XLINGOTALKER: CROSS-LINGUAL PERSONALIZED SPEECH-DRIVEN 3D FACIAL ANIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual personalized 3D facial animation aims to generate speech-synchronized facial motion in a new language while preserving a speaker's characteristic articulation patterns. Existing methods typically rely on language-matched speech-motion pairs and therefore cannot address cases where target-language motion is unavailable. We formulate this problem as *weakly supervised cross-lingual personalized speech-driven 3D facial animation*, where paired speech-motion data are available only in a source language, together with unpaired target-language speech from the same speaker. The key challenge is to separate speaker-specific motion style from language-dependent acoustic variation. We observe that audio-derived style representations vary substantially across languages, whereas motion-derived speaker representations remain more consistent. Motivated by this finding, we propose *XLingoTalker*, which grounds audio-derived style in a shared audio-motion space. A style-guided multi-head mixture-of-experts decoder maps the grounded representation to personalized facial dynamics, while a style-replacement consistency objective leverages unpaired target-language speech to preserve speaker-specific motion across languages. We further introduce *XLingo-12*, the first multilingual 3D talking-face dataset featuring controlled bilingual recordings from 12 speakers for fixed-identity cross-lingual evaluation. Experiments show that XLingoTalker better preserves personalized facial dynamics while maintaining speech-motion synchronization.

1 INTRODUCTION

Speech-driven 3D facial animation generates facial motion from speech and is a core component of digital humans and virtual avatars. Beyond reproducing speech-motion correspondence, personalized animation aims to capture how a particular individual characteristically moves while speaking, including lip and jaw dynamics, motion amplitude, and expressive behavior (Zhou et al., 2020; Chai et al., 2022; Thambiraja et al., 2023; Fu et al., 2024). Extending this capability across languages is challenging because the generated motion must satisfy two objectives that are related but distinct: articulation must adapt to the phonetic structure of the target language, while the speaker-specific manner of motion should remain unchanged.

This problem is further constrained by the limited availability of paired speech-motion supervision. Target-language speech can be collected relatively easily, whereas acquiring corresponding 3D facial motion requires specialized capture equipment and substantial post-processing. We therefore formulate a new task, termed *weakly supervised cross-lingual personalized speech-driven 3D facial animation*. For each speaker, paired speech-motion data are available only in a source language, together with a limited amount of target-language speech without motion annotations (Fig. 1(a)). At inference time, the model must animate target-language speech while preserving the personalized dynamics learned from source-language motion. This setting differs from multilingual facial animation, which assumes paired speech-motion data in both languages (Fig. 1(b)), and from direct source-to-target transfer, which applies a source-language model to target-language speech and is therefore vulnerable to cross-lingual style shifts (Fig. 1(c)). The central question is thus how to transfer speaker-specific facial dynamics across languages without access to target-language motion.

The main bottleneck lies in the representation of personalization. Existing approaches typically rely on identity embeddings (Fan et al., 2022; Cudeiro et al., 2019), reference-based style representa-

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.

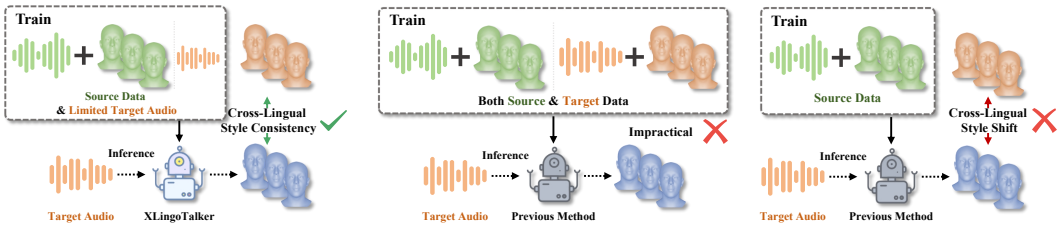


Figure 1: (a) XLingoTalker grounds limited unpaired target-language speech in source-language facial motion, preserving speaker-specific dynamics **without** target-language motion capture. (b) Fully supervised bilingual training requires costly motion capture for every target language, whereas (c) direct source-to-target transfer is prone to personalized style drift.

tions (Sun et al., 2024; Thambiraja et al., 2023), or audio-derived style representations (Kim et al., 2025; Han et al., 2025). Identity embeddings encode categorical speaker information but provide limited access to fine-grained motion behavior. Reference-based representations contain richer motion cues, yet they are often influenced by sample-specific factors such as emotion, speaking rate, and local expression. Audio-derived representations are attractive because they can be extracted from unpaired target-language speech, but they tend to entangle speaker-specific motion characteristics with language-dependent acoustic variation. As a result, the inferred style may change when the language changes, even for the same speaker. This phenomenon is consistent with language-mismatch effects reported in cross-lingual speaker recognition (Qin et al., 2021; Silnova et al., 2022; Nam et al., 2022).

These limitations suggest that cross-lingual personalization should not rely on audio style alone. Facial motion provides a more stable anchor because it directly reflects the speaker-specific dynamics to be preserved, while audio remains necessary at inference time because target-language motion is unavailable. The key is therefore to combine the accessibility of audio with the stability of motion.

Motivated by this insight, we propose *XLingoTalker*, which learns a shared audio-motion style space and uses source-language motion to regularize style representations extracted from both source- and target-language speech. This alignment reduces sensitivity to language-dependent acoustics while retaining speaker-specific motion characteristics. A style-guided multi-head mixture-of-experts decoder then maps the aligned style to personalized facial dynamics without suppressing speech-dependent articulation.

We further introduce a style-replacement consistency constraint to exploit unpaired target-language speech. By enforcing consistent speaker-level motion characteristics under a shared style condition, it preserves personalization across languages while allowing articulation to vary with speech content. Together, these components enable speech-synchronized target-language animation that remains faithful to the speaker's characteristic motion.

For evaluation, we introduce *XLingo-12*, a controlled bilingual 3D talking-face dataset with synchronized speech-motion recordings from the same 12 speakers in both languages. Its fixed-identity design separates language variation from speaker variation and enables direct evaluation of cross-lingual personalized motion preservation. Under our weakly supervised protocol, target-language speech is used during training, while its paired motion is withheld for evaluation.

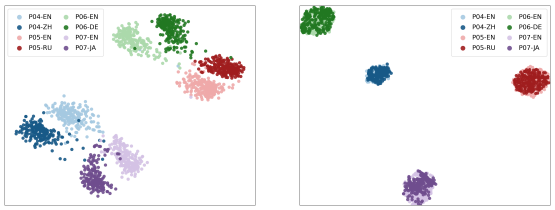
Our main contributions are threefold:

- We introduce the task of weakly supervised cross-lingual personalized speech-driven 3D facial animation, where only source-language speech-motion pairs and unpaired target-language speech are available for each speaker.
- We propose XLingoTalker, a motion-grounded framework that integrates cross-lingual audio-motion style alignment, style-guided multi-head mixture-of-experts decoding, and style-replacement consistency for robust personalization across languages.
- We present XLingo-12, the first multilingual 3D talking-face dataset for cross-lingual personalization. Extensive experiments validate the effectiveness of our approach in motion accuracy, temporal consistency, and speaker-specific style preservation.

