Anchoring Speech with Semantics: A Multimodal Adapter Mechanism for Automatic Speech Recognition in Low-Resource Languages

Kuan-Tang Huang[1,3,*], Cheng-Yeh Yang[1,*], Chien-Chun Wang[1]
Hung-Shin Lee[1,†], Hsin-Min Wang[2], Berlin Chen[1]

[1]National Taiwan Normal University, Taiwan
[2]Academia Sinica, Taiwan
[3]EZAI, Taiwan
[*]Equal contribution [†]Corresponding author

Abstract

Low-resource ASR remains difficult because scarce transcripts provide limited supervised evidence for target-side generation. To address this gap, we propose SAMA-ASR, a lightweight adapter mechanism that augments the decoder with semantic anchors from auxiliary translations and an acoustic anchor from speech; in principle, the mechanism can be applied to similar encoder–decoder multitask speech models. Through cross-modal adaptation, SAMA-ASR conditions decoder states on translation-derived semantic embeddings and a speech embedding, combining utterance-level meaning with speech-grounded evidence before token prediction. At evaluation time, these semantic anchors can be generated automatically by an upstream speech-to-text translator rather than supplied as oracle translations. Experiments on two 30-hour datasets covering the low-resource Sinitic varieties Taiwanese Hokkien and Hakka show that SAMA-ASR improves over acoustic, prior prompt-based, and semantic-only translation-guided baselines and remains effective in practical automatic semantic-anchor settings; translator-capacity analyses show that useful semantic anchors can be produced by a compact ST model.

1 Introduction

Automatic Speech Recognition (ASR) aims to transcribe acoustic speech signals into accurate text sequences. Although modern encoder–decoder architectures have substantially advanced ASR (Watanabe et al., 2017), robust recognition for many low-resource languages and dialects remains a major gap because labeled speech is scarce (Cheng et al., 2025; Yang et al., 2025; Li and Niehues, 2026).

On the encoder side, multilingual speech pretraining can partially mitigate acoustic scarcity by transferring shared phonetic representations (Conneau et al., 2021; Bapna et al., 2022). However, when only a few target-language transcripts are available, the decoder receives limited supervised evidence for target-variety lexical choices, orthographic conventions, common constructions, and meaning-compatible continuations.

To address this problem, we propose the Semantic-Aware Multimodal Adapter for ASR (SAMA-ASR)[1], a lightweight decoder-side module that conditions ASR decoding on two complementary signals: semantic anchors encoded from auxiliary translations and an acoustic anchor derived from the speech signal. Concretely, SAMA-ASR lets decoder hidden states query both translation-derived semantic embeddings and a speech embedding at each block, so local token decisions use utterance-level meaning while remaining tied to acoustic evidence. This explicit acoustic path acts as a grounding constraint: it discourages semantically plausible but acoustically unsupported continuations when semantic anchors are noisy or incomplete. Using translations as semantic anchors is practical because such dominant- or standard-language text appears across diverse low-resource speech settings, often when target-variety transcripts are unavailable (Xiao et al., 2023; Plüss et al., 2023; Zanon Boito et al., 2022). From an implementation perspective, SAMA-ASR leaves the pretrained backbone frozen and, as an adapter, can in principle be attached to other end-to-end encoder–decoder ASR architectures.

This semantic-anchoring design is also useful for autoregressive decoding: semantic anchors are supplied before generation, providing global meaning context before partial hypotheses become reliable and reducing exposure to error accumulation (Bengio et al., 2015; Arora et al., 2022).

For practical inference without oracle translations, we introduce an upstream speech-to-text translation (ST) model to generate semantic an-

[1]Code is available at https://github.com/610494/sama.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 31516–31531
October 24-29, 2026 ©2026 Association for Computational Linguistics

chors automatically. Experiments show that a compact Whisper Small ST model already yields clear gains; compared with the frozen Whisper Medium ASR backbone used by SAMA-ASR (769M parameters), this 244M-parameter semantic-anchor generator keeps the added model cost moderate. Closest to this setting, prior translation-guided ASR improves Taiwanese Hokkien ASR with Mandarin auxiliary translations, but assumes that oracle translations are provided and lacks an explicit acoustic-anchor path (Yang et al., 2026). In summary, the main contributions of this study are four-fold:

- **A semantic-anchoring formulation for low-resource ASR:** We recast extremely low-resource encoder–decoder ASR as a target-side generation problem and identify paired speech–translation data as a practical supervision source for improving decoder decisions when target-language transcripts are scarce.
- **A lightweight multimodal adapter architecture:** We introduce SAMA-ASR, a decoder-side adapter that lets autoregressive states attend to both translation-derived semantic anchors and an explicit acoustic anchor, improving semantic conditioning while preserving speech-grounded transcription and keeping the pretrained backbone frozen.
- **A realistic non-oracle inference pipeline:** We study automatic semantic-anchor generation with upstream ST models, showing that SAMA-ASR does not require oracle translations at test time and that compact translators can provide useful anchors with moderate added cost.
- **Cross-dataset validation and auxiliary-language analysis:** We validate SAMA-ASR on two low-resource Sinitic target varieties, Taiwanese Hokkien and Hakka, and provide analyses across data scales, translator capacities, auxiliary languages, and multilingual anchor compositions to clarify when semantic anchoring helps.

2 Related Work

2.1 Low-Resource ASR

Data scarcity remains a primary bottleneck for under-resourced ASR, including dialectal settings (Besacier et al., 2014; Cheng et al., 2025; Yang et al., 2025). Standard mitigation strategies include acoustic augmentation (e.g., speed perturbation and SpecAugment (Ko et al., 2015; Park et al., 2019)) and resource construction from loosely aligned media or audiobook sources (Chen et al., 2020; Yeroyan and Karpov, 2024). Such resource-construction pipelines reduce annotation cost, but subtitle or long-form alignment can introduce label and segmentation noise.

A complementary direction designs ASR algorithms around the linguistic properties or resources of a specific target language, such as tone-aware modeling, character-set adaptation, or language-specific transfer strategies (Coto-Solano, 2022; Getman et al., 2024; Peng et al., 2026). These language-specific studies show that target-language expertise can be valuable, but their assumptions may not transfer unchanged to languages with different phonology, scripts, orthographic conventions, or related-language resources. In contrast, SAMA-ASR is not tied to a handcrafted linguistic feature or a single target language: our experiments show consistent gains on both Taiwanese Hokkien and Hakka, and our auxiliary-language analyses further suggest that semantic anchors can remain useful across translation languages when anchor quality and composition are favorable.

Joint ASR–ST and Parameter-Efficient Adaptation. Joint ASR–ST models exploit the complementarity between recognition and translation through jointly optimized decoders and cross-task information exchange (Le et al., 2020). Separately, Meta-Adapter and LoRA-Whisper improve low-resource or multilingual ASR through parameter-efficient adaptation (Hou et al., 2021; Song et al., 2024). SAMA-ASR is complementary to these directions: rather than jointly optimizing ASR and ST objectives, it keeps the backbone frozen and injects translation-derived semantic and speech-derived acoustic evidence into the decoder, while its combination with LoRA shows that semantic–acoustic anchoring can further complement parameter-efficient adaptation.

