# SIGNDUBBER: PERSONALIZED AND SYNCHRONIZED SPEECH SYNTHESIS FROM SIGN LANGUAGE VIDEOS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Sign language translation has largely focused on converting signing videos into text, while the more practically desired task of synthesizing faithful and expressive speech from sign language remains underexplored. To bridge this gap, we introduce a new task: generating personalized speech that matches the signer's appearance (timbre), signing rhythm, and visible emotion from the sign video and its translated text. To tackle this, we present SignDubber, a multimodal conditional speech synthesis framework that jointly models fine-grained video understanding and cross-modal alignment for expressive speech generation. Specifically, we adopt a flow-matching-based audio generation model conditioned on text embeddings from SpeechT5 and video features extracted by an audio-synchronization-aware video encoder, achieving effective cross-modal alignment. Since existing datasets mostly provide neutral signing with text annotations and lack paired expressive speech, we design a creative data synthesis pipeline that leverages pre-trained audio-visual generation models to obtain qualified training pairs. Extensive experiments show that our proposed method generalizes robustly to real-world videos, accurately capturing identity, rhythm, and emotion. Both evaluation metrics and user studies confirm that our approach significantly outperforms multiple representative baselines, demonstrating its practical value.

## 1 INTRODUCTION

Sign languages convey linguistic content not only through hand movements, but also through body posture, head movement, and facial expression, which encode syntax, emphasis, rhythm, and emotion. Existing continuous sign language recognition and translation methods typically map video to gloss sequences or written-language sentences (Camgoz et al., 2018; 2020; Shi et al., 2022), primarily addressing the semantic question of "what was said." Although text conveys the basic content, it does not preserve how that content was expressed. When sign language videos are presented to audiences who do not understand sign language, subtitles alone omit the vocal characteristics associated with the signer's apparent identity, the speed and pauses embedded in the motion, and the emotion conveyed through facial and bodily expression. Written translation therefore cannot fully reproduce the communicative experience of the original sign language video.

We study *sign language video dubbing*: given a silent sign language video and an available translation, the goal is to generate speech matched to the video. An ideal output must satisfy four objectives. First, its spoken content should agree with the translation. Second, it should sound natural and realistic, with a timbre compatible with the signer's apparent identity. Third, its speaking rate, pauses, emphasis, onset, and offset should match the rhythm of the signing motion. Finally, the type and intensity of vocal emotion should agree with the signer's facial and bodily expression. In this task, text specifies "what to say," while video provides evidence about "who is speaking" and "how it is expressed."

This task differs from both conventional text-to-speech (TTS) and general movie dubbing. TTS systems can synthesize intelligible speech from text, but their standard inputs do not include the person's appearance, motion, or expression in the video. Movie-dubbing methods commonly require reference audio and rely on lip motion associated with the target utterance to constrain duration and synchronization (Wu et al., 2023; Cong et al., 2025). A signer's lip motion, however, is not strictly aligned with target-language speech and is therefore unreliable as pronunciation or temporal-



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.

alignment evidence. The model must instead extract rhythm and emotion from upper-body motion and facial expression, combine visual information about the person to produce a compatible timbre, and rely on text to ensure accurate spoken content.

The first challenge is the lack of suitable training data. Existing sign language datasets primarily support recognition and translation and typically provide video, glosses, and translated text, but no target speech that simultaneously matches signer identity, motion duration, and emotional expression. Moreover, mainstream datasets are largely drawn from weather forecasts or controlled laboratory settings, where signing is notably de-emotionalized and information-oriented. To construct usable supervision, we select sign language clips from OpenASL (Shi et al., 2022) and extract full-body pose with DWPose (Yang et al., 2023). Because a single generator struggles to satisfy pose preservation, silent lips, text accuracy, duration alignment, and emotional expression simultaneously, we use a decoupled video/audio pipeline based on LTX-2 (HaCohen et al., 2026). The video branch generates silent emotional sign language video from the first frame, pose, and emotion condition. The audio branch uses the same first frame and pose to generate reference speech whose onset and offset align with the motion and whose timbre is compatible with the signer. IndexTTS2 (Zhou et al., 2025) then enhances the target emotion while preserving the reference rhythm as much as possible. Candidate selection and mild time stretching align the final audio and video durations. This process yields video–text–audio triplets that share signer identity, motion, and emotion.

Building on these data, we propose SignDubber. The model follows the multimodal flow-matching backbone of MMAudio (Cheng et al., 2025), replaces the original text condition with representations from a frozen SpeechT5 encoder (Ao et al., 2022), and structurally removes both the CLIP text and visual encoder streams (Radford et al., 2021). We make these changes because CLIP text encoding emphasizes semantic understanding, whereas SpeechT5 can provide more direct guidance for pronunciation. Generic CLIP visual representations focus on high-level video semantics and do not adequately provide the frame-level temporal constraints required by signing motion. In our model, Synchformer (Iashin et al., 2024) extracts local temporal conditions from the sign language video. Text and audio tokens interact through joint attention, while temporal visual features directly modulate the audio stream. The network predicts a conditional vector field in the audio-VAE latent space, and an Euler solver integrates this field to generate speech. This design assigns content and pronunciation primarily to the text branch, while the video branch provides motion rhythm, signer information, and emotional cues.

