in the experiment record. We intentionally remove the underspecified open-resource comparison and do

not characterize the internal split as speaker-independent. The record does not preserve enough information to report run-to-run variance, significance tests, complete hardware and latency measurements, or all visual decoding hyperparameters. Accordingly, this paper makes empirical system claims rather than theorem, proof, or statistical-significance claims. A release should include configurations, seeds, checkpoint selection, the instruction-construction script, and a datasheet for any non-public evaluation set.

Claim–evidence alignment. The aggregate comparisons show that the *complete* interface improves its corresponding upstreams. The dual-decoder result supports a representation benefit only on the internal test, while the CMLR strata associate reliable Pinyin with correction; neither identifies Pinyin's incremental prompt-time effect. We therefore treat "Pinyin provides the complete gain," "multi-checkpoint sampling is superior," and "the method is LLM-family independent" as hypotheses. This avoids equating a correlated subgroup with an intervention, two same-family backbones with broad generalization, or an end-to-end gain with an unablated component.

Falsifiable implications. The mechanism yields concrete tests. Holding character hypotheses and optimization fixed, masked or shuffled Pinyin should hurt most when Pinyin supports the target; oracle Pinyin should define an upper bound that widens as predicted PER rises; and character-only refinement should approach the joint model on identical-Pinyin sentences, where phonetics is non-discriminative. If character-only matches UF-VSR across all strata, the gain belongs to generic error correction; if oracle Pinyin adds little, lexical reasoning or the character upstream is the bottleneck. These outcomes can confirm or falsify our explanation on the same visual test set.

Validity and deployment risks. Adaptation uses VSR training utterances and excludes evaluation references. Because CMLR and model pretraining are public, undocumented pretraining exposure still cannot be ruled out. Confidence-based invocation could reduce unsupported rewrites and computation, but calibration and gating are untested. A 4B refiner adds memory and latency, and speaker, demographic, or domain imbalance may affect both visual errors and lexical preferences. Deployment therefore requires per-speaker results, calibration, latency, and subgroup analysis in addition to CER.

7 Conclusion

UF-VSR exposes a Mandarin visual recognizer's phonetic and lexical hypotheses to an adapted language model. Its novelty is not the probability identity or phoneme-to-text reconstruction alone, but the combination of parallel character/Pinyin prediction, a retained lexical anchor, and VSR-specific refinement. The complete system reduces CER across the reported upstreams; gains concentrate where at least one input remains trustworthy. Pinyin constrains phonetically distinguishable alternatives, while sentence context resolves homophones. With Top-1 error propagation and component attribution stated explicitly, the evidence supports a useful Mandarin VSR interface and motivates confidence-aware decoding and controlled prompt ablations.

REFERENCES

C. Chen, D. Wang, and T. F. Zheng. Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

C. Chen, Z. Liu, X. Li, L. Li, and D. Wang. Cnvsrc 2023: The first chinese continuous visual speech recognition challenge. *arXiv preprint arXiv:2406.10313*, 2024.

S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao. Deep learning-based automated lip-reading: A survey. *IEEE Access*, 9:121184–121205, 2021.

Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proc. Interspeech 2022*, pp. 2063–2067, 2022. doi: 10.21437/Interspeech.2022-9996.

Zhifu Gao, Shiliang Zhang, Ming Lei, and Ian McLoughlin. SAN-M: Memory equipped self-attention for end-to-end speech recognition. In *Proc. Interspeech 2020*, pp. 6–10, 2020. doi: 10.21437/Interspeech.2020-1447.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Y. Li, X. Qiao, X. Zhao, H. Zhao, W. Tang, M. Zhang, and H. Yang. Large language model should understand pinyin for chinese asr error correction. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Z. Liu, X. Li, C. Chen, L. Li, and D. Wang. Cnvsrc 2024: The second chinese continuous visual speech recognition challenge. *arXiv preprint arXiv:2506.02010*, 2025a.

Z. Liu, X. Li, L. Guo, L. Li, and D. Wang. Leveraging large language models in visual speech recognition: Model scaling, context-aware decoding, and iterative polishing. *arXiv preprint arXiv:2506.02012*, 2025b.

D. Luo, D. Xie, Y. Zhang, W. Xie, and B. Sun. Hard sample semantic reconstruction for mandarin visual speech recognition. *Digital Signal Processing*, 160:105066, 2025.

P. Ma, S. Petridis, and M. Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4(11):930–939, 2022a.

P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic. Training strategies for improved lip-reading. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8472–8476. IEEE, 2022b.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-avsr: Audio-visual speech recognition with automatic labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096889.

R. Ma, M. Qian, M. Gales, and K. Knill. Asr error correction using large language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2025.

Wesley Mattheyses, Lukas Latacz, and Werner Verhelst. Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis. *Speech Communication*, 55(7–8):857–876, 2013. doi: 10.1016/j.specom.2013.02.005.

Y.-H. Park, R.-H. Park, and H.-M. Park. Swinlip: An efficient visual speech encoder for lip reading using swin transformer. *Neurocomputing*, pp. 130289, 2025.

K. R. Prajwal, T. Afouras, and A. Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5162–5172, 2022.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

X. Shi, Y. Yang, Z. Li, Y. Chen, Z. Gao, and S. Zhang. Seaco-paraformer: A non-autoregressive asr system with flexible and effective hotword customization ability. *arXiv preprint arXiv:2308.03266*, 2023.

M. K. K. Teng, H. Zhang, and T. Saitoh. Phoneme-level visual speech recognition via point-visual fusion and language model reconstruction. *arXiv preprint arXiv:2507.18863*, 2025.

M. Thomas, E. Fish, and R. Bowden. Vallr: Visual asr language model for lip reading. *arXiv preprint arXiv:2503.21408*, 2025.

A. D. Tur, A. Moumen, and M. Ravanelli. Progres: Prompted generative rescoring on asr n-best. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 600–607. IEEE, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

D. Wang, L. Li, Z. Liu, X. Li, K. Li, and H. Bu. Cnvsrc 2025: The third chinese continuous visual speech recognition challenge. https://cnceleb.org/competition, 2025. Accessed: 2025-10-01.

J. Yeo, S. Han, M. Kim, and Y. M. Ro. 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*, pp. 11391–11406, 2024.

Xingxuan Zhang, Hanjiang Gong, Xiyang Dai, Fan Yang, Nian Liu, and Ming Liu. Understanding pictograph with facial features: End-to-end sentence-level lip reading of chinese. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9211–9218, 2019.

Y. Zhao, R. Xu, and M. Song. A cascade sequence-to-sequence model for chinese mandarin lip reading. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pp. 1–6, 2019.

Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6917–6924, 2020.

Y. Zhao, J. Huang, J. Hu, X. Wang, Y. Mao, D. Zhang, Z. Jiang, Z. Wu, B. Ai, A. Wang, W. Zhou, and Y. Chen. Swift: A scalable lightweight infrastructure for fine-tuning. *arXiv preprint arXiv:2408.05517*, 2024. URL https://arxiv.org/abs/2408.05517.