2 Related Work

Speech-driven 3D facial animation aims to synthesize realistic and temporally coherent facial motions from speech signals. Early studies investigated the correspondence between acoustic units and visual articulation to model the mapping from speech to facial motion (Zhou et al., 2018; Pham et al., 2017; Karras et al., 2017; Cudeiro et al., 2019; Fanelli et al., 2010; Richard et al., 2021). More recent works (Daněček et al., 2023; Fan et al., 2022; Xing et al., 2023; Peng et al., 2023) adopt a Transformer-based or autoregressive paradigm to model temporal dependencies in facial motion, with CodeTalker (Xing et al., 2023) further introducing a VQ-VAE-based (Van Den Oord et al., 2017) discrete motion codebook to constrain the motion space and improve generation realism. Another line of work (Stan et al., 2023; Chen et al., 2023; Ma et al., 2025; Aneja et al., 2023) explores diffusion-based generative models to capture diverse facial motion distributions and improve generation fidelity. Although these methods improved speech-motion alignment and generation fidelity, they largely overlook speaker-specific facial motion characteristics.

Personalized speech-driven 3D facial animation goes beyond generating plausible speech-synchronized motion by capturing the speaker-specific facial dynamics that characterize an individual's manner of speaking. Existing methods typically model such personalized characteristics using identity labels, reference motions, or audio-derived style representations. Identity-based methods such as VOCA and FaceFormer condition motion generation on subject-specific one-hot codes, which provide only coarse categorical cues and are insufficient for capturing fine-grained speaker-specific motion characteristics (Cudeiro et al., 2019; Fan et al., 2022). To obtain richer personalized cues beyond identity labels, motion-based methods (Thambiraja et al., 2023; Sun et al., 2024; Liu et al., 2024) extract speaker-specific characteristics from reference facial sequences. However, their reliance on reference motion limits practical applicability in speech-only scenarios at inference. A more prevalent line of work achieves personalized facial animation by extracting speaker-specific style cues directly from audio. MemoryTalker (Kim et al., 2025) derives speaking-style representations from input speech to modulate retrieved motion features, while PESTalk (Han et al., 2025) further incorporates voiceprint and emotional cues to capture speaker-dependent expressive patterns. Despite these advances, existing personalized methods are largely developed under monolingual or language-matched settings, leaving the cross-lingual preservation of speaker-specific motion styles insufficiently addressed. Unlike these audio-derived styles, which shift with the input language, we ground speaker style in facial motion and align audio to motion-derived speaker anchors, yielding a style representation that remains consistent under language shift and enables weakly supervised use of unpaired target-language speech.



(a) Audio-derived style embeddings (b) Motion-derived style embeddings

Figure 2: Audio-derived style drifts across languages, whereas motion-derived style remains speaker-consistent. The t-SNE visualization compares (a) PESTalk style embeddings with (b) motion features extracted by a speaker classifier trained only on source-language motion. Colors indicate speaker identities, while lighter and darker shades denote English and non-English utterances, respectively. Motion features demonstrate cross-lingual consistency, motivating their use as cross-lingual style anchors.

3 Empirical Observation and Motivation

Cross-lingual personalization requires a style representation that remains consistent across languages while discriminating between speakers. We first examine whether an existing audio-derived style representation exhibits this property. Specifically, we extract the style embeddings used by PESTalk (Han et al., 2025) from bilingual utterances and visualize them using t-SNE (Van der Maaten & Hinton, 2008). As shown in Fig. 2, embeddings from the same speaker form distinct language-dependent clusters. This separation suggests that the audio-derived style representation retains substantial linguistic variation, causing the inferred style to shift when the input language

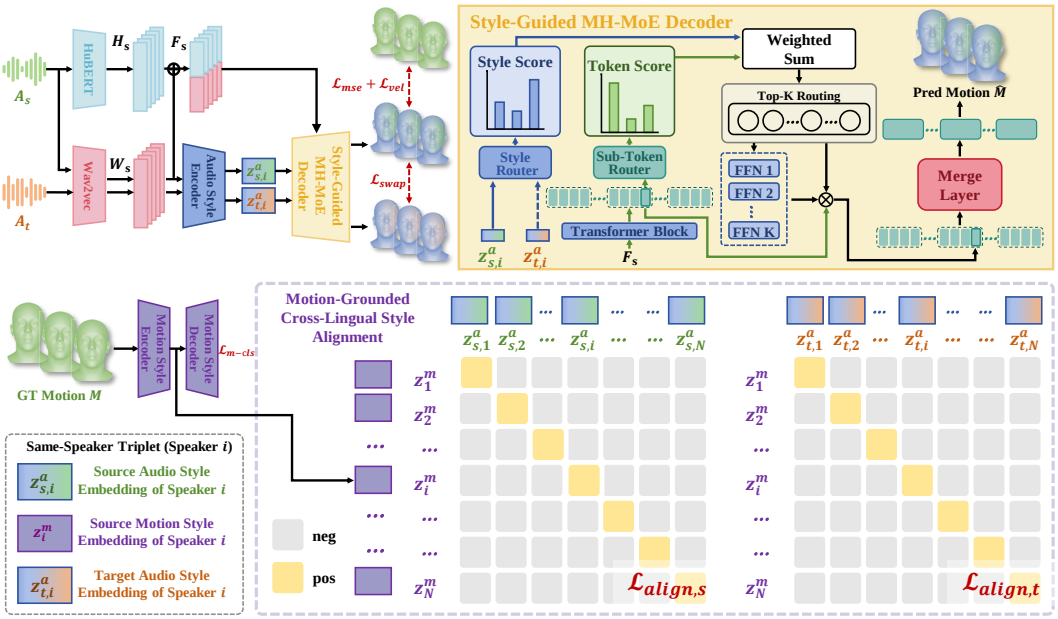


Figure 3: Overview of XLingoTalker. Rather than directly using language-sensitive audio embeddings as motion style codes, XLingoTalker grounds source- and target-language styles in same-speaker source motion through batch-level audio-motion contrastive alignment. The aligned style serves as a global routing prior for the multi-head MoE decoder, while token-adaptive routing preserves content-dependent articulation. This framework enables style extraction and personalized motion generation from the target-language utterances under the weakly-supervised setting.

changes. We then ask *whether facial motion offers a more stable basis for cross-lingual personalization*. To investigate this, we train a lightweight speaker classifier using source-language motion only and extract its intermediate features from motion sequences in both languages. Despite never observing target-language motion during training, the resulting representations cluster primarily by speaker rather than language: motions from the same speaker remain closely grouped across languages, while different speakers are clearly separated.