Our contributions are as follows:

- We present the first study on a novel task: synthesizing personalized speech from sign language video and translated text. This work bridges a critical gap between current sign-language video translation research and its ultimate real-world application.
- We propose SignDubber, a multimodal conditional speech generation model that captures faithful timbre, rhythmic prosody, and emotional expressiveness from the input video to render expressive speech.
- We develop a scalable data construction pipeline for pairing sign-language videos with faithfully aligned speech audio, which ensures robust generalization capability to real-world cases.

## 2 Related Work

### 2.1 Sign Language Recognition and Translation

In early neural sign language translation (SLT) systems, the task was often decomposed into two cascaded stages: first, continuous sign language recognition (SLR) to convert signs into glosses, and second, translation from glosses into written language. Neural Sign Language Translation formulated video-to-text mapping as a neural machine translation problem (Camgoz et al., 2018), while Sign Language Transformers jointly optimized gloss recognition and translation with a Transformer (Camgoz et al., 2020). For visual modeling, TwoStream-SLR/SLT explicitly combined RGB and body-keypoint streams to represent appearance, hand shape, and body motion (Chen et al., 2022b). Because gloss annotation requires specialist knowledge and is expensive, recent work has increasingly investigated end-to-end gloss-free SLT. GloFE uses common concepts extracted from

translation text as weak intermediate supervision (Lin et al., 2023), while GFSLT-VLP narrows the semantic gap between video and text through vision-language pretraining (Zhou et al., 2023). SignLLM subsequently discretizes video into symbolic sequences more accessible to large language models (Gong et al., 2024); FLa-LLM uses staged training to alleviate the tendency of language model optimization to dominate visual representation learning (Chen et al., 2024b); and SpaMo connects off-the-shelf encoders for spatial configurations and motion dynamics to a large language model (Hwang et al., 2025). These methods still output glosses or written text and are evaluated mainly with recognition or translation metrics such as WER and BLEU. We instead assume that the translated text is already available and focus on vocal identity, motion rhythm, and emotion not explicitly represented in writing. Therefore, SignDubber is a downstream speech-presentation module for a sign language recognition/translation system, rather than a new recognizer or translator.

### 2.2 Text-to-Speech

Modern TTS has expanded from intelligibility alone to zero-shot voice transfer, prosody control, and efficient generation. Among systems based on discrete speech tokens, CosyVoice uses a language model to generate semantic tokens and conditional flow matching to reconstruct speech (Du et al., 2024), while MaskGCT generates semantic and acoustic tokens in two masked-prediction stages without explicit phoneme alignment or duration prediction (Wang et al., 2024). Another line of work applies flow matching directly in a continuous acoustic space. E2 TTS performs speech infilling from character and filler sequences without a separate duration model, grapheme-to-phoneme conversion, or monotonic alignment search (Eskimez et al., 2024); F5-TTS further introduces a text ConvNeXt and inference-time sampling strategies to improve convergence and robustness (Chen et al., 2024a). For controllable synthesis, IndexTTS2 supports zero-shot voice cloning and disentangles timbre from emotion control, achieving state-of-the-art emotional controllability (Zhou et al., 2025), while Qwen3-TTS supports multilingual synthesis, zero-shot voice cloning, and natural-language instruction control (Hu et al., 2026). These systems benefit from large-scale speech corpora and objectives focused on pronunciation clarity, making them strong baselines for content accuracy. Conventional TTS, however, generates speech from text, reference audio, or attribute prompts without directly observing frame-level signing motion, and thus cannot recover pauses, motion emphasis, or evolving emotion directly from the video.

### 2.3 Video Dubbing

Video dubbing uses visual information (mainly lip motion) together with scripts and reference speech to constrain speech. V2C: Visual Voice Cloning explicitly formulated visual voice cloning in which reference audio specifies timbre and reference video specifies emotion (Chen et al., 2022a). VideoDubber predicts the remaining duration during machine translation to mitigate asynchrony caused by cross-lingual text-length variation (Wu et al., 2023). StyleDubber replaces frame-level dubbing with phoneme-level multiscale style learning and uses a phoneme-guided lip aligner to balance pronunciation completeness and synchronization (Cong et al., 2024). Uni-Dubbing aligns audio-visual representations with frozen encoders to support cross-modal, cross-lingual zero-shot transfer (Lei et al., 2024). More recently, EmoDubber combines lip-prosody alignment, pronunciation enhancement, and user-controllable emotional flow matching (Cong et al., 2025), while DiFlow-Dubber first learns discrete speech generation from large-scale TTS data and then adapts to video dubbing through facial-prosody mapping and a cross-modal synchronizer (Nguyen et al., 2026). FunCineForge uses a multimodal large language model to jointly represent permissible speaking intervals, scripts, scene categories, faces, and textual clues, supporting speaker switching and text-driven emotion control in monologue, narration, dialogue, and multi-speaker scenes (Liu et al., 2026). Difficult data acquisition and additional visual synchronization constraints generally make the WER of video dubbing substantially higher than that of TTS.