2.2 Text-Guided ASR

Contextual biasing strengthens ASR decoders by injecting target-language priors or phrase-level constraints during recognition (Zhao et al., 2019; Le et al., 2021; Sun et al., 2023). These contextual-biasing methods are effective when the domain is known, but they usually require predefined phrase lists, target-language text, or task-specific biasing inventories, making them less suitable for open-domain low-resource ASR. LLM-based ASR fur-

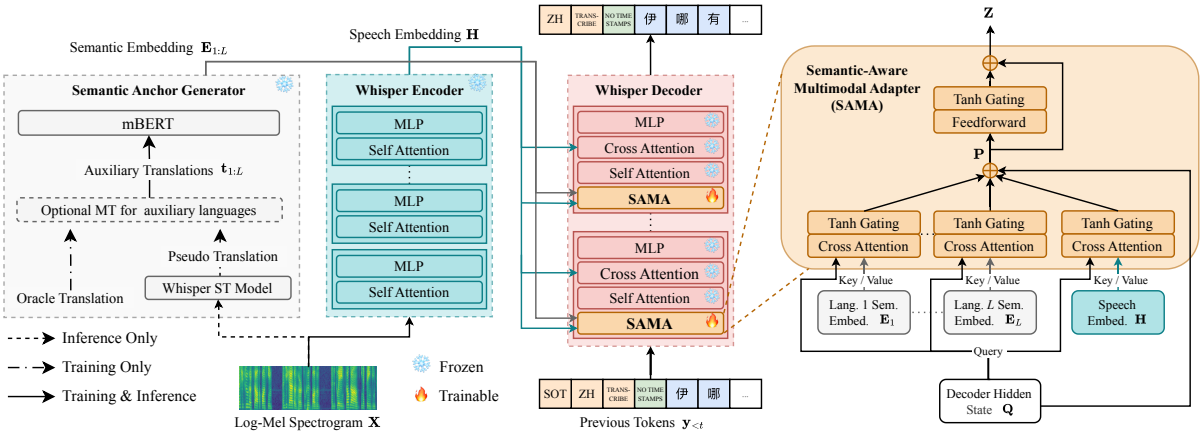

Figure 1: **Overview of the SAMA-ASR framework.** Snowflakes mark frozen modules and flames mark trainable modules. Automatic translations are produced by ST, encoded by mBERT, and fused with the frozen Whisper speech embedding through gated cross-attention. MT is used only in the optional multilingual diagnostic setting.

ther brings reasoning and instruction-following ability (Chen et al., 2024; Hsu et al., 2025), but at billion-parameter-scale decoding or fusion cost. Closest to our work, TG-ASR uses translation embeddings for Taiwanese Hokkien ASR (Yang et al., 2026), but assumes translations are available at inference time. SAMA-ASR instead targets a more general setting: the method does not require a predefined domain, phrase list, or oracle translation at test time, because semantic anchors can be generated automatically by a million-parameter-scale ST model. SAMA-ASR then uses these anchors as utterance-level semantic guidance while adding an explicit acoustic-grounding path, so decoder decisions remain tied to the speech signal rather than relying on translation-only guidance.

Distillation vs. Anchoring. Text-guided speech distillation transfers linguistic knowledge from pretrained text or language-model teachers into ASR models during training (Choi and Park, 2022; Hentschel et al., 2024). Depending on the method, the teacher signal may be latent representation alignment or language-model token-probability targets, but it is used as a training-time supervision signal rather than an inference-time input. By contrast, SAMA-ASR uses translated text as a runtime semantic anchor and combines it with an acoustic anchor through decoder-side adapter modules during decoding.

Multi-Source Text Fusion. SAMA-ASR also differs from source-text-assisted ASR for simultaneous interpretation, which uses an auxiliary source-language text encoder alongside the speech encoder (Taniguchi et al., 2022). That setting assumes source text is available at recognition time, whereas our practical setting must first generate auxiliary translations from speech.

Cross-Modal Adapter Design. SAMA-ASR's parameter-efficient design follows the same broad family as Flamingo-style gated cross-attention adapters and Whisper-Flamingo (Alayrac et al., 2022; Rouditchenko et al., 2024), while its multimodal grounding motivation is related to audio-visual ASR fusion work (Sterpu et al., 2018; Ma et al., 2021). However, rather than fusing visual cues with speech, SAMA-ASR fuses semantic text anchors with an acoustic feature stream to separate meaning-level guidance from surface-form grounding. Because SAMA-ASR attends over token-level semantic embeddings from mBERT (Devlin et al., 2019), it preserves local translation structure instead of relying only on sentence-level semantic vectors such as LaBSE (Feng et al., 2022).

3 Methodology

We present SAMA-ASR as a lightweight decoder-side adapter mechanism for enhancing low-resource ASR with semantic and acoustic anchors. The central idea is: scarce target-language transcripts make the decoder's next-token distribution difficult to estimate reliably, so SAMA-ASR conditions decoder states on translation-derived semantic embeddings and a speech embedding, using the semantic embeddings for utterance-level meaning and the speech embedding to keep decoding grounded when translations are noisy or incomplete. At inference time, semantic anchors are generated automatically by an upstream speech-to-text translation model, so deployment does not require oracle translations. In our implementation, Whisper serves as the ASR backbone and mBERT serves as the semantic text encoder, although this

adapter-based anchoring mechanism can in principle be attached to other encoder–decoder ASR architectures. We first formalize this probabilistic framing, then describe semantic-anchor generation and the SAMA-ASR architecture with its training objective.

3.1 Problem Framing and Objective

Given a speech utterance $\mathbf{X}$ and a target transcription $\mathbf{y} = (y_1, \ldots, y_N)$ of length N, a standard autoregressive ASR decoder models

$$p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{X}) = \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(y_i \mid \mathbf{y}_{<i}, \mathbf{X}). \tag{1}$$

The ASR conditional distribution must learn both how acoustic evidence in $\mathbf{X}$ aligns with output tokens and an implicit target-language prior from the observed transcript history $\mathbf{y}_{<i}$. In low-resource settings, the paired transcript set is too small to estimate this distribution reliably:

$$\widehat{p}_{\text{low}}(y_i \mid \mathbf{y}_{<i}, \mathbf{X}) \not\approx p^*(y_i \mid \mathbf{y}_{<i}, \mathbf{X}). \tag{2}$$

Here p^* denotes the true target conditional distribution that the model is intended to approximate, while $\widehat{p}_{\text{low}}$ denotes the distribution learned from limited transcripts. This distributional mismatch reflects a limited-sample estimation error: scarce transcripts provide an incomplete sample of the underlying target-language distribution, so the decoder may place excessive probability mass on tokens that are locally plausible in the observed data but inconsistent with the utterance meaning or the speech signal.

To reduce this mismatch while preserving the ASR target, SAMA-ASR augments the decoder's conditioning information. For an utterance, we define a set of anchors

$$\mathcal{A}(\mathbf{X}, \mathcal{T}) = \{\mathbf{E}_1, \ldots, \mathbf{E}_L, \mathbf{H}\}, \tag{3}$$

where $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_L\}$, L is the number of auxiliary translations, $\mathbf{E}_l$ is the semantic embedding of an auxiliary translation $\mathbf{t}_l$, and $\mathbf{H}$ is the speech embedding produced by the frozen speech encoder. For brevity, we write $\mathcal{A}$ for $\mathcal{A}(\mathbf{X}, \mathcal{T})$ when the utterance and translations are clear from context. In the main setting, $L = 1$ with Mandarin as the auxiliary translation, but the formulation can be extended to $L > 1$ with multiple auxiliary translations; we evaluate this extension in Appendix E.

The adapted decoder therefore estimates

$$p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{y} \mid \mathbf{X}, \mathcal{A}) = \prod_{i=1}^{N} p_{\boldsymbol{\theta},\boldsymbol{\psi}}(y_i \mid \mathbf{y}_{<i}, \mathbf{X}, \mathcal{A}), \tag{4}$$

where $\boldsymbol{\theta}$ denotes the frozen ASR backbone parameters and $\boldsymbol{\psi}$ denotes the trainable SAMA-ASR parameters. Semantic anchors narrow the distribution toward meaning-compatible hypotheses, while the acoustic anchor prevents the model from drifting toward translation-like or acoustically unsupported outputs. Together, these anchors compensate for the limited-sample distributional bias by adding semantic evidence while preserving acoustic grounding for the same ASR target. Because this anchor evidence is available throughout autoregressive decoding, it also mitigates early-decoding fragility after generation begins: the first few continuation tokens need not rely only on a very short and potentially unreliable history $\mathbf{y}_{<i}$. After early mistakes, the anchors continue to provide utterance-level semantic and acoustic references, reducing the decoder's dependence on its own erroneous partial history.