These results demonstrate that, under language shift, audio-derived style embeddings are unreliable descriptors of speaker style, whereas motion-derived representations exhibit stronger cross-lingual stability while remaining discriminative across speakers. This motivates the core design of XLingoTalker: motion-grounded cross-lingual style learning. Specifically, source-language motion defines speaker-specific anchors in a shared audio-motion style space, and both source- and target-language audio representations are aligned with these anchors. XLingoTalker can thus infer personalized style from target-language speech while mitigating language-induced style drift.

4 Method

4.1 Problem Formulation

Building on the observation that facial motion provides a more stable cross-lingual basis for personalization, we develop XLingoTalker under a *weakly supervised cross-lingual* setting. Motion supervision is confined to a source language, while the target language contributes only a limited amount of speech from the same speakers. Specifically, for each of N speakers, the training set of speaker i is defined as

$$\mathcal{T}^i = \left\{ \left(\mathbf{A}_{\mathrm{s}}^{i,k}, \mathbf{M}_{\mathrm{s}}^{i,k} \right) \right\}_{k=1}^{K_{\mathrm{s}}} \cup \left\{ \mathbf{A}_{\mathrm{t}}^{i,k} \right\}_{k=1}^{K_{\mathrm{t}}}, \tag{1}$$

where $\mathbf{A}_{\mathrm{s}}^{i,k}$ and $\mathbf{A}_{\mathrm{t}}^{i,k}$ denote source- and target-language audio, respectively, and $\mathbf{M}_{\mathrm{s}}^{i,k} \in \mathbb{R}^{T \times V \times 3}$ denotes the 3D facial motion paired with $\mathbf{A}_{\mathrm{s}}^{i,k}$. We assume $K_{\mathrm{t}} \ll K_{\mathrm{s}}$, with no target-language motion available during training. The target-language audio shares speaker identity with the source data but need not be aligned in linguistic transcript. Source-language pairs provide direct motion supervision, whereas target-language speech provides speaker-level cues for cross-lingual style adaptation. Given unseen target-language audio $\mathbf{A}_{\mathrm{t}}$, our objective is to generate facial motion that follows its

linguistic transcript while preserving the personalized dynamics learned from source-language motion:

$$\hat{\mathbf{M}}_{\mathrm{t}} = G_\theta\left(\mathbf{A}_{\mathrm{t}}\right). \tag{2}$$

As illustrated in Fig. 3, XLingoTalker realizes this objective by grounding cross-lingual personalization in source-language facial motion. A shared audio-motion style space aligns source- and target-language audio with motion-derived speaker anchors, yielding motion-relevant style representations robust to language shift. These representations guide expert routing in a multi-head mixture-of-experts decoder, preserving personalized dynamics without suppressing speech-dependent articulation. Style replacement consistency further transfers source-language motion supervision to target-language style representations, enabling personalized animation without paired target-language motion.

4.2 Motion-Grounded Cross-Lingual Style Alignment

Audio-derived style representations may entangle speaker-specific dynamics with language-dependent acoustic and articulatory patterns, making them unstable under language shift. To obtain a cross-lingually robust personalization signal, we construct a shared audio-motion style space anchored by speaker-specific facial dynamics extracted from source-language motion. Given an audio $\mathbf{A}_r$ from language $r \in \{\mathrm{s}, \mathrm{t}\}$, a frozen Wav2Vec 2.0 encoder (Baevski et al., 2020; Conneau et al., 2020) extracts frame-level acoustic features:

$$\mathbf{W}_r = E_{\mathrm{wav2vec}}\left(\mathbf{A}_r\right), \qquad r \in \{\mathrm{s}, \mathrm{t}\}. \tag{3}$$

A shared Transformer-based (Vaswani et al., 2017) style encoder E_{a} aggregates these features into an audio-style embedding:

$$\mathbf{z}_r^{\mathrm{a}} = E_{\mathrm{a}}\left(\mathbf{W}_r\right), \qquad r \in \{\mathrm{s}, \mathrm{t}\}. \tag{4}$$

In parallel, the paired source-language motion $\mathbf{M}_{\mathrm{s}}$ is encoded by a motion-style encoder E_{m}:

$$\mathbf{z}^{\mathrm{m}} = E_{\mathrm{m}}\left(\mathbf{M}_{\mathrm{s}}\right), \tag{5}$$

where E_{m} includes vertex projection and temporal encoding, while $\mathbf{z}_r^{\mathrm{a}}, \mathbf{z}^{\mathrm{m}} \in \mathbb{R}^{B\times C}$ denote audio- and motion-derived style embeddings. We organize this space through speaker-level contrastive learning inspired by CLIP (Radford et al., 2021). Each mini-batch contains B distinct speakers, with one source audio, its paired motion, and one target-language audio sampled per speaker. The target audio shares only speaker identity with the source pair and need not be aligned in linguistic content. Audio and motion embeddings from the same speaker constitute positive pairs, while cross-speaker combinations serve as in-batch negatives. For each language r, the audio-motion similarity matrix is

$$\mathbf{S}_r = \mathrm{Norm}\left(\mathbf{z}_r^{\mathrm{a}}\right)\mathrm{Norm}\left(\mathbf{z}^{\mathrm{m}}\right)^{\top}, \qquad r \in \{\mathrm{s}, \mathrm{t}\}, \tag{6}$$

where $\mathrm{Norm}(\cdot)$ denotes row-wise ℓ_2 normalization. We optimize a bidirectional contrastive loss:

$$\mathcal{L}_{\mathrm{align},r} = \frac{1}{2}\left[\mathrm{CE}\left(\frac{\mathbf{S}_r}{\tau}, \mathbf{y}\right) + \mathrm{CE}\left(\frac{\mathbf{S}_r^{\top}}{\tau}, \mathbf{y}\right)\right], \quad r \in \{\mathrm{s}, \mathrm{t}\}, \tag{7}$$

where τ is a temperature parameter and $\mathbf{y}$ indexes the diagonal speaker-matched pairs. This objective pulls source- and target-language audio toward the same motion-derived speaker anchor while separating different speakers.

To strengthen the discriminability of the motion anchors, we further predict speaker identity from the motion-style embedding:

$$\mathcal{L}_{\text{m-cls}} = \mathrm{CE}\left(Q_{\mathrm{m}}\left(\mathbf{z}^{\mathrm{m}}\right), \mathbf{c}\right), \tag{8}$$

where Q_{m} is a speaker classifier and $\mathbf{c}$ denotes the ground-truth identities. The resulting shared space is organized by speaker-specific facial dynamics rather than language-dependent acoustics, providing motion-grounded style representations for subsequent personalized generation.

4.3 Style-Guided Multi-Head MoE Decoder

The aligned embedding $\mathbf{z}_r^{\mathrm{a}}$ provides a motion-grounded description of speaker style. The remaining challenge is to translate this global style into personalized motion without entangling it with

Table 1: Comparison with existing facial animation datasets. XLingo-12 combines multilingual coverage, explicit speaker identities, and same-speaker bilingual speech-motion data for cross-lingual personalization.