## 3 Method

We aim to generate faithful and expressive speech audio from sign language videos and their corresponding translated text, where the timbre conforms to the signer's appearance, rhythm aligns with the signing movements, and emotion matches the visible expression. As depicted in Figure 1, we propose SignDubber, a multimodal conditional speech generation model that interprets the signing

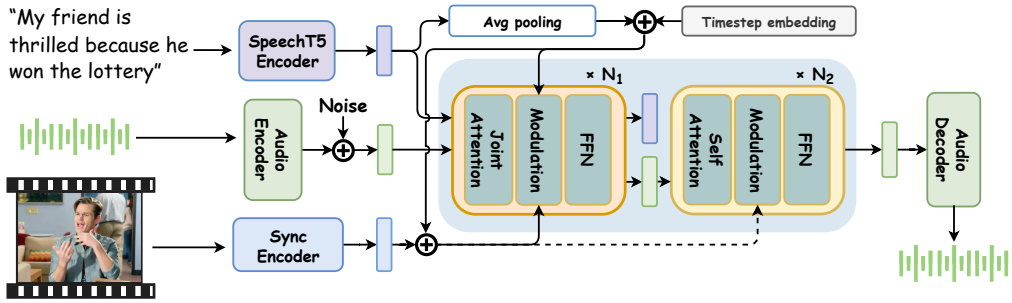


Figure 1: Overview of SignDubber. A diffusion Transformer generates the target audio under multimodal conditioning, where sign-language video encodes speaking style (e.g. timbre, emotion, and rhythm) and narration text encodes linguistic content.

video in a fine-grained manner, while the translated text serves as the speech content. It shares certain similarities with text-to-speech (TTS) tasks, yet it additionally requires the sign language video to be tightly utilized to determine the speaking style. In the following sections, we elaborate on the model designs, training process, and training data construction.

### 3.1 Multimodal Conditional Speech Generation

We formulate speech generation using a Diffusion Transformer (Peebles & Xie, 2023). To facilitate cross-modal interaction between text and audio, we adopt the MM-DiT block design from SD3 (Esser et al., 2024), which employs a dual-branch architecture to enable simultaneous interaction and joint evolution of both modalities across Transformer blocks. Additionally, the global text embedding representation is injected into each Transformer block via adaptive layer normalization (adaLN), alongside the timestep embedding.

**Modality-Specific Representation.** *Video Tokens*: Video frames are resized and center-cropped to $224 \times 224$. For clips shorter than 10 seconds, we pad the video to 10 seconds by repeating the final frame. A frozen Synchformer (Iashin et al., 2024) extracts 768-dimensional features at an effective rate of 24 fps, yielding $F_s \in \mathbb{R}^{240\times768}$. These video features are upsampled via nearest-neighbor interpolation to match the frame rate of the synchronized audio features. *Audio Tokens*: The target audio is represented as a 44.1 kHz monaural waveform, zero-padded on the right to 441,344 samples. A frozen audio VAE (Kingma & Welling, 2014) encodes a 128-bin mel spectrogram into 40-dimensional latents at approximately 43.07 fps, producing $x \in \mathbb{R}^{431\times40}$ for a 10-second clip. The VAE outputs are normalized channel-wise using the training-set statistics before being fed into the flow-matching network. *Text Tokens*: Since the input text prompt serves as speech narration, we adopt SpeechT5 (Ao et al., 2022) as the text encoder over the widely used LLM T5 (Raffel et al., 2020), as its embeddings are better suited for speech synthesis. Input texts are tokenized by the SpeechT5 tokenizer, and the resulting embeddings are truncated or padded to 255 tokens, giving $F_t \in \mathbb{R}^{255\times768}$. Particularly, attention masks are employed to omit the effect of padding tokens during encoding, as well as during the subsequent trainable projection and global pooling.

**Synchronization Module.** The core challenge lies in effectively injecting conditional video features to achieve synchrony between the synthesized speech and the signer's hand movements. As this objective aligns with existing video-to-audio generation tasks, we follow the practice of MMAudio (Cheng et al., 2025) and use Synchformer to extract video features as conditions for audio generation. Since Synchformer is pre-trained for video-audio alignment, its extracted features are inherently compatible with audio features and well-suited for establishing cross-modal temporal correspondences. Specifically, we fuse these video features with the global text-narration representation and use the combined signal to modulate the backbone audio features via an adaptive layer normalization (adaLN) layer on a per-frame basis. In contrast to prior work, we remove the CLIP image features originally fed into the joint attention layer alongside the text and audio tokens. This design choice is motivated by our insight that, unlike in MMAudio, the text modality here serves as speech narration rather than scene-level semantics; thus, incorporating CLIP visual features would

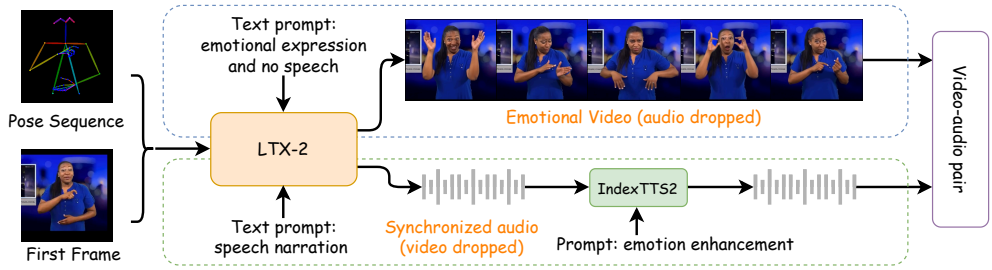


Figure 2: Synthetic data construction pipeline. Video and audio generation are decoupled into two independent passes, differing primarily in the text prompt used.

interfere with speech accuracy and audio-visual synchrony. We validate this phenomenon through our ablation experiments in Sec. 5.3.