3.2 Semantic Anchor Generation

The left side of Figure 1 shows how SAMA-ASR obtains semantic anchors before ASR decoding. The auxiliary translation $\mathbf{t}_1$ is associated with the input speech; in our Taiwanese Hokkien and Hakka experiments, $\mathbf{t}_1$ is instantiated as Mandarin because paired Mandarin translations are available. Accordingly, SAMA-ASR assumes paired speech–translation data during adaptation and does not directly apply when auxiliary translations are entirely unavailable. During SAMA-ASR training, $\mathbf{t}_1$ is the paired oracle translation from the training split, which provides clean semantic conditioning rather than a pseudo-transcription target. Using oracle translations during adaptation avoids propagating translator errors into the adapter, analogous to teacher forcing with clean conditioning context. Appendix B supports this choice in our ablation setting: the train–test translation-source mismatch is small with 30 hours of paired data, while oracle-translation training becomes more robust than pseudo-translation training when paired data is reduced. For practical inference without human-provided translations, we start from a pretrained Whisper model and fine-tune it independently for speech-to-text translation, using the same training speech utterances as ASR adaptation but the corre-

sponding auxiliary translations as output labels. We optimize this ST generator with a standard cross-entropy translation objective and run it on each test utterance to produce a pseudo translation.

For the $L > 1$ diagnostic analyses, the Mandarin auxiliary translations, instantiated by oracle translations in those analyses, are translated into English, Hindi, Spanish, and French with SeamlessM4T (Communication et al., 2023), forming multilingual auxiliary translations $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_L\}$ with up to five total translations. A frozen mBERT encoder then converts each auxiliary translation into token-level semantic embeddings $\mathbf{E}_{1:L}$ (Devlin et al., 2019). Oracle-translation settings use the same mBERT encoding step but replace the ST-generated pseudo translation with the oracle Mandarin translation.

3.3 SAMA-ASR Architecture

Given semantic embeddings $\mathbf{E}_{1:L}$, the speech input $\mathbf{X}$, and the anchor set $\mathcal{A} = \{\mathbf{E}_1, \ldots, \mathbf{E}_L, \mathbf{H}\}$ defined above, SAMA-ASR keeps the original Whisper encoder and decoder frozen and inserts an additional SAMA-ASR module into each decoder block, as shown in Figure 1. The Whisper encoder maps $\mathbf{X}$ to a speech embedding $\mathbf{H}$, while the semantic-anchor generator described in Section 3.2 provides the semantic embeddings.

Within each decoder block, SAMA-ASR is inserted before decoder self-attention, so the current decoder state can query both translation-derived semantic embeddings and the speech embedding before the frozen decoder updates its autoregressive token-history representation. Let $\mathbf{Q}^{(b)} \in \mathbb{R}^{S \times D}$ denote the decoder hidden states at the SAMA-ASR insertion point of block b, where S is the output-side sequence length and D is the decoder width. Before attention, each anchor representation $\mathbf{A}_k \in \mathcal{A}$ is mapped to the decoder width. SAMA-ASR then uses parallel cross-attention branches so that each anchor can be queried and gated independently. For an anchor $\mathbf{A}_k$, the decoder state obtains

$$\mathbf{C}_k = \mathrm{CrossAttn}_k(\mathbf{Q}^{(b)}, \mathbf{A}_k, \mathbf{A}_k), \tag{5}$$

where $\mathbf{A}_k$ provides both keys and values, and each branch has its own trainable attention projections. The contribution of each branch is controlled by a learnable scalar gate:

$$\widehat{\mathbf{C}}_k = \tanh(\alpha_k)\mathbf{C}_k. \tag{6}$$

The gated contexts are summed with the original decoder state to form the intermediate state

$$\mathbf{P}^{(b)} = \mathbf{Q}^{(b)} + \sum_{\mathbf{A}_k \in \mathcal{A}} \widehat{\mathbf{C}}_k. \tag{7}$$

Finally, SAMA-ASR applies a gated feed-forward adapter branch,

$$\mathbf{Z}^{(b)} = \mathbf{P}^{(b)} + \tanh(\alpha_{\text{ffn}})\,\mathrm{FFN}(\mathbf{P}^{(b)}), \tag{8}$$

and passes $\mathbf{Z}^{(b)}$ to the frozen self-attention sublayer of the original Whisper decoder block. Following Flamingo-style gated cross-attention/dense blocks (Alayrac et al., 2022) and their Whisper adaptation (Rouditchenko et al., 2024), all gate parameters $\{\alpha_k, \alpha_{\text{ffn}}\}$ are initialized to zero, making each inserted SAMA-ASR module initially behave as an identity adapter, $\mathbf{Z}^{(b)} \approx \mathbf{Q}^{(b)}$, before it learns to use semantic and acoustic evidence. The SAMA-ASR adapter design augments rather than replaces Whisper's original frozen encoder–decoder attention: the pretrained acoustic pathway remains intact, while the additional acoustic branch provides a trainable gated memory that can counterbalance noisy semantic anchors before autoregressive self-attention.

Training Objective. Following parameter-efficient low-resource ASR adaptation work (Hou et al., 2021), we optimize only the SAMA-ASR modules and their projection layers to reduce overfitting under scarce target-variety transcripts. SAMA-ASR is trained with a standard cross-entropy loss on the target-language transcription:

$$\mathcal{L}_{\text{ASR}} = -\sum_{i=1}^{N} \log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(y_i \mid \mathbf{y}_{<i}, \mathbf{X}, \mathcal{A}). \tag{9}$$

4 Experimental Setup

4.1 Datasets and Metrics

We evaluate SAMA-ASR on two low-resource Sinitic ASR settings: Taiwanese Hokkien and Hakka. Both settings provide paired dialect speech and Mandarin auxiliary translations, allowing us to study the realistic question of how much paired data is available rather than assuming a separate large resource for the translation component. For both datasets, transcriptions follow the corpus convention of Traditional Chinese character-based notation. We report character error rate (CER) over the corpus-provided character sequence, since Taiwanese Hokkien and Hakka lack standardized word segmentation and explicit word delimiters.

Taiwanese Hokkien (YT-THDC). YT-THDC is a previously introduced Taiwanese Hokkien corpus consisting of YouTube drama audio paired with open-caption Mandarin translations (Yang et al., 2026). These Mandarin translations are used as loosely aligned auxiliary translations rather than deterministic pseudo-transcripts (Chen et al., 2020).

Taiwanese Hakka (HAT). For Hakka, we construct a controlled 30-hour low-resource subset from the Hakka Across Taiwan corpus (Liao et al., 2023), matching the scale of YT-THDC while preserving dialectal variation. Appendix A provides dataset construction details and the split statistics in Table 5 for the controlled 30-hour low-resource setting used for both datasets.

4.2 Baselines and Protocol

Table 1 contains three comparisons: audio-only adaptation, prior text-guided baselines under the same practical automatic-translation setting, and composition with LoRA. Self Attn.+Prompt supplies automatic Mandarin through Whisper's native prompt interface (Peng et al., 2023), TG-ASR serves as a strong semantic-only baseline by encoding the same text as decoder-side anchors (Yang et al., 2026), and SAMA-ASR further adds the acoustic anchor. The prompt baseline uses the same automatic Mandarin translation source as the practical TG-ASR and SAMA-ASR rows; implementation details are in Appendix A. For the practical comparisons, the upstream ST generator only produces auxiliary Mandarin strings shared by Self Attn.+Prompt, TG-ASR, and SAMA-ASR; no oracle test translations are used, and the generator parameters do not participate in ASR decoding.