	VOCASET	BIWI	MultiTalk	RAVDESS	HDTF	**XLingo-12**
Duration	0.5h	1.4h	423.2h	2.8h	15.8h	**5.5h**
Num of IDs	12	14	–	24	300+	**12**
Multilingual	✗	✗	✓	✗	✗	✓
Paired data	✗	✗	✗	✗	✗	✓

frame-level articulation. Rather than directly blending style and content features, we cast personalized generation as style-guided expert specialization. Building on mixture-of-experts (Shazeer et al., 2017) and multi-head routing (Wu et al., 2024), our decoder D_{moe} uses speech content to resolve local articulatory variation and motion-grounded style to select speaker-characteristic motion patterns.

For audio $\mathbf{A}_r$, we complement its Wav2Vec 2.0 representation $\mathbf{W}_r$ with phonetic features extracted by HuBERT (Hsu et al., 2021):

$$\mathbf{H}_r = E_{\text{hub}}\left(\mathbf{A}_r\right), \qquad r \in \{\text{s}, \text{t}\}. \tag{9}$$

The acoustic and phonetic streams are fused into a content representation:

$$\mathbf{F}_r = \text{PPE}\left(\left[\phi_{\text{h}}\left(\mathbf{H}_{\text{r}}\right); \phi_{\text{w}}\left(\mathbf{W}_{\text{r}}\right)\right]\right), \tag{10}$$

where ϕ_{h} and ϕ_{w} are linear projections, $[\cdot;\cdot]$ denotes channel-wise concatenation, and $\text{PPE}(\cdot)$ denotes periodic positional encoding. While $\mathbf{F}_r$ encodes temporally varying speech content, $\mathbf{z}^{\text{a}}_u$ supplies the global style condition, where $u \in \{\text{s}, \text{t}\}$ denotes the language from which the style is inferred. The decoder comprises stacked style-guided multi-head MoE layers. Following the multi-head routing paradigm (Wu et al., 2024), each hidden token is partitioned into multiple representation subspaces, allowing different expert groups to model complementary facial dynamics. At layer ℓ, the routing distribution for the h-th sub-token at temporal position n is defined as:

$$\mathbf{p}^{\ell}_{r\leftarrow u,n,h} = \alpha\, \mathbf{p}_{\text{con}}\left(\mathbf{x}^{\ell}_{r\leftarrow u,n,h}\right) + (1-\alpha)\, \mathbf{p}_{\text{sty}}\left(\mathbf{z}^{\text{a}}_u\right), \tag{11}$$

where α balances content-adaptive and style-guided routing. The content branch responds to local phonetic context, whereas the style branch biases the routing toward experts that encode the speaker's characteristic motion patterns. Following sparse MoE computation (Shazeer et al., 2017), only the top-K experts are activated for each sub-token. The generation process is summarized as:

$$\hat{\mathbf{M}}_{r\leftarrow u} = D_{\text{moe}}\left(\mathbf{F}_r, \mathbf{z}^{\text{a}}_u\right), \qquad r, u \in \{\text{s}, \text{t}\}. \tag{12}$$

This factorization enables style to shape the motion pathway without overriding the articulation encoded by speech. It also unifies the three decoding modes required by our framework: $\hat{\mathbf{M}}_{\text{s}\leftarrow\text{s}}$ for supervised source reconstruction, $\hat{\mathbf{M}}_{\text{s}\leftarrow\text{t}}$ for weakly supervised style replacement, and $\hat{\mathbf{M}}_{\text{t}\leftarrow\text{t}}$ for target-language inference. This factorization is what makes weak supervision tractable. The content branch inherits cross-lingual articulation from the multilingual encoders, so the only quantity that must be learned is the personalized manner of motion. Both weakly supervised objectives act on precisely this quantity through the style branch, where $\mathcal{L}_{\text{align,t}}$ anchors target-language audio to the speaker's motion and $\mathcal{L}_{\text{swap}}$ carries source-language motion supervision across languages.

4.4 Cross-Lingual Style Replacement Consistency

The motion-grounded alignment constrains source- and target-language style embeddings at the representation level, but target-language speech still lacks direct supervision from facial motion. We bridge this gap through cross-lingual style replacement consistency. Given the same source-language content representation $\mathbf{F}_{\text{s}}$, we decode it using source- and target-language style embeddings from the same speaker:

$$\hat{\mathbf{M}}_{\text{s}\leftarrow\text{s}} = D_{\text{moe}}\left(\mathbf{F}_{\text{s}}, \mathbf{z}^{\text{a}}_{\text{s}}\right), \quad \hat{\mathbf{M}}_{\text{s}\leftarrow\text{t}} = D_{\text{moe}}\left(\mathbf{F}_{\text{s}}, \mathbf{z}^{\text{a}}_{\text{t}}\right). \tag{13}$$

Because the speech content is fixed and both style embeddings describe the same speaker, replacing $\mathbf{z}^{\text{a}}_{\text{s}}$ with $\mathbf{z}^{\text{a}}_{\text{t}}$ should preserve the generated motion. We therefore impose consistency on both facial geometry and temporal dynamics:

$$\mathcal{L}_{\text{swap}} = \text{MSE}\left(\hat{\mathbf{M}}_{\text{s}\leftarrow\text{s}}, \hat{\mathbf{M}}_{\text{s}\leftarrow\text{t}}\right) + \text{MSE}\left(\Delta\hat{\mathbf{M}}_{\text{s}\leftarrow\text{s}}, \Delta\hat{\mathbf{M}}_{\text{s}\leftarrow\text{t}}\right), \tag{14}$$

Table 2: Quantitative comparison under *English-to-non-English* and *non-English-to-English* settings. Bold indicates the best, and underline indicates the second best. ↓ means lower is better, and ↑ means higher is better.

Method	*English to non-English*							*Non-English to English*						
	LVE↓ ($\times 10^{-5}$)	FVE↓ ($\times 10^{-6}$)	Lip-max↓ ($\times 10^{-5}$)	LDTW↓ ($\times 10^{-5}$)	FDTW↓ ($\times 10^{-6}$)	FID↓ ($\times 10^{-1}$)	SCS↑	LVE↓ ($\times 10^{-5}$)	FVE↓ ($\times 10^{-6}$)	Lip-max↓ ($\times 10^{-5}$)	LDTW↓ ($\times 10^{-5}$)	FDTW↓ ($\times 10^{-6}$)	FID↓ ($\times 10^{-1}$)	SCS↑
FaceFormer	0.3383	0.8071	0.5581	0.2899	0.7367	5.954	0.8159	0.3088	0.7527	0.5363	0.2704	0.7320	4.259	0.8117
CodeTalker	0.3223	0.7559	0.5325	0.2671	0.7153	5.568	0.8655	0.3172	0.7525	0.5401	0.2680	0.7186	4.315	0.7469
SelfTalk	0.3161	0.7679	0.5434	0.2697	0.7390	4.960	0.7820	0.3804	0.7743	0.6943	0.3330	0.7451	4.476	0.7837
Imitator	0.3096	0.6847	0.5390	0.2652	0.6607	4.864	0.8361	0.2517	0.6238	0.4416	0.2144	0.6007	4.792	0.8163
MemoryTalker	0.2833	0.6705	0.4987	0.2469	0.6463	4.797	0.8682	0.2526	0.6959	0.4467	0.2194	0.6670	4.662	0.7720
Ours	**0.2270**	**0.5753**	**0.3941**	**0.1932**	**0.5467**	**3.804**	**0.9100**	**0.2294**	**0.5940**	**0.4026**	**0.1975**	**0.5209**	**3.434**	**0.9073**