Furthermore, our model comprises two types of Transformer blocks: $N_1$ multimodal blocks that allow text and audio interaction, and $N_2$ audio-only blocks, implemented by removing the text-modality streams (where joint attention reverts to self-attention). The audio-only blocks are employed to explicitly enhance audio fidelity and audio-visual synchrony. During inference, the generated audio latent is decoded to a mel spectrogram by the frozen audio VAE and then converted to a waveform using the frozen 44.1 kHz BigVGAN-v2 vocoder (Lee et al., 2023).

### 3.2 Training Objective and Sampling

The training audio is encoded by the frozen VAE as a diagonal Gaussian distribution, from which the data latent $x_1$ is sampled. We standardize each latent channel using training-set statistics. Given Gaussian noise $x_0 \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}(0, 1)$, the linear conditional probability path is

$$x_t = (1 - t)x_0 + tx_1. \tag{1}$$

The network $v_\theta$ predicts the path velocity and is optimized with the flow-matching mean-squared error (Lipman et al., 2023):

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0, x_1, t}\left[\|v_\theta(x_t, t, V, Y) - (x_1 - x_0)\|_2^2\right]. \tag{2}$$

Here, $x_t$ is a point along the vector-field trajectory, $t$ is the timestep, $V$ is the silent sign language video, and $Y$ is the corresponding English translation. The loss is averaged over latent time steps and channels. At inference, integration starts from Gaussian noise and uses 25 Euler steps to reach $t = 1$.

## 4 Data Construction

Our training framework requires triplets consisting of sign-language video, text translation, and speech audio. However, existing publicly available sign-language video datasets are primarily curated for sign-language-to-gloss recognition or text translation tasks, thus providing only {video, text} pairs. A naive solution is to synthesize the missing speech audio via Text-to-Speech (TTS) models. Furthermore, existing sign-language datasets are predominantly composed of neutral expressions and rarely cover the diverse emotions encountered in daily life.

To address these challenges, we propose a novel data synthesis pipeline that leverages the pre-trained text-to-audio-video generation model LTX-2 (HaCohen et al., 2026), conditioned on skeleton sequences. For each paired {video, text} sample, we extract per-frame skeletons using DWPose (Yang et al., 2023). These skeletons serve as the action condition, while the text prompt specifies speech content and emotional attributes, aiming to produce synchronized video and audio with good visual-audio synchrony, consistent appearance and timbre, and diverse emotional expressions. However, our preliminary experiments show that generating video and audio jointly in a single forward pass

suffers from two issues: (i) the text prompt struggles to control the emotion in the audio; (ii) the generated video tends to exhibit pronounced speech-like lip motion, which provides a direct pronunciation shortcut that competes with hand and upper-body dynamics. Based on these observations, we decouple video and audio generation while retaining the same pose condition as a shared reference. The overall pipeline is shown in Figure 2.

**Emotional video without spoken text.** We condition the LTX model on the source first frame, the pose video, and an emotion prompt. The text prompt excludes spoken text and instead reinforces typical emotions through positive descriptions of eyebrow shape, eyelid position, gaze, facial muscles, and mouth state. We retain only the generated video track from this branch.

**Synchronized audio under visual context.** We condition the LTX model on the source first frame and pose video. Its prompt specifies the full spoken content and emphasizes a voice consistent with the person's age, gender presentation, identity, and physical appearance. Additionally, we rely on the ability of the LTX model to align the audio's speaking rate, rhythm, pauses, and emphasis with the visible motion. We filter outputs by objective metrics, removing approximately 10% of the generated samples, and extract their audio tracks to serve as timbre and rhythm references. In addition, we use IndexTTS2 (Zhou et al., 2025) to enhance the synthesized audio pieces with typical emotions, corresponding to their video counterparts.

**Dataset and split.** OpenASL is an open-domain ASL–English translation dataset containing 288 hours of video from more than 200 performers and consisting of signing clips paired with English text from online videos (Shi et al., 2022). We use the OpenASL videos and English translations to construct target speech suitable for our task. We first remove clips shorter than 3 seconds or longer than 10 seconds. For each source video, we generate angry, happy, sad, fearful, and surprised versions. At last, we obtain approximately 0.2 million video–text–audio triplets. The split is determined before filtering and emotion augmentation. Using the original OpenASL sample ID as the grouping key, we assign 90%, 5% and 5% of source videos to the training, validation and testing sets, respectively. All five emotional versions of a source video are written to the same split, preventing the same motion and text from crossing splits only because their emotions differ.

## 5 EXPERIMENTS

**Implementation Details.** Our diffusion Transformer is trained from scratch, and we implement two model scales, the smaller with 0.1B parameters and the larger with 0.8B parameters. The small model is trained on 4 GPUs with a global batch size of 1536 and an initial AdamW learning rate of $8 \times 10^{-5}$. The large model is trained on 8 GPUs with a global batch size of 768 and an initial learning rate of $4 \times 10^{-5}$. Both models are trained for 300,000 iterations with 1,000 steps of linear warmup. Audio is sampled at 44.1 kHz and has a maximum duration of 10 seconds. We use 25 Euler steps at inference time.

**Test Dataset.** We evaluate our method on real sign-language videos that have no overlap with the data used for our training dataset. We use the publicly available EmoSign dataset (Chua et al., 2025) that contains 200 real ASL videos with English text and emotional tendencies, with an average duration of approximately 4.8 seconds, but it does not provide target speech corresponding to the person in each video.