5 Results and Discussion

5.1 Main Results

Following Sec. 4.2, Table 1 tests whether semantic anchors help beyond audio-only and prior prompt-based baselines, whether acoustic grounding improves over the strong semantic-only TG-ASR baseline, and whether SAMA-ASR composes with LoRA. All rows in the table use per-utterance greedy decoding with beam size 1 and temperature 0. Because practical ASR inference must generate hypotheses without reference-token history, we evaluate all systems in the same free-running regime. This free-running evaluation includes TG-ASR, whose original report used teacher-forced evaluation; the TG-ASR numbers here are therefore intended for controlled comparison with SAMA-ASR rather than direct comparison with the originally reported teacher-forced scores.

Unimodal vs. Multimodal. Standard acoustic adaptation yields only limited gains ($\leq$4.32% relative improvement), suggesting that sparse acoustic supervision alone does not sufficiently reshape decoder-side prediction. The prior prompt-based baseline shows that automatic Mandarin prompts help, while the strong TG-ASR baseline reduces CERs further to 24.38% on YT-THDC and 24.97% on HAT, indicating that adapter-level semantic anchoring is stronger than prompt-based conditioning. Adding the acoustic anchor in SAMA-ASR further improves the practical setting to 23.48% and 22.48%, indicating that acoustic grounding supplies complementary fine-grained evidence rather than merely correcting noisy translations. Appendix D provides a qualitative example.

Oracle Trends. Oracle evaluation clarifies the role of anchor fidelity. On YT-THDC, replacing automatic Mandarin translations with oracle Mandarin translations improves TG-ASR from 24.38% to 20.15%, and adding the acoustic anchor in SAMA-ASR further reduces CER to 16.79%. The larger oracle-condition gain suggests that reliable semantic anchors amplify acoustic grounding: once the semantic stream narrows the hypothesis space, acoustic cues can more effectively resolve remaining surface-form ambiguities. On HAT, oracle TG-ASR (24.28%) still trails practical SAMA-ASR (22.48%), but oracle SAMA-ASR reaches 19.63%, again showing that acoustic grounding remains essential when translation semantics diverge from colloquial dialectal forms.

Compatibility with a Strong PEFT Baseline. LoRA is a strong PEFT baseline, reaching 23.60% CER on YT-THDC and 19.30% on HAT, so standalone SAMA-ASR is not meant to replace it. Instead, SAMA-ASR acts as a compositional semantic-acoustic module: LoRA+SAMA-ASR further reduces CER to 20.64% and 16.82%, corresponding to 12.54% and 12.85% relative improvements over LoRA alone.

5.2 Controlled Acoustic-Anchor Ablation

To verify that the acoustic anchor contributes speech-grounded information rather than merely additional model capacity, we keep the complete SAMA-ASR architecture and trainable parameter count fixed while corrupting only the acoustic-anchor input. **Random Noise** replaces the acoustic

Method	Train	Eval	Aud.	YT-THDC			HAT		
				CER ↓	Rel. Impr. ↑ vs. Base	Rel. Impr. ↑ vs. Audio	CER ↓	Rel. Impr. ↑ vs. Base	Rel. Impr. ↑ vs. Audio
Baseline									
Self Attn.	–	–	No	33.49	–	–	27.78	–	–
Cross Attn.	–	–	Yes	32.89	1.79	–	26.58	4.32	–
Prior Text-Guided Baselines + SAMA-ASR *(Practical Evaluation)*									
Self Attn. + Prompt (Peng et al., 2023)	–	Auto	No	27.24	18.66	17.18	25.45	8.39	4.25
TG-ASR (Yang et al., 2026)	Oracle	Auto	No	24.38	27.20	25.87	24.97	10.12	6.06
SAMA-ASR	Oracle	Auto	Yes	**23.48**	**29.89**	**28.61**	**22.48**	**19.08**	**15.43**
LoRA Baseline + SAMA-ASR									
LoRA	–	–	–	23.60	29.53	28.25	19.30	30.53	27.39
LoRA + SAMA-ASR	Oracle	Auto	Yes	**20.64**	**38.37**	**37.25**	**16.82**	**39.45**	**36.72**
Diagnostic Oracle Evaluation									
TG-ASR	Oracle	Oracle	No	20.15	39.83	38.74	24.28	12.60	8.65
SAMA-ASR	Oracle	Oracle	Yes	**16.79**	**49.87**	**48.95**	**19.63**	**29.34**	**26.15**

Table 1: **Main ASR results on YT-THDC and HAT.** CER values are percentages; lower is better. Relative-improvement columns report CER reduction rates (%) against the context-only Self Attn. baseline ("vs. Base") and the audio-only Cross Attn. baseline ("vs. Audio"); higher is better. "Oracle"/"Auto" denote the auxiliary-translation source, and "Aud." indicates whether an explicit acoustic anchor is used. Bold practical SAMA-ASR results are significant at $p < 0.05$ by Bisani and Ney (2004): standalone SAMA-ASR is compared with all non-LoRA baselines above it, and LoRA+SAMA-ASR is compared with LoRA.

Acoustic-Anchor Condition	YT-THDC	HAT
Self Attn. baseline	33.49	27.78
Cross Attn. baseline	32.89	26.58
Random Noise (matched mean/std)	42.62	30.91
Shuffle Across Utterances	45.49	35.59
SAMA-ASR	**23.48**	**22.48**

Table 2: **Controlled acoustic-anchor ablation.** CERs (%; lower is better). The controlled variants retain the complete SAMA-ASR architecture and trainable parameter count; the assigned corruption is used during both training and evaluation.

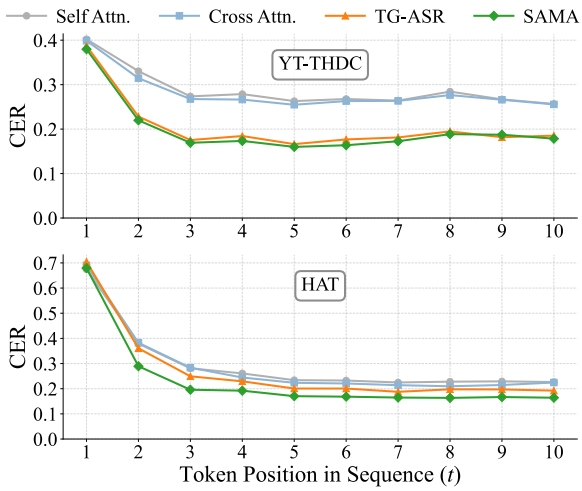

Figure 2: **Position-wise CER analysis of autoregressive cold-start fragility.** Average CER is reported over the first ten target-token positions on the test sets. TG-ASR is the strong semantic-only translation-guided baseline, while SAMA-ASR adds the acoustic-anchor path.

anchor with Gaussian noise matched to the original features' mean and standard deviation, whereas **Shuffle Across Utterances** preserves real acoustic features but assigns them to different utterances within each mini-batch. Each corruption is applied during both training and evaluation.

Table 2 shows that both controls substantially degrade performance relative to SAMA-ASR. Random noise raises CER to 42.62% on YT-THDC and 30.91% on HAT, while utterance shuffling further degrades performance to 45.49% and 35.59%, compared with 23.48% and 22.48% for SAMA-ASR. Both controls also fall below the Self Attn. and Cross Attn. baselines, showing that the acoustic branch alone does not explain the gain. The shuffled condition is particularly diagnostic because it preserves real acoustic representations and their statistics while breaking only utterance alignment. Together, these results isolate the source of SAMA-ASR's gain: correctly aligned speech-derived evidence rather than generic acoustic features or additional capacity.