where Δ denotes the first-order temporal difference. By holding content constant, this constraint specifically suppresses language-induced variation in the style condition. More importantly, it propagates source-language motion supervision to target-language style inference without requiring target-language motion annotations. For paired source-language data, the reconstruction objectives are:

$$\mathcal{L}_{\text{mse}} = \text{MSE}\left(\hat{\mathbf{M}}_{\text{s}\leftarrow\text{s}}, \mathbf{M}_{\text{s}}\right), \; \mathcal{L}_{\text{vel}} = \text{MSE}\left(\Delta\hat{\mathbf{M}}_{\text{s}\leftarrow\text{s}}, \Delta\mathbf{M}_{\text{s}}\right). \tag{15}$$

The complete training objective combines motion reconstruction, cross-lingual style replacement, audio-motion alignment, and motion-style classification:

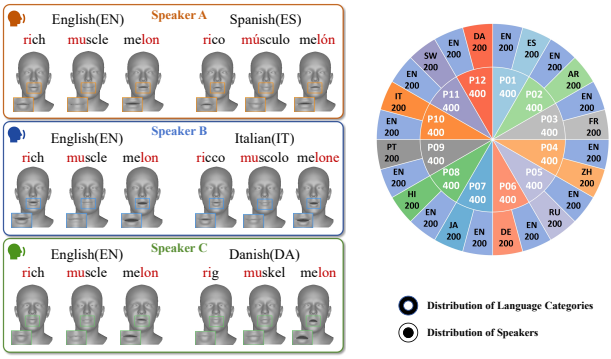
$$\mathcal{L} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{swap}}\mathcal{L}_{\text{swap}} + \lambda_{\text{align}} \sum_{r\in\{\text{s,t}\}} \mathcal{L}_{\text{align},r} + \lambda_{\text{m-cls}}\mathcal{L}_{\text{m-cls}}. \tag{16}$$

5 Experiment

5.1 XLingo-12 Dataset

Cross-lingual personalization requires observations of the same speaker across languages to disentangle speaker dynamics from linguistic variation. As existing datasets lack such controlled coverage as demonstrated in Table 1, we construct **XLingo-12**, a benchmark comprising 5.5 hours of bilingual speech and 3D facial motion from 12 speakers. Each speaker contributes 200 English utterances and 200 utterances in one of 12 non-English languages, yielding 4,800 clips in total. The utterances are paired with facial motion within each language but are not text-aligned across languages.

As shown in Fig. 4, this balanced within-speaker design exposes language-dependent articulation while preserving identity-specific motion characteristics, providing the controlled structure required to study cross-lingual style consistency. 3D facial motions are estimated from publicly accessible recordings using EMOCA (Daněček et al., 2022). We discard identity-related FLAME shape coefficients and retain only shape-normalized motion parameters for vertex reconstruction, reducing the influence of static facial geometry. We evaluate bidirectional transfer under the weakly supervised setting. *English to non-English* uses paired English speech-motion data as source supervision and non-English speech alone for adaptation; *non-English to English* reverses these roles. Target-language motion is withheld from training and used exclusively for evaluation. The two protocols test whether personalized dynamics remain stable regardless of which language provides motion supervision. The two protocols also serve as a built-in control against speaker–language confounding, since



(a) Cross-lingual motion patterns

(b) Data distribution

Figure 4: Composition of XLingo-12. (a) Same-speaker bilingual samples exhibit language-specific articulation while retaining characteristic facial dynamics. (b) XLingo-12 contains 12 speakers, each with 200 English and 200 non-English utterances, yielding balanced bilingual coverage for cross-lingual personalization.

in the non-English-to-English direction the target language is shared by all 12 speakers and thus carries no speaker-discriminative information. Upon acceptance, we will release XLingo-12 to facilitate future research in cross-lingual personalized facial animation.

5.2 Implementation Details

All our experiments are conducted on a single NVIDIA RTX 3090 GPU. For each of the 12 speakers, we use 160/40 paired source-language speech-motion sequences for supervised training/validation, 10 target-language utterances for motion-free adaptation and 190 target-language utterances as the held-out test set. The same protocol is applied to both English-to-non-English and non-English-to-English settings to assess cross-lingual personalized style preservation under limited target-language speech supervision. We adopt the pretrained Wav2Vec 2.0 model (Baevski et al., 2020; Conneau et al., 2020) as the acoustic encoder and employ the pretrained HuBERT model (Hsu et al., 2021) to provide auxiliary linguistic representations. Our MH-MoE decoder consists of 4 layers, each containing 16 experts with 4 routing heads, and the top-K routing parameter is set to $K = 4$. The routing coefficient is set to $\alpha = 0.7$, assigning a larger weight to token-conditioned routing while retaining style-guided guidance. We optimize the model using Adam with a learning rate of 1×10^{-4}. λ_{mse} and λ_{vel} are both set to 1, while λ_{swap}, λ_{align}, and $\lambda_{\text{m-cls}}$ are set to 0.05, 5×10^{-4}, and 5×10^{-4}, respectively. For both experimental settings, the model is trained for 100 epochs.