**Evaluation Metrics.** Since the EmoSign dataset lacks ground-truth target speech, we primarily rely on reference-free evaluation. We first use Whisper (Radford et al., 2023) to transcribe the generated speech and compute the Word Error Rate (WER), where lower scores reflect more accurate spoken content. Beyond intelligibility, it is also essential to assess the alignment between the generated speech and the sign movements in the video. As no standard metric is directly applicable to this task, we propose two heuristic metrics: (i) **Audio-video boundary alignment**. We identify the effective start and end timestamps of the speech, denoted as $t_s^a$ and $t_e^a$ and correspondingly the start and end timestamps of the sign movements, denoted as $t_s^v$ and $t_e^v$. The boundary misalignment is then computed as $\text{Bnd-Err} = \frac{|t_s^a - t_s^v| + |t_e^a - t_e^v|}{2}$. (ii) **Cross-modal rhythm correlation.** For sign movement rhythm, we compute the distribution of frame-to-frame keypoint displacement

| Method | WER (%) ↓ | Bnd-Err (s) ↓ | CRC-W1 ↓ | Timbre/Nat.-MOS ↑ | Emotion-MOS ↑ | Rhythm/Sync-MOS ↑ |
|---|---|---|---|---|---|---|
| Qwen3-TTS | **1.59** | 0.76 | 0.17 | $3.59 \pm 0.16$ | $3.38 \pm 0.16$ | $3.09 \pm 0.20$ |
| IndexTTS2 | 1.66 | 0.75 | 0.13 | $3.67 \pm 0.16$ | $3.40 \pm 0.17$ | $3.08 \pm 0.17$ |
| Qwen3.5-Omni-Plus | 1.98 | 0.93 | 0.18 | $3.57 \pm 0.16$ | $3.18 \pm 0.16$ | $2.68 \pm 0.18$ |
| FunCineForge | 16.14 | **0.08** | 0.07 | $3.48 \pm 0.17$ | $3.16 \pm 0.15$ | $3.04 \pm 0.15$ |
| SignDubber-small | 5.90 | **0.08** | **0.06** | $3.93 \pm 0.15$ | $\mathbf{3.82 \pm 0.16}$ | $\mathbf{3.83 \pm 0.14}$ |
| SignDubber-large | 6.71 | **0.08** | **0.06** | $\mathbf{4.00 \pm 0.16}$ | $3.73 \pm 0.15$ | $3.72 \pm 0.16$ |

Table 1: Unified evaluation on EmoSign.

from skeleton keypoints and resample the resulting motion sequence to $N$ temporal bins, denoted as $\mathcal{M} = \{m_i\}_{i=1}^{N}$, where $m_i$ denotes the mean keypoint displacement in the $i$-th temporal bin. Accordingly, for the speech audio, we compute the short-time RMS with a 25 ms window and a 10 ms hop and resample the resulting audio rhythm sequence to the same $N$ temporal bins, yielding $\mathcal{A} = \{a_i\}_{i=1}^{N}$. To measure the correlation between the two rhythms, we treat $M$ and $A$ as probability masses on a normalized time axis and compute their Wasserstein-1 distance:

$$\text{CRC-W1} = \frac{1}{N} \sum_{i=1}^{N} \left| \sum_{j=1}^{i} p_j^v - \sum_{j=1}^{i} p_j^a \right|, \tag{3}$$

where $p_i^v = \frac{m_i}{\sum_{j=1}^{N} m_j}$ and $p_i^a = \frac{a_i}{\sum_{j=1}^{N} a_j}$. A smaller distance indicates stronger rhythmic correspondence.

**Subjective evaluation.** There is no unique objective ground truth for whether a voice suits a person's appearance, whether its emotion matches the video, or whether cross-modal synchronization is perceptually natural. We therefore conduct a subjective evaluation. We select 14 test videos from EmoSign, preferentially choosing those with rich and diverse emotional content as well as sufficient total duration, and recruit 12 participants. For each video, dubbed outputs from different methods are presented anonymously and in randomized method order. Participants score every output on each dimension using a five-point mean opinion score (MOS), where 1 denotes very poor quality and 5 denotes very good quality:

- **Timbre and naturalness (Timbre/Nat.-MOS):** whether the speech is natural and realistic, and whether its timbre is compatible with the signer's apparent identity;
- **Emotion (Emotion-MOS):** whether the type and intensity of the vocal emotion agree with the signer's facial expression and motion;
- **Rhythm and synchronization (Rhythm/Sync-MOS):** whether speaking rate, pauses, and emphasis match the signing rhythm, and whether speech and motion end together.

We report the mean scores from the 12 participants together with 95% confidence intervals. Objective and subjective metrics are presented in the same table (Table 1) to compare content accuracy, temporal alignment, and perceptual quality for the same set of methods.

### 5.1 Comparative Evaluation

No existing model fully matches this task, so we select four strong baselines with complementary capabilities. FunCineForge (Liu et al., 2026) is a dedicated video-dubbing system conditioned on silent video, script, textual clues, and reference speech, with control over speaker attributes, emotion, and complex scene types. Qwen3-TTS (Hu et al., 2026) and IndexTTS2 (Zhou et al., 2025) are pure-TTS baselines representing natural-language descriptive control and disentangled timbre–emotion control, respectively. We use the multimodal large language model Kimi K3 (Team et al., 2026) to read each sign language video and generate appearance, emotion, and speaking-style prompts for these three systems. Because FunCineForge and IndexTTS2 require reference audio, we crop the Qwen3-TTS output to less than one second as a timbre reference, preventing direct inheritance of its full-utterance rhythm. Qwen3.5-Omni-Plus represents a general omni-modal model that directly understands video and generates speech (Team, 2026).