5.3 Autoregressive Cold-Start Analysis

Figure 2 examines the autoregressive vulnerability motivating SAMA-ASR by measuring CER by output-token position against non-semantic self-attention, audio-only cross-attention, and the strong semantic-only TG-ASR baseline. Both datasets show a clear cold-start pattern: the first few positions are much harder than later ones, confirming that early decisions are fragile when little target-side history is available. SAMA-ASR does not substantially improve the first position, where the ini-

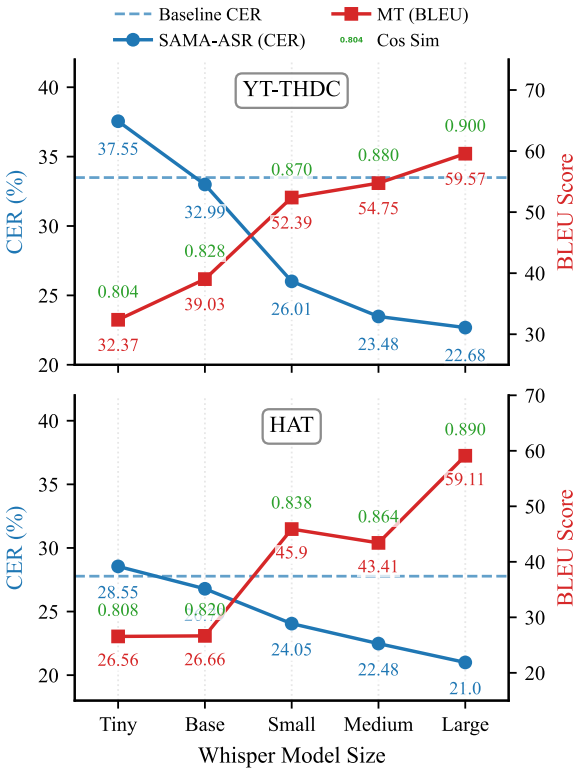

Figure 3: **Evaluation-time translator-capacity analysis.** Values are SAMA-ASR CERs using automatic Mandarin translations from translators of different sizes. Dashed lines denote context-only CER baselines; red values report translator BLEU, and green values report average cosine similarity between mBERT embeddings of automatic and oracle Mandarin translations.

tial surface token is still strongly tied to acoustic onset and utterance alignment. Its advantage emerges from positions 2–3 onward, supporting our framing that semantic anchors help once the decoder begins conditioning on a fragile partial history rather than acting as an oracle for the first token. Compared with TG-ASR, SAMA-ASR shows the clearest additional benefit on HAT, where the acoustic-anchor path yields a stronger early-position separation than semantic-only conditioning. On YT-THDC, the two translation-guided methods are closer, indicating a smaller but still favorable incremental effect from acoustic grounding rather than position-wise dominance.

5.4 Robustness

We evaluate robustness under two practical constraints: varying the evaluation-time translator capacity while fixing ASR adaptation, and reducing the paired-data budget used for both ASR adaptation and translator training. The latter setting is stricter because data scarcity weakens the acoustic adapter, the upstream translator, and the resulting auxiliary translations at the same time.

Method	Data	YT-THDC	HAT
Self Attention	10 min	**61.32**	96.44
Cross Attention	10 min	66.18	78.27
SAMA-ASR	10 min	65.63	**72.67**
Self Attention	1 hr	55.92	68.37
Cross Attention	1 hr	54.97	57.20
SAMA-ASR	1 hr	**53.45**	**54.68**
Self Attention	10 hr	38.10	35.99
Cross Attention	10 hr	39.04	33.61
SAMA-ASR	10 hr	**30.04**	**30.93**

Table 3: **Paired-data budget analysis.** CERs (%) when the same subset trains both ASR adaptation and the translator; SAMA-ASR uses automatic Mandarin translations at evaluation.

Upstream Translation Robustness. Figure 3 shows that SAMA-ASR does not require an oracle-quality translator at inference time. Base and larger translators outperform the context-only baseline on both datasets; even with modest BLEU, mBERT-space similarity to oracle translations remains high. BLEU is not a monotonic proxy for downstream utility: on HAT, the Small translator has higher BLEU than Medium (45.90 vs. 43.41), yet its lower mBERT cosine similarity (0.838 vs. 0.864) coincides with worse CER (24.05 vs. 22.48). The BLEU–CER mismatch suggests that SAMA-ASR is more sensitive to semantic-neighborhood compatibility in the mBERT space consumed by the adapter than to surface n-gram overlap alone. Appendix C visualizes these mBERT neighborhoods.

End-to-End Data-Scarcity Robustness. Table 3 evaluates a stricter scenario where the same paired-data subset trains both the ASR adapter and the translator. With 1 hour and 10 hours, SAMA-ASR gives the best CERs on both datasets. The data-scarcity results show that semantic anchors need not be oracle translations, but they must remain sufficiently meaning-compatible for the decoder-side benefit to outweigh translation noise.

Practical Boundary Conditions. The benefits of semantic anchoring are not uniform under the most resource-constrained conditions. With a Tiny upstream translator, SAMA-ASR can fall below the context-only baseline, indicating that anchors with insufficient semantic fidelity may not provide reliable guidance. Similarly, in the 10-minute paired-data setting, SAMA-ASR performs best on HAT and improves over Cross Attn. on YT-THDC, but trails Self Attn. on YT-THDC (65.63% vs. 61.32% CER). These cases delineate the practical operating boundary of SAMA-ASR: the method

Text Encoder	Train	Eval	Aud.	YT-THDC	HAT
Random Init.	Oracle	Auto	Yes	27.39	27.00
Whisper Emb.	Oracle	Auto	Yes	33.91	28.13
mBERT	Oracle	Auto	Yes	**23.48**	**22.48**

Table 4: **Text-encoder ablation.** CERs (%) in audio-enhanced SAMA-ASR with oracle training anchors and automatic evaluation anchors.

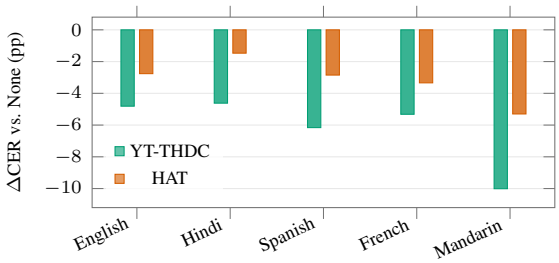

Figure 4: **Auxiliary-language oracle-translation diagnostic.** CER reductions are relative to the context-only baseline; bars below zero are better. Non-Mandarin auxiliary translations are produced by MT from oracle Mandarin translations.

requires either sufficient paired supervision or an upstream translator whose outputs remain meaning-compatible enough to offset uncertainty in low-resource decoding.

5.5 Semantic Prior Ablation

Table 4 isolates the role of the text encoder in the audio-enhanced SAMA-ASR setting. Replacing frozen mBERT with random embeddings substantially degrades performance, and Whisper embeddings perform even worse, showing that the gains do not come merely from adding another conditioning stream. Rather, SAMA-ASR depends on cross-lingual semantic structure in the mBERT semantic encoder, which constrains decoder hypotheses while acoustic evidence determines the final surface form.

5.6 Auxiliary Languages

Figure 4 shows that semantic anchoring is not limited to Mandarin: all MT-derived auxiliary languages outperform the context-only baseline, including typologically distant languages such as English, Spanish, and French. Because these MT-derived auxiliary translations are translated from oracle Mandarin translations, the auxiliary-language experiment is a controlled diagnostic of cross-lingual semantic transfer rather than a deployment-time claim about arbitrary automatic auxiliary translations. Appendix E further compares these trends with URIEL (Littell et al., 2017) typological distances.