5.3 Quantitative Evaluation

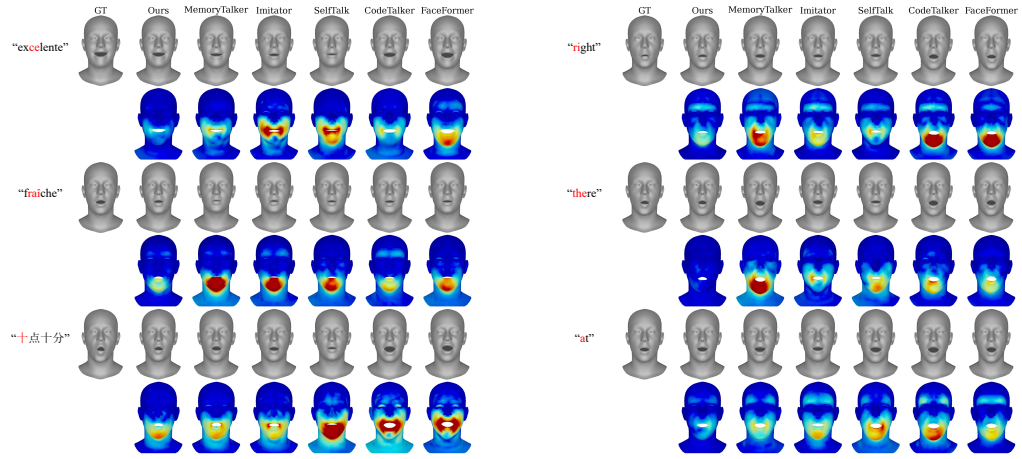
Metrics. We evaluate the generated animation along three complementary dimensions. For **geometric accuracy**, Lip Vertex Error (LVE) and Face Vertex Error (FVE) measure the mean Euclidean error over the lip region and the full facial mesh, respectively, while Lip-max (Richard et al., 2021) captures severe local articulation deviations through the maximum lip-region error. For **temporal fidelity**, Lip Dynamic Time Warping (LDTW) and Face Dynamic Time Warping (FDTW) (Thambiraja et al., 2023) quantify trajectory-level discrepancies in lip and full-face motion. We further assess **motion realism** using the Fréchet Inception Distance (FID) (Heusel et al., 2017), computed on rendered facial images following Kim et al. (2025). For **style preservation**, Style Cosine Similarity (SCS) (Fu et al., 2024) measures the agreement between generated and ground-truth embeddings extracted by a facial-motion classifier. Note that the evaluator shares no parameters with E_m or Q_m and is trained only on the training split, so its judgments are not shaped by the training objectives of XLingoTalker and remain unbiased on the held-out target-language test set.

Comparison with State-of-the-Art Methods. Table 2 reports results in both *English to non-English* and *non-English to English* transfer directions. Our method ranks first on all seven metrics in both settings, demonstrating that its advantage is not tied to a particular source language. For English-to-non-English transfer, XLingoTalker reduces LVE, Lip-max, and LDTW by 19.9%, 21.0%, and 21.8%, respectively, compared with the best competing result for each metric. In the reverse direction, it ranks first on all metrics and reduces FID by 19.4%.

Most notably, XLingoTalker achieves the best SCS values in both settings, exceeding the strongest baselines by 4.18 and 9.10 percentage points, respectively. The stable SCS performance across transfer directions indicates that the learned personalization signal does not rely on language identity, since in the reverse direction the target language is shared by all speakers and offers no shortcut. Moreover, the simultaneous improvements in SCS and lip-related metrics show that stronger style preservation is not obtained at the expense of speech-dependent articulation. These results support the central premise of our method: grounding cross-lingual audio style in facial motion enables personalized dynamics to remain stable under language shift while preserving low error against held-out target-language motion estimates.

5.4 Qualitative Evaluation

Visual Comparisons. Fig. 5 shows that XLingoTalker remains faithful to target-language articulation in both transfer directions. Competing methods frequently over- or under-articulate challenging phonetic segments, producing generic mouth configurations that deviate from the reference. In contrast, our results recover the characteristic lip closure, opening, and shape of each utterance, with errors remaining localized and substantially lower around the mouth. The consistent advantage



(a) English-to-non-English transfer

(b) Non-English-to-English transfer

Figure 5: Qualitative comparisons under two cross-lingual generation settings. Our method produces better articulation and lower mouth-region errors in both transfer directions, demonstrating robust cross-lingual speech-motion alignment.

Table 3: Ablation study of key components under the *English-to-non-English* setting. Bold indicates the best result; ↓ means lower is better.

Metric	w/o $\mathbf{A}_t$	w/o $\mathcal{L}_{\text{align}}$	w/o $\mathcal{L}_{\text{swap}}$	w/o $\mathbf{p}_{\text{sty}}$	**Full**
FVE ↓ ($\times 10^{-6}$)	0.6378	0.6286	0.5955	0.6348	**0.5753**
Lip-max ↓ ($\times 10^{-5}$)	0.4459	0.4323	0.4200	0.4463	**0.3941**
LVE ↓ ($\times 10^{-5}$)	0.2504	0.2447	0.2467	0.2429	**0.2270**

across both language directions suggests that motion-grounded style learning preserves personalization without compromising language-specific articulation.

5.5 Ablation Studies

Table 3 evaluates the key components of XLingoTalker under English-to-non-English transfer. Removing target-language speech $\mathbf{A}_{\text{t}}$ increases FVE, Lip-max, and LVE by 10.9%, 13.1%, and 10.3%, respectively. This result confirms that limited speech-only target data provide a meaningful adaptation signal when grounded in source-language motion, despite the absence of target-language motion. Since this variant never observes target-language audio, its above-baseline SCS confirms that the cross-lingual style is grounded in motion rather than language cues. Meanwhile, the two style-learning objectives operate at complementary levels. Without motion-grounded alignment $\mathcal{L}_{\text{align}}$, all errors increase, with LVE degrading by 7.8%, indicating that unconstrained audio style remains vulnerable to language-dependent variation. Removing $\mathcal{L}_{\text{swap}}$ yields Lip-max degradation of 6.6%. Representation alignment alone therefore does not guarantee consistent decoding behavior, suggesting output-level style replacement is crucial for suppressing localized articulation failures induced by cross-lingual style shifts. Disabling style-guided routing $\mathbf{p}_{\text{sty}}$ also causes broad degradation. This shows that a robust style embedding is insufficient unless it actively shapes generation. By steering expert selection, the proposed router converts motion-grounded style into speaker-specific computational pathways rather than treating it as a passive conditioning vector.

6 Conclusion

We presented *XLingoTalker*, a weakly supervised framework that grounds cross-lingual personalization in speaker-specific facial motions, eliminating the need for paired speech-motion data of target-language. It unifies motion-grounded audio-style alignment, style-guided expert routing, and cross-lingual style consistency to preserve personalized motion across languages. We further introduce XLingo-12 for same-speaker bilingual evaluation. Extensive experiments validate facial motion as a robust cross-lingual bridge and a practical alternative to multilingual motion capture.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research execution through code generation and refinement. We have not used generative AI tools for literature retrieval, research ideation, generating synthetic datasets, or proving mathematical claims. Additionally, we used generative AI tools to draft, revise, and polish parts of the manuscript. All AI-assisted text was carefully reviewed and revised by the authors, and all AI-assisted code was manually inspected, tested, and verified before being used in our experiments. The experimental results and scientific conclusions were independently validated by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in neural information processing systems*, volume 33, pp. 12449–12460, 2020.

Yujin Chai, Tianjia Shao, Yanlin Weng, and Kun Zhou. Personalized audio-driven 3d facial animation via style-content disentanglement. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1803–1820, 2022.

Peng Chen, Xiaobao Wei, Ming Lu, Yitong Zhu, Naiming Yao, Xingyu Xiao, and Hui Chen. Diffusiontalker: Personalization and acceleration for speech-driven 3d face diffuser, 2023.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10101–10111, 2019.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20311–20322, 2022.

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael J Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement, 2023.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18770–18780, 2022.

Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010.

Hui Fu, Zeqing Wang, Ke Gong, Keze Wang, Tianshui Chen, Haojie Li, Haifeng Zeng, and Wenxiong Kang. Mimic: Speaking style disentanglement for speech-driven 3d facial animation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 1770–1777, 2024.