The pure-TTS baselines retain a clear advantage in content accuracy. Qwen3-TTS and IndexTTS2 achieve WERs of 1.59% and 1.66%, Qwen3.5-Omni-Plus reaches 1.98%, and SignDubber-small,

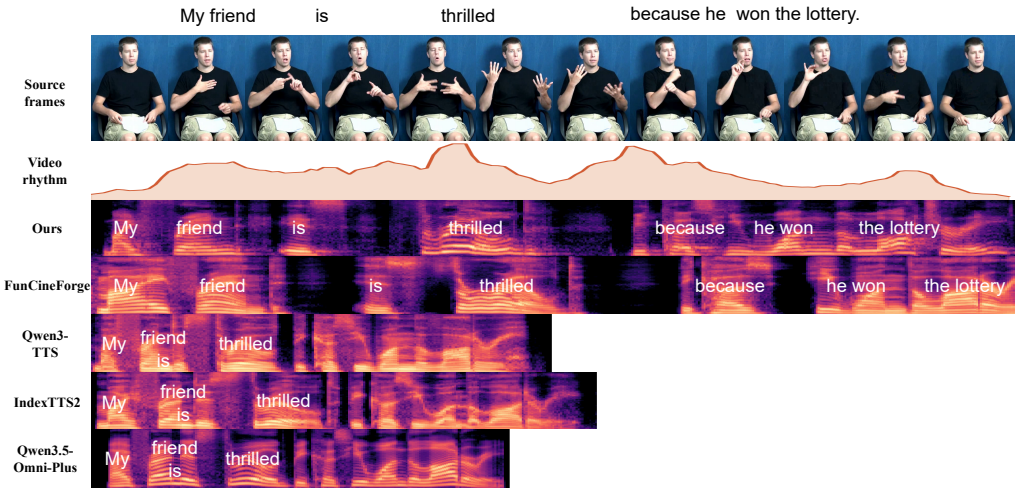


Figure 3: Qualitative synchronization comparison on a real EmoSign sample.

SignDubber-large, and FunCineForge obtain 5.90%, 6.71%, and 16.14%, respectively. Pure TTS can devote model capacity to text-to-speech pronunciation mapping, while general omni-modal models also benefit from large-scale speech-generation training. Dedicated video-dubbing systems must simultaneously satisfy video timing, identity, and emotion constraints, making content accuracy more susceptible to multi-objective trade-offs.

Temporal metrics show the opposite trend. FunCineForge obtains a Bnd-Err of 0.08 and CRC-W1 of 0.07, substantially outperforming both pure-TTS systems and Qwen3.5-Omni-Plus and approaching SignDubber's 0.08 and 0.06. Its explicit video-dubbing formulation therefore controls speaking intervals and global activity distributions well. However, its Rhythm/Sync-MOS is only 3.04 and does not improve in step with the objective synchronization metrics. After listening to the generated results, we believe that although the speech performs well on objective metrics such as temporal interval, endpoint alignment, and energy distribution, it is overly mechanical in prosodic features such as syllable duration, stress, and pauses, resulting in unnatural prosodic perception in some cases. In addition, its WER is overly high. These may explain the discrepancy between objective metrics and subjective ratings.

The subjective results further demonstrate the advantages of our models. SignDubber-small reaches an Emotion-MOS of 3.82 and a Rhythm/Sync-MOS of 3.83, while SignDubber-large achieves the highest Timbre/Nat.-MOS of 4.00. Notably, among the four external baselines, the two pure-TTS systems obtain higher Emotion-MOS values (3.38 and 3.40) than Qwen3.5-Omni-Plus and FunCineForge (3.18 and 3.16). This suggests that explicit emotion control in TTS is more stable than emotion generation in dubbing models, although it remains weaker than SignDubber. Scaling is not uniformly beneficial: the large model slightly improves timbre and naturalness but performs worse than the small model in WER, emotion, and rhythm, suggesting that current performance is constrained more by supervision and task trade-offs than by backbone capacity alone.

### 5.2 Qualitative Analysis

Figure 3 illustrates the temporal correspondence between signing motion and generated speech for a real sample. The top row arranges source-video frames and translated phrase groups along time, the second row shows the video rhythm $\mathcal{M}$, and the remaining rows show spectrograms for each method. Word labels help visualize how utterance content is allocated over the signing sequence.