6 Conclusion

We present SAMA-ASR, a lightweight decoder-side adapter that treats low-resource ASR as both an acoustic adaptation problem and a target-side generation problem. SAMA-ASR conditions a frozen encoder–decoder ASR backbone on translation-derived semantic anchors while retaining an explicit acoustic anchor. Experiments on Taiwanese Hokkien and Hakka show consistent gains over acoustic and prior prompt-based baselines and a strong semantic-only TG-ASR baseline, as well as further improvements when combined with LoRA. Analyses indicate that reliable semantic anchors help early autoregressive decoding and that compact upstream ST models can provide useful anchors without oracle translations at test time. Overall, acoustic-grounded semantic anchoring offers a practical way to exploit paired speech–translation data when target-language transcripts are scarce.

Limitations

While SAMA-ASR demonstrates promising results, we acknowledge several limitations. First, the practical pipeline requires an upstream ST pass, semantic encoding, and SAMA-ASR decoding, which increases inference cost compared with context-only or audio-only baselines; Appendix F quantifies the resulting latency, model-size, and memory trade-offs. Although the measured real-time factor remains below 1.0 in our offline setting, future work should optimize the latency–accuracy trade-off through model distillation, shared encoders, or anchor caching for more latency-sensitive deployment. Second, SAMA-ASR depends on the quality of the automatic semantic anchors. Our translator-capacity and paired-data-budget analyses show that useful anchors need not be oracle translations, but they must remain sufficiently meaning-compatible; in extremely low-data or severely hallucinated settings, noisy anchors can reduce the benefit or mislead the decoder. Settings without paired auxiliary translations require an alternative source of semantic supervision and are outside the scope of the current framework. This motivates future work on confidence estimation, anchor filtering, and joint training strategies that reduce train–test translation-source mismatch. Finally, our empirical validation is limited to two low-resource Sinitic target varieties, Taiwanese Hokkien and Hakka. The non-Mandarin auxiliary-language and multilingual-anchor analyses are diagnostic because they use MT-derived translations from oracle Mandarin rather than independent deployment-time ST systems. Generalizing acoustic-grounded semantic anchoring to typologically diverse targets, especially languages with agglutinative or polysynthetic morphologies, requires broader multilingual evaluation.

Acknowledgments

This work was supported by the National Science and Technology Council of Taiwan under Grants NSTC 112-2221-E-001-009-MY3 and NSTC 115-2634-F-001-006.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Proc. NeurIPS*.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of ACL*.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mSLAM: Massively multilingual joint pre-training for speech and text. In *Arxiv Preprint arXiv:2202.01374*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proc. NeurIPS*.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proc. ICASSP*.

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Siniscalchi, Pin-Yu Chen, Ensiong Chng, and Chao-Han Huck Yang. 2024. It's never too late: Fusing acoustic Information into large language models for automatic speech recognition. In *Proc. ICLR*.

Pin-Yuan Chen, Chia-Hua Wu, Hung-Shin Lee, Shao-Kang Tsao, Ming-Tat Ko, and Hsin-Min Wang. 2020. Using taigi dramas with mandarin chinese subtitles to improve taigi speech recognition. In *Proc. O-COCOSDA*.

Yao-Fei Cheng, Li-Wei Chen, Hung-Shin Lee, and Hsin-Min Wang. 2025. Exploring the impact of data quantity on ASR in extremely low-resource languages. In *Proc. O-COCOSDA*.

Kwanghee Choi and Hyung-Min Park. 2022. Distilling a pretrained language model to a multilingual ASR model. In *Proc. Interspeech*.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. SeamlessM4T: Massively multilingual & multimodal machine translation. In *Arxiv Preprint arXiv:2308.11596*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech*.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proc. COLING*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proc. ACL*.

Yaroslav Getman, Tamas Grosz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource ASR: A case study on Northern Sámi. In *Proc. Interspeech*.

Michael Hentschel, Yuta Nishikawa, Tatsuya Komatsu, and Yusuke Fujita. 2024. Keep decoding parallel with effective knowledge distillation from language models to end-to-end speech recognisers. In *Proc. ICASSP*.

Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki. 2021. Meta-adapter: Efficient cross-lingual adaptation with meta-learning. In *Proc. ICASSP*.

Chan-Jan Hsu, Yi-Chang Chen, Feng-Ting Liao, Pei-Chen Ho, Yu-Hsiang Wang, Po-Chun Hsu, and Da-shan Shiu. 2025. Let's fuse step by step: A generative fusion decoding algorithm with LLMs for robust and instruction-aware ASR and OCR. In *Findings of ACL*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech*.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Michael L. Seltzer. 2021. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. In *Proc. Interspeech*.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proc. COLING*.

Zhaolin Li and Jan Niehues. 2026. Multimodal in-context learning for ASR of low-resource languages. In *Findings of ACL*.

Yuan-Fu Liao, Shaw-Hwa Hwang, You-Shuo Chen, Han-Chun Lai, Yao-Hsing Chung, Li-Te Shen, Yen-Chun Huang, Chi-Jung Huang, Hsu Wen Han, Li-Wei Chen, Pei-Chung Su, and Chao-Shih Huang. 2023. Taiwanese Hakka across Taiwan corpus and Formosa speech recognition challenge 2023 - Hakka ASR. In *Proc. O-COCOSDA*.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proc. EACL*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proc. ICLR*.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In *Proc. ICASSP*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech*.

An-Ci Peng, Kuan-Tang Huang, Tien-Hong Lo, Hung-Shin Lee, Hsin-Min Wang, and Berlin Chen. 2026. Efficient dialect-aware modeling and conditioning for low-resource Taiwanese Hakka speech processing. In *Proc. LREC*.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. In *Proc. Interspeech*.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proc. ACL*.

Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. In *Proc. Interspeech*.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-efficient and extensible multilingual ASR. In *Proc. Interspeech*.

George Sterpu, Christian Saam, and Naomi Harte. 2018. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proc. ICMI*.

Guangzhi Sun, Xianrui Zheng, Chao Zhang, and Philip C. Woodland. 2023. Can contextual biasing remain effective with whisper and GPT-2? In *Proc. Interspeech*.

Shuta Taniguchi, Tsuneo Kato, Akihiro Tamura, and Keiji Yasuda. 2022. Transformer-based automatic speech recognition with auxiliary input of source language text toward transcribing simultaneous interpretation. In *Proc. Interspeech*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech

recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Cihan Xiao, Henry Li Xinyuan, Jinyi Yang, Dongji Gao, Matthew Wiesner, Kevin Duh, and Sanjeev Khudanpur. 2023. HK-LegiCoST: Leveraging non-verbatim transcripts for speech translation. In *Proc. Interspeech*.

Cheng-Yeh Yang, Chien-Chun Wang, Li-Wei Chen, Hung-Shin Lee, Hsin-Min Wang, and Berlin Chen. 2026. TG-ASR: Translation-guided learning with parallel gated cross attention for low-resource automatic speech recognition. In *Proc. LREC*.

Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025. Enhancing low-resource ASR through versatile TTS: Bridging the data gap. In *Proc. ICASSP*.

Ara Yeroyan and Nikolay Karpov. 2024. Enabling ASR for low-resource languages: A comprehensive dataset creation approach. In *Arxiv Preprint arXiv:2406.01446*.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estève. 2022. Speech Resources in the Tamasheq Language. In *Proc. LREC*.

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-fusion end-to-end contextual biasing. In *Proc. Interspeech*.

A Additional Experimental Details

We confirm that all pre-trained models and datasets employed in our experiments are publicly available. Our usage of these existing artifacts is strictly consistent with their intended research purposes and complies with their respective academic licenses. Furthermore, the public benchmark datasets utilized in this study have been previously curated by their creators to exclude Personally Identifiable Information (PII) and offensive content. Any new artifacts or evaluation scripts generated from this study are carefully reviewed to ensure anonymity and are intended solely for non-commercial, academic research.