Tianshun Han, Benjia Zhou, Ajian Liu, Yanyan Liang, Du Zhang, Zhen Lei, and Jun Wan. Pestalk: Speech-driven 3d facial animation with personalized emotional styles. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7893–7901, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, volume 30, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (ToG)*, 36(4):1–12, 2017.

Hyung Kyu Kim, Sangmin Lee, and Hak Gu Kim. Memorytalker: Personalized speech-driven 3d facial animation via audio-guided stylization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11241–11251, 2025.

Qingju Liu, Hyeongwoo Kim, and Gaurav Bharaj. Content and style aware audio-driven facial animation, 2024.

Zhiyuan Ma, Xiangyu Zhu, Chen Qian, Shukai Chen, Li Gao, Guojun Qi, Zhaoxiang Zhang, and Zhen Lei. Diffspeaker: Speech-driven 3d facial animation with diffusion transformer. In *2025 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2025.

Kihyun Nam, Youkyum Kim, Jaesung Huh, Hee Soo Heo, Jee-weon Jung, and Joon Son Chung. Disentangled representation learning for multilingual speaker recognition, 2022.

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20687–20697, 2023.

Hai X Pham, Yuting Wang, and Vladimir Pavlovic. End-to-end learning for 3d facial animation from raw waveforms of speech, 2017.

Xiaoyi Qin, Chao Wang, Yong Ma, Min Liu, Shilei Zhang, and Ming Li. Our learned lessons from cross-lingual speaker verification: The crmi-dku system description for the short-duration speaker verification challenge 2021. In *Proc. Interspeech 2021*, pp. 2317–2321, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1173–1182, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017.

Anna Silnova, Themos Stafylakis, Ladislav Mosner, Oldrich Plchot, Johan Rohdin, Pavel Matejka, Lukas Burget, Ondrej Glembek, and Niko Brummer. Analyzing speaker verification embedding extractors and back-ends under language and channel mismatch, 2022.

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11, 2023.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (ToG)*, 43(4):1–9, 2024.

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20621–20631, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in neural information processing systems*, volume 30, 2017.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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.

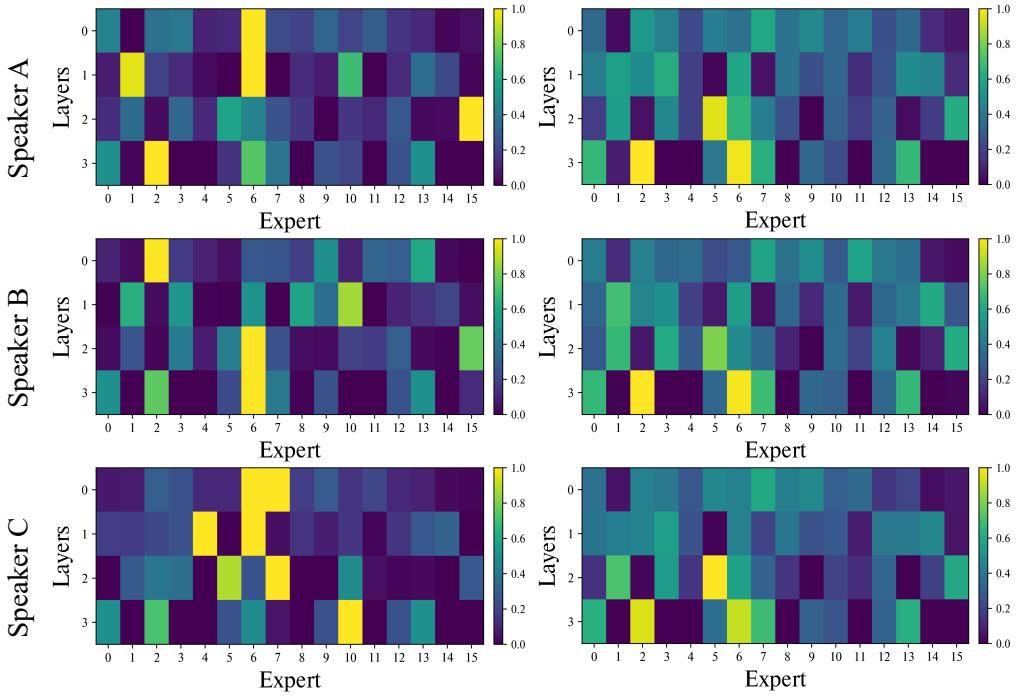
Xun Wu, Shaohan Huang, Wenhui Wang, Shuming Ma, Li Dong, and Furu Wei. Multi-head mixture-of-experts. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 94073–94096. Curran Associates, Inc., 2024. doi: 10.52202/079017-2984.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12780–12790, 2023.

Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (ToG)*, 37(4):1–10, 2018.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020.

A Expert Routing Analysis



(a) Style-guided Routing (b) Token-Only Routing

Figure 6: Motion-grounded style induces speaker-specific expert pathways. Expert activations for 3 speakers uttering the same sentence are visualized across decoder layers. (a) Style-guided routing retains shared experts for common articulation while eliciting distinct speaker-dependent activation patterns. (b) Token-only routing produces largely similar expert preferences across speakers, indicating content-dominated computation with limited personalization.

Fig. 6 compares expert utilization with and without style-guided routing for three speakers uttering the same sentence, which controls for linguistic content. Under token-only routing in Fig. 6(b), the three speakers activate similar experts across layers. This shared pattern indicates that content-dependent routing alone largely maps identical linguistic input to the same computational pathways, despite the additional expert capacity. Introducing the motion-grounded style prior yields more speaker-dependent routing patterns in Fig. 6(a). Some experts remain shared across speakers to model common articulatory structure, while the dominant experts and their layer-wise combinations vary with speaker style. The content branch therefore retains local phonetic computation, while the style branch redirects selected representation subspaces toward speaker-characteristic motion patterns. This visualization provides qualitative evidence that style conditioning actively affects expert selection rather than serving as a passive feature appended to the decoder. Together with the degradation observed after removing style-guided routing in the main paper, these patterns support the role of the proposed router in translating motion-grounded style into personalized generation pathways.

B Target-Language Data Efficiency

We examine how XLingoTalker scales with the amount of motion-free target-language speech. The number of target-language utterances per speaker is varied over $\{0, 2, 5, 10, 20\}$, while the complete source-language speech–motion supervision and held-out test set remain fixed. Using a fixed random seed, we construct nested target-language subsets to ensure a controlled comparison. The zero-target-language-utterance setting removes target-language adaptation entirely, providing a stringent test of whether the motion-grounded style learned from the source language can survive language shift on its own.

Table 4: Target-language data efficiency. XLingoTalker degrades gracefully without target-language adaptation and improves consistently with more motion-free target-language speech, demonstrating robust and data-efficient cross-lingual personalization.