Qwen3-TTS, IndexTTS2, and Qwen3.5-Omni-Plus concentrate their main acoustic activity in the first half of the video and leave a long silent region afterward, showing that text or video-level prompts alone do not naturally fill the signing duration. FunCineForge and our method both extend speech to the video endpoint and achieve better local synchronization, consistent with their lower Bnd-Err and CRC-W1. However, our method demonstrates better local rhythm alignment: (1) The

| Setting | WER (%) ↓ | Bnd-Err (s) ↓ | CRC-W1 ↓ | Timbre/Nat.-MOS ↑ | Emotion-MOS ↑ | Rhythm/Sync-MOS ↑ |
|---|---|---|---|---|---|---|
| w/ CLIP | 8.08 | 0.09 | 0.09 | $3.70 \pm 0.18$ | $3.64 \pm 0.17$ | $3.65 \pm 0.18$ |
| w/ CLIP w/o Sync | 8.72 | 0.11 | 0.08 | $3.48 \pm 0.18$ | $3.39 \pm 0.17$ | $3.59 \pm 0.15$ |
| w/o Sync | 13.53 | 0.58 | 0.11 | $2.21 \pm 0.20$ | $2.63 \pm 0.16$ | $2.39 \pm 0.18$ |
| Ours | **5.90** | **0.08** | **0.06** | $\mathbf{3.93 \pm 0.15}$ | $\mathbf{3.82 \pm 0.16}$ | $\mathbf{3.83 \pm 0.14}$ |

Table 2: CLIP/Sync ablation on small version. Objective and subjective metrics follow the protocol in Table 1.

two peaks on the rhythm map (corresponding to the words "thrilled" and "because", respectively) are more closely aligned with the high-energy positions in our method; (2) the pause between "friend" and "is" in the speech generated by FunCineForge is excessively long, which is consistent with our analysis in the quantitative evaluation; (3) FunCineForge pronounces "the lottery" too late, by which time the hand movements in the video frames have already ended, and the rhythm map also tends toward low values. For more informative visual comparison, please refer to the video results provided in the supplementary material.

### 5.3 Ablation Studies

For our task, the most critical aspect in designing a multimodal conditional speech generation model is how to utilize the video signal. Here we conduct a comparative analysis of the design choices. Specifically, we discuss two injection pathways: one extracts semantic features from sparse frames via the CLIP image encoder and performs joint attention together with text and audio tokens; the other extracts features via Synchformer and injects them in a frame-by-frame modulation manner. We have a total of four combination choices: **Ours** denotes our choice, which uses only Synchformer; **w/ CLIP** uses both CLIP and Synchformer as two-stream video conditions; **w/ CLIP w/o Sync** indicates using only CLIP; **w/o Sync** uses neither CLIP nor Synchformer, forming a pure text-conditioned model. Each setting is independently trained from scratch with the same conditioning configurations applied during both training and validation. All other experimental parameters—including data split, optimizer, batch size, learning rate, and number of training steps—are kept consistent across settings. This design directly examines whether adding CLIP is beneficial, as well as the contribution of per-timestamp synchronous conditioning with and without CLIP. The ablation experiments are uniformly conducted on the small version.

As shown in Table 2, ours achieves the best WER, CRC-W1, and all three MOS scores. Relative to Ours, adding the CLIP stream increases WER from 5.90% to 8.08% and CRC-W1 from 0.06 to 0.09, while reducing rhythm/synchronization MOS from 3.83 to 3.65; timbre and emotion MOS also show no improvement. This suggests that CLIP's global visual semantics weaken the backbone's use of phonetic information from SpeechT5 and local temporal synchronization cues from Synchformer, which supports our choice. We also provide typical video results in the supplementary material, which support the same conclusion as the quantitative comparison.

The synchronization condition is more critical. Removing Sync from the architecture without CLIP reduces the model to text-only generation: WER rises from 5.90% to 13.53% and CRC-W1 from 0.06 to 0.11, while the three MOS values decrease by 1.72, 1.19, and 1.44. Even with CLIP added, removing Sync worsens WER and all three MOS metrics, showing that video-level semantics cannot replace time-aligned motion conditions. Overall, the ablations show that the combination of SpeechT5 and Synchformer is better suited to this task than adding CLIP.

## 6 Conclusion

We introduce sign language video dubbing and present SignDubber, a multimodal framework that generates personalized, expressive, and synchronized speech from sign language videos and their translated text. SignDubber jointly uses linguistic content and visual cues to model both what should be spoken and how it should be delivered. We also develop a scalable data synthesis pipeline to address the lack of training data for this task. Experiments on real-world sign language videos demonstrate the effectiveness of our approach in matching the signer's appearance, visible emotion, and signing rhythm, advancing sign language translation from written output toward more natural and accessible spoken communication.

### AI USE STATEMENT

In this work, we used LLM-based assistance to identify potentially relevant literature and to draft and polish portions of the manuscript for clarity and readability. We did not use generative AI tools for research ideation, method or experimental design, theoretical modeling, or the formulation or proof of mathematical claims. The authors manually verified suggested references against the original publications, checked all AI-assisted technical descriptions and interpretations against the implementation and experimental records, and reviewed all generated or edited text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

This work uses publicly available research datasets and pretrained models under their respective licenses and terms of use, and our subjective evaluation reports only aggregated ratings from 12 participants. In this work, identity refers to visually conditioned vocal characteristics that are perceptually compatible with the signer's appearance, rather than recovery of the signer's authentic voice or biometric identity. Such appearance-to-voice generation may reflect or amplify stereotypes related to age, gender expression, race, disability, or accent, and generated speech could be misinterpreted as the signer's true voice. We therefore recommend explicit disclosure that the audio is synthetic, preservation of the original signing video and text as authoritative references, signer consent for deployment, and appropriate safeguards against impersonation or misuse. This system is intended as an assistive presentation tool rather than a replacement for Deaf-led translation or interpretation.