Dataset Construction. YT-THDC follows the previously introduced Taiwanese Hokkien split and preprocessing protocol (Yang et al., 2026); its Mandarin translations are used as loosely aligned auxiliary translations rather than verbatim transcripts. For Hakka, we use the Hakka Across Taiwan (HAT) corpus (Liao et al., 2023). Although the full HAT corpus contains approximately 600 hours covering Sixian and Hailu dialects, we construct a controlled low-resource subset by randomly sampling 30 hours of speech and restricting utterance durations to 2–10 seconds. The controlled-subset design keeps the evaluation focused on low-resource conditions while preserving dialectal diversity.

Optimization Details. All ASR systems use the original pretrained Whisper tokenizer, which supports multilingual text including Chinese characters, so no external vocabulary mapping is required. All models are optimized with AdamW (Loshchilov and Hutter, 2019) using a learning rate of 1.0×10^{-4}, a batch size of 32, and a weight decay of 0.01. We train for 32k steps on YT-THDC and 20k steps on HAT to account for differences in utterance count. A linear learning-rate scheduler with a 10% warm-up phase is applied. Audio inputs are truncated or padded to 10 seconds and converted into 80-channel log-mel spectrograms. For SAMA-ASR, both acoustic and cross-lingual text embeddings are projected to dimension $D = 1024$.

Prompt Baseline Decoding. Self Attn.+Prompt uses the same automatic Mandarin translations as the practical TG-ASR and SAMA-ASR rows. For each test utterance, we remove Whisper control tokens from the generated translation and pass the remaining Mandarin string to Whisper's native `prompt` field in `DecodingOptions`. Prompted decoding uses `task=transcribe`, `language=zh`, no timestamps, beam size 1, and temperature 0; because Whisper prompts are specified per decoding option, prompted utterances are decoded one at a time and the prompt is reset for each utterance.

Split	YT-THDC		HAT	
	Duration	**# Utts.**	**Duration**	**# Utts.**
Train	27.51	50,984	27.50	19,120
Test	2.79	4,859	2.50	1,725
Total	30.30	55,843	30.00	20,845

Table 5: **Low-resource dataset statistics.** Durations are in hours.

Model Size and Computational Budget. We implemented all experiments using PyTorch. The complete architecture contains approximately 1.3B parameters in total, of which roughly 404M are trainable parameters during our training phase. Unless otherwise specified, all experiments use a single NVIDIA RTX 3090 (24GB) GPU. The computational budget for training a single experimental run is estimated to be approximately 12 GPU hours.

AI Assistant Usage. During the preparation of this manuscript, we used Prism AI and Gemini to assist with language polishing, grammar correction, and code generation. We explicitly state that all AI-generated text and code were strictly reviewed, verified, and thoroughly tested by the human authors. These tools were not used to make autonomous decisions about methodology, evaluation, or interpretation of results. No generative AI was used to conceive novel scientific ideas, experimental designs, or draw conclusions. The authors assume full responsibility for the accuracy, originality, and integrity of the final manuscript and the implemented codebase.

B Robustness under Limited Training Data and Noisy Translation Resources

Limited Training Data. Table 6 addresses the train–test mismatch discussed in Section 3.3: main experiments train SAMA-ASR with oracle translations but evaluate practical rows with automatic translations. Pseudo-translation training matches inference more closely by using translator-generated translations.

Under the full 30-hour setting, the two strategies are close and mixed: oracle-translation training is slightly better on YT-THDC, while pseudo-translation training is better on HAT. The 30-hour ablation suggests that the train–test translation-

Train Set	Train Translation	YT-THDC	HAT
30 hr	Pseudo	23.93	**21.51**
30 hr	Oracle	**23.48**	22.48
10 hr	Pseudo	50.50	39.53
10 hr	Oracle	**30.04**	**30.93**
1 hr	Pseudo	56.29	65.97
1 hr	Oracle	**53.45**	**54.68**
10 min	Pseudo	67.16	81.19
10 min	Oracle	**65.63**	**72.67**

Table 6: **Training-translation source ablation.** All rows evaluate with automatic Mandarin auxiliary translations; values are CERs (%). Pseudo-translation training uses translator-generated auxiliary translations, while oracle-translation training uses paired Mandarin translations.

System / ST Source	BLEU ↑	CER ↓
Cross Attn. baseline	–	32.89
SAMA-ASR + Tiny	32.37	32.56
SAMA-ASR + Base	39.03	31.90
SAMA-ASR + Small	52.39	26.80
SAMA-ASR + Medium	54.75	24.74
SAMA-ASR + Large	59.57	24.14

Table 7: **Robustness to noisy translation resources on YT-THDC.** Both training and evaluation use pseudo Mandarin translations generated by the indicated ST model; values are BLEU and CER (%).

source mismatch is not large when enough paired data is available to train a usable translator. When paired data is reduced, however, oracle-translation training consistently outperforms pseudo-translation training because the translator-generated translations become much noisier. We therefore use oracle-translation training in the main experiments to avoid propagating translator errors into SAMA-ASR, while keeping evaluation translations automatic for practical deployment.

Noisy Translation Resources. We further consider a separate setting where the full 30-hour training set is retained, but both SAMA-ASR training and evaluation use pseudo translations generated by ST models of different capacities.

Table 7 shows that SAMA-ASR remains effective when semantic anchors are automatically generated throughout both training and evaluation. Even Tiny provides a small improvement over the Cross Attn. baseline, while stronger ST models progressively reduce CER as translation quality improves. Together with the limited-data analysis above, these results distinguish two sources of difficulty: reducing paired data weakens both adaptation and translation, whereas this setting isolates translation noise while retaining the full training set.

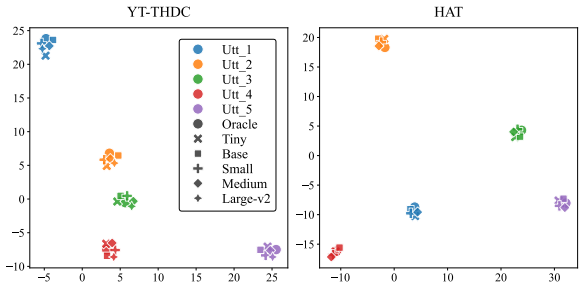

Figure 5: **Semantic-neighborhood visualization of automatic translations.** UMAP shows mBERT embeddings for five sampled test utterances per dataset. Colors denote utterances, marker shapes denote translation sources, and axes are for visualization only.

C Semantic-Neighborhood Visualization

To further examine why automatic semantic anchors can remain useful despite imperfect translation quality, we visualize the mBERT embedding neighborhoods of oracle and automatic Mandarin translations. Figure 5 shows UMAP projections for five sampled test utterances from each dataset. Automatic translations for the same utterance tend to remain close to the corresponding oracle translation, supporting the interpretation from the translator-capacity analysis in Figure 3 that approximate semantic-neighborhood compatibility can guide decoding when the generated anchor remains meaning-compatible.

D Detailed Case Study

To qualitatively inspect the error patterns behind the observed gains, Table 8 presents a representative YT-THDC example in which a short target phrase is acoustically confusable with several semantically incompatible alternatives.

As shown in Table 8, the target sequence 「到銀行」 (*kàu gîn-hâng*) is a difficult case because it can be confused with phonetically similar expressions in Taiwanese Hokkien. The two baselines without semantic anchoring, **Self Attn.** and **Cross Attn.**, therefore make pronunciation-driven substitutions. **Self Attn.** replaces the target phrase with 「教囡仔」 (*kà gín-á*, meaning "teaching children"), while **Cross Attn.** maps the same target slot to 「教議員學」 (*kà gī-uân hˡk*, meaning "teaching councilors to learn"). Both outputs preserve local phonetic similarity, but they are semantically

System	Transcription and English Translation (Gloss)
REF	干焦會當靠到銀行抑是公家機關看診 *(Can only rely on arriving at banks or government agencies for medical consultations.)*
Self Attn.	干焦會當靠教囡仔抑是公家機關看證 *(Can only rely on teaching children or government agencies for verification.)*
Cross Attn.	就按呢咱靠教議員學抑是公家機關看證 *(Just like this, we rely on teaching councilors to learn or government agencies for verification.)*
TG-ASR	干焦會當靠交警方抑是公家機關看證 *(Can only rely on turning over to the police or government agencies for verification.)*
SAMA-ASR	干焦會當靠到銀行抑是公家機關看證 *(Can only rely on arriving at banks or government agencies for verification.)*

Table 8: **YT-THDC error correction case study.** Red highlights mark the target phrase and corresponding substitutions. Baselines produce acoustically or semantically plausible substitutions, whereas SAMA-ASR recovers 「到銀行」(*kàu gîn-hâng*, arriving at banks) by combining semantic guidance with acoustic grounding.

incompatible with the utterance.