# of Tgt-Lang. Utt.	FVE↓ ($\times 10^{-6}$)	Lip-max↓ ($\times 10^{-5}$)	LVE↓ ($\times 10^{-5}$)	LDTW↓ ($\times 10^{-5}$)	FDTW↓ ($\times 10^{-6}$)	FID↓ ($\times 10^{-4}$)	SCS↑
0	0.6378	0.4459	0.2503	0.2218	0.6074	4.328	0.8724
2	0.6177	0.4291	0.2423	0.2115	0.5889	4.164	0.8861
5	0.5875	0.4039	0.2302	0.1998	0.5623	3.943	0.9025
10	0.5753	0.3941	0.2270	0.1932	0.5467	3.804	0.9100
20	0.5712	0.3914	0.2259	0.1911	0.5412	3.756	0.9134

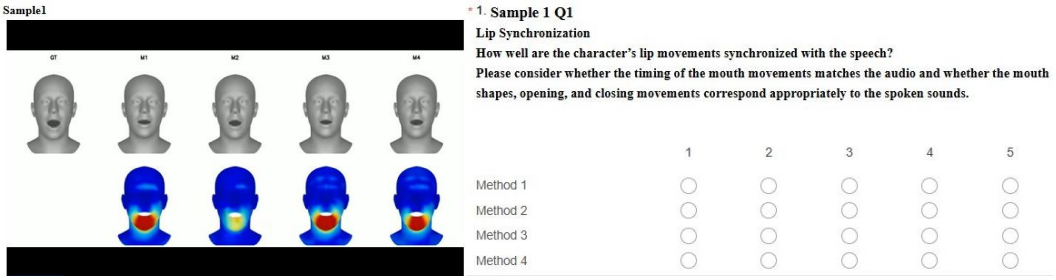


Figure 7: An example of user-study questions. Participants viewed the target-language reference motion (GT) alongside four anonymized generated results (M1–M4) and their error heatmaps. For each result, they provided separate five-point ratings for lip synchronization, motion naturalness, and speaker-specific style preservation. The screenshot shows the lip-synchronization question.

As shown in Table 4, performance improves consistently as more target-language speech becomes available, yet remains robust in the low-data regime. Remarkably, even without target-language adaptation, XLingoTalker retains an SCS of 0.8724 and competitive geometric and temporal accuracy. This graceful degradation supports our central premise: anchoring personalization in source-language facial dynamics yields a style representation that is substantially less sensitive to language-specific acoustics, enabling meaningful zero-shot transfer rather than collapsing under an unseen language condition. Only two target-language utterances already improve every metric, and five utterances recover most of the benefit obtained with ten. Increasing the budget from 10 to 20 yields only marginal gains. The saturation suggests that target-language speech primarily calibrates the motion-grounded style representation to residual language-specific variation, rather than relearning personalization from scratch. XLingoTalker therefore achieves its strongest trade-off with a small speech-only adaptation set, validating both the data efficiency of our weakly supervised formulation and the robustness of motion-grounded cross-lingual style modeling.

C User Study

C.1 Participants

The study involved 20 participants with higher-education backgrounds. All reported good foreign-language proficiency, and most reported proficiency in at least two foreign languages. Before the study, participants received instructions on the evaluation criteria and how to assess the visualized results.

C.2 User Study Protocol

We selected six speakers and randomly sampled three utterances per speaker in each transfer direction, yielding 36 evaluation cases across English-to-non-English and non-English-to-English transfer. For each case, participants viewed 3D-rendered videos of the ground-truth motion and the outputs of four methods, together with the corresponding error-heatmap videos. For each video,

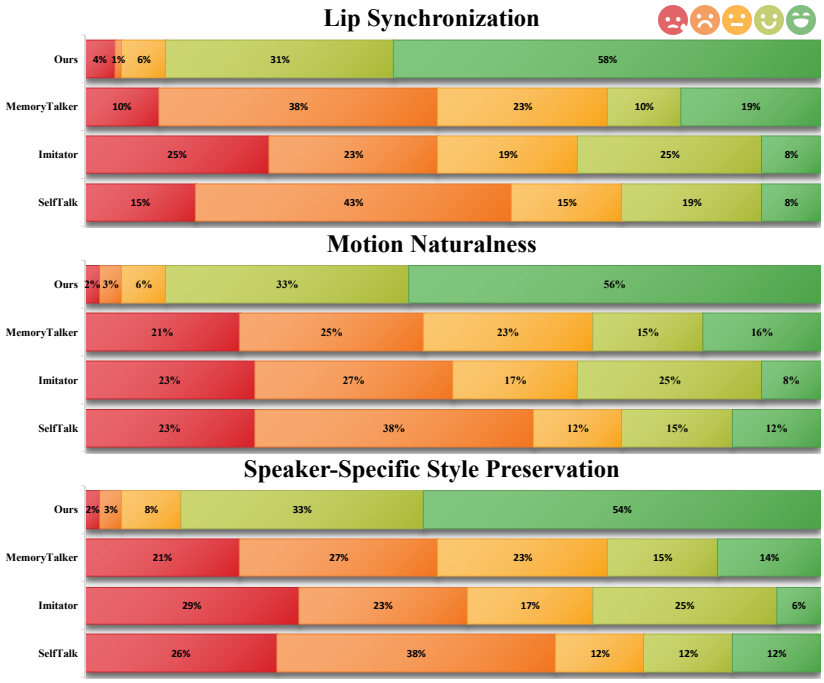


Figure 8: User-study results across six speakers and both transfer directions. XLingoTalker is consistently preferred for lip synchronization, realism, and motion style, demonstrating perceptually robust personalization without compromising animation quality.

participants independently rated every method on a five-point scale (1 as poor and 5 as excellent) for each of three dimensions: lip synchronization, motion naturalness, and speaker-specific style preservation as shown in Fig. 7. Lip synchronization measures the temporal and articulatory agreement between the input speech and the generated lip motion. Motion naturalness measures whether the facial motion appears smooth, coherent, and plausible throughout the utterance. Style preservation measures how faithfully the generated motion retains the target speaker's characteristic motion amplitude, articulation dynamics, and expressive behavior. To keep the evaluation manageable, the perceptual study included our method and three baselines; two other baselines with visibly weaker results were omitted. The presentation order of the four methods was randomized.

C.3 Results and Analysis

Fig. 8 presents the distribution of five-point ratings for the four methods included in the perceptual study. XLingoTalker receives ratings of 4 or 5 in 89% of lip-synchronization assessments and 89% of motion-naturalness assessments, compared with at most 33% for the strongest baseline in either dimension. For speaker-specific style preservation, 87% of its ratings are 4 or 5, versus at most 31% for the baselines. The highest rating of 5 is also the most common response for XLingoTalker across all three dimensions, accounting for 58%, 56%, and 54% of its respective ratings. These distributions suggest that participants perceived the generated lip movements as better aligned with the speech and the overall facial motion as more natural. The favorable style ratings further indicate closer perceived agreement with the displayed target-language reference motion. Taken together, the results provide complementary perceptual evidence for the quantitative comparisons.