By contrast, the text-guided model (**TG-ASR**) has semantic context, but lacks an explicit acoustic anchor. The nearby mention of "government agencies" (公家機關) encourages it to predict the semantically associated but acoustically unsupported phrase "turning over to the police" (「交警方」).

SAMA-ASR avoids both types of errors by using the two anchors for complementary roles. The semantic anchor rejects phonetically plausible but meaning-incompatible hypotheses such as 「教囡仔」 and 「教議員學」, while the acoustic anchor prevents text-side hallucinations such as 「交警方」. Consequently, the model recovers the correct phrase 「到銀行」, illustrating why semantic guidance and acoustic verification are most effective when used together.

Language	Syntactic	Phonological	Inventory
Distances from Taiwanese Hokkien			
Sinitic Family			
Mandarin	**0.58**	**0.53**	0.65
Indo-European Family			
English	0.62	0.64	**0.55**
Hindi	0.70	0.62	0.59
Spanish	0.66	0.62	0.63
French	0.68	0.65	0.57
Distances from Hakka			
Sinitic Family			
Mandarin	**0.52**	**0.41**	0.55
Indo-European Family			
English	0.58	0.52	**0.54**
Hindi	0.68	0.49	0.58
Spanish	0.64	0.49	0.56
French	0.71	0.55	0.55

Table 9: **URIEL linguistic distances.** Cosine distances are measured from Taiwanese Hokkien and Hakka; lower is closer.

E Typological and Multilingual Analyses

To investigate whether auxiliary-language performance is determined solely by linguistic similarity, we compare Figure 4 with the URIEL typological distances in Table 9. Mandarin has the lowest syntactic and phonological distances to both Taiwanese Hokkien and Hakka, consistent with its strong performance as an oracle translation. However, the non-Sinitic results do not follow a simple monotonic relationship with typological distance. For example, French yields stronger gains than English on both datasets despite being syntactically and phonologically farther from the corresponding target variety. The typological comparison supports our interpretation that SAMA-ASR operates in a continuous semantic space and can exploit cross-lingual conceptual information in this oracle-reference-derived diagnostic setting rather than relying only on surface-level similarity.

We further evaluate whether multiple auxiliary translations provide complementary semantic evidence. As shown in Figure 6, the benefit of multilingual anchoring depends strongly on translation composition. The best-first strategy consistently outperforms the worst-first strategy, especially when only one additional language is introduced ($L = 2$), indicating that carefully selected auxiliary translations provide useful complementary semantics beyond Mandarin. As more languages are added, the trend becomes dataset-dependent: some settings saturate earlier, while Hakka remains competitive up to larger multilin-

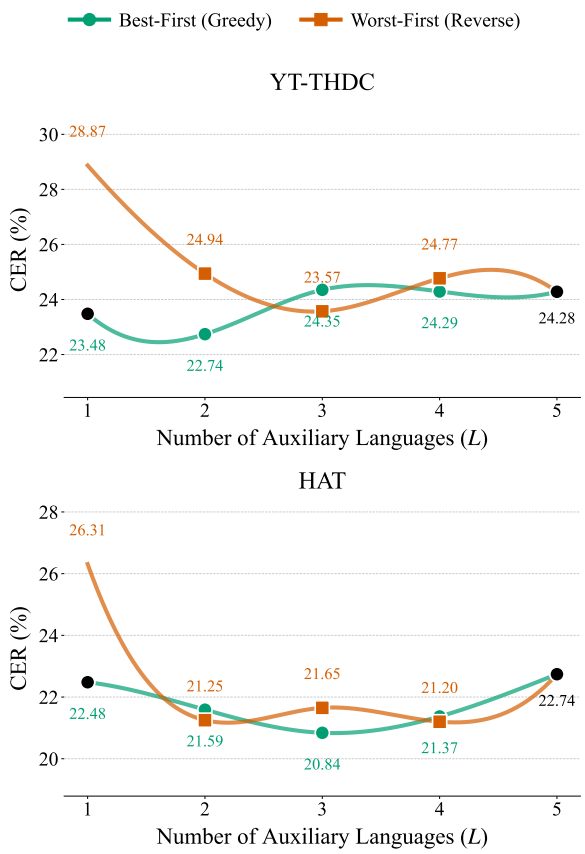

Figure 6: **Multilingual translation-composition diagnostic.** Best-first and worst-first orders add auxiliary translations by oracle single-language performance; L denotes the number of auxiliary translations.

gual sets and degrades only when all five languages are included. These results suggest that multilingual anchoring is not a simple "more languages is better" effect; rather, SAMA-ASR benefits from high-quality complementary translations, whereas redundant or less compatible translations can dilute the attention signal.

F Inference Efficiency and Deployment Cost Analysis

To quantify the accuracy–efficiency trade-off of semantic-anchor generation and acoustic grounding, we report real-time factor (RTF), total model parameters, and peak GPU memory on the full YT-THDC test set. RTF is defined as wall-clock processing time divided by input-audio duration. All efficiency measurements use batch size 1, exclude the first 10 warm-up utterances from timing, and follow the same per-utterance greedy decoding protocol as the main experiments.

Table 10 first isolates the cost of the anchoring modules by assuming that translations are already

Method	Sem.	Aud.	CER ↓	RTF ↓	Params (B)	VRAM (GB)
Self Attn.	No	No	33.49	0.216	0.942	3.66
TG-ASR (Oracle)	Yes	No	20.15	0.307	1.244	4.86
SAMA-ASR (Oracle)	Yes	Yes	16.79	0.391	1.345	5.22

Table 10: **Inference cost with provided translations on YT-THDC.** Oracle translations are available at inference time in this controlled setting, isolating the overhead of semantic and acoustic anchoring from upstream ST generation. "Sem." and "Aud." indicate the explicit semantic- and acoustic-anchor branches.

ST Model	CER ↓	RTF ↓	Params (B)	VRAM (GB)
Tiny	32.56	0.447	1.383	5.35
Base	31.90	0.466	1.418	5.53
Small	26.80	0.489	1.586	6.13
Medium	23.48	0.603	2.109	8.17
Large-v2	22.68	0.674	2.889	11.29

Table 11: **End-to-end deployment cost on YT-THDC.** Each pipeline generates one automatic Mandarin semantic anchor with the indicated upstream ST model, encodes it with mBERT, and performs SAMA-ASR decoding. MT-based multilingual-anchor expansion is not included.

available. Compared with TG-ASR, SAMA-ASR provides a 16.7% relative CER reduction while increasing the total parameter count by 8.1% and peak GPU memory by 7.4%. Its RTF increases from 0.307 to 0.391, but remains well below 1.0.

Table 11 measures the complete practical pipeline with automatic semantic anchors. Increasing ST capacity consistently improves downstream CER at the cost of greater latency, model size, and GPU memory. The Small translator offers a lower-cost operating point at 0.489 RTF, while the Medium translator obtains the 23.48% CER used in the main practical comparison at 0.603 RTF. All evaluated configurations remain faster than real time, allowing the semantic-anchor generator to be selected according to the desired accuracy–efficiency trade-off.