### REPRODUCIBILITY STATEMENT

The model architecture, modality representations, flow-matching objective, and sampling procedure are described in Section 3; the synthetic-data construction and grouped train–validation split are described in Section 4; and the training settings, evaluation protocol, and ablations are reported in Section 5. Additional backbone configurations, emotion-control details, baseline prompts and other implementation details are provided in Appendix A.

## REFERENCES

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5723–5738. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.393.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793, 2018. doi: 10.1109/CVPR.2018.00812.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10023–10033, 2020.

Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2C: Visual voice cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21242–21251, 2022a.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024a. doi: 10.48550/arXiv.2410.06885.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems*, volume 35, 2022b. doi: 10.52202/068431-1240.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 7071–7081. ELRA and ICCL, 2024b.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28901–28911, 2025.

Phoebe Chua, Cathy Mengying Fang, Takehiko Ohkawa, Raja Kushalnagar, Suranga Nanayakkara, and Pattie Maes. Emosign: A multimodal dataset for understanding emotions in american sign language. *39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks*, 2025. Available at `https://huggingface.co/datasets/catfang/emosign`.

Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton van den Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. StyleDubber: Towards multi-scale style learning for movie dubbing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6767–6779. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.404.

Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton van den Hengel, Jian Yang, and Qingming Huang. Emodubber: Towards high quality and emotion controllable movie dubbing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15863–15873, 2025.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024. doi: 10.48550/arXiv.2407.05407.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS. *arXiv preprint arXiv:2406.18009*, 2024. doi: 10.48550/arXiv.2406.18009.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. LLMs are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18362–18372, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, and Yaki Bitterman et al. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL `https://arxiv.org/abs/2601.03233`.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-TTS technical report. *arXiv preprint arXiv:2601.15621*, 2026. doi: 10.48550/arXiv.2601.15621.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pp. 3901–3920. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.197.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. In *International Conference on Learning Representations*, 2023.

Songju Lei, Xize Cheng, Mengjiao Lyu, Jianqiao Hu, Jintao Tan, Runlin Liu, Lingyu Xiong, Tao Jin, Xiandong Li, and Zhou Zhao. Uni-dubbing: Zero-shot speech synthesis from visual articulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10082–10099. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.543.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 12904–12916. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.722.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jiaxuan Liu, Yang Xiang, Han Zhao, Xiangang Li, and Zhenhua Ling. FunCineForge: A unified dataset toolkit and model for zero-shot movie dubbing in diverse cinematic scenes. *arXiv preprint arXiv:2601.14777*, 2026. doi: 10.48550/arXiv.2601.14777.

Ngoc-Son Nguyen, Thanh V. T. Tran, Jeongsoo Choi, Hieu-Nghia Huynh-Nguyen, Truong-Son Hy, and Van Nguyen. DiFlowDubber: Discrete flow matching for automated video dubbing via cross-modal alignment and synchronization. *arXiv preprint arXiv:2603.14267*, 2026. doi: 10.48550/arXiv.2603.14267.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6365–6379. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.427.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, and M. C. et al. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024. doi: 10.48550/arXiv.2409.00750.

Yihan Wu, Junliang Guo, Xu Tan, Chen Zhang, Bohan Li, Ruihua Song, Lei He, Sheng Zhao, Arul Menezes, and Jiang Bian. Videodubber: Machine translation with speech-aware length control for video dubbing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13772–13779, 2023. doi: 10.1609/aaai.v37i11.26613.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4210–4220, 2023.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20871–20881, 2023.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. IndexTTS2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*, 2025. doi: 10.48550/arXiv.2506.21619.

# A Appendix

## A.1 Comparison with the Data Synthesis Pipeline

We compare SignDubber directly with the LTX-2 data synthesis pipeline to determine whether the trained model merely inherits the capabilities of its synthetic-data generator or learns a stronger mapping from sign language video to speech. All methods are evaluated on the same unfiltered testing set of OpenASL. For LTX-2, we use the synchronized audio produced directly by its audio branch, before dataset-level quality filtering and IndexTTS2 emotion enhancement. Since these outputs do not undergo the subsequent emotion-enhancement stage, we restrict the comparison to the objective metrics used in the main evaluation: WER, Bnd-Err, and CRC-W1.

We additionally measure whether the generated voice is compatible with the apparent identity of the signer. For each method, we randomly sample 100 outputs from its complete result set and manually exclude outputs whose voices are severely inconsistent with the person shown in the corresponding video. We define the identity success rate (Suc. Rate) as the number of retained outputs divided by 100.

Table 3: Comparison with raw LTX-2 outputs on the unfiltered synthetic dataset.

| Method | WER (%) ↓ | Bnd-Err (s) ↓ | CRC-W1 ↓ | Suc. Rate ↑ |
|---|---|---|---|---|
| LTX-2 | 10.86 | 0.35 | 0.10 | 0.87 |
| SignDubber-small | **6.33** | **0.09** | **0.06** | 0.98 |
| SignDubber-large | 7.17 | **0.09** | **0.06** | **1.00** |

As shown in Table 3, both SignDubber variants consistently outperform the raw LTX-2 outputs in content accuracy, temporal alignment, and identity compatibility. This comparison is meaningful because LTX-2 already provides non-trivial identity consistency and local synchronization through its visual and pose conditions. The gains therefore indicate that SignDubber learns capabilities beyond those of the data construction pipeline itself.

Two factors may contribute to these improvements. First, the synthesized data are filtered according to the objective content and temporal-alignment criteria reflected by WER, Bnd-Err, and CRC-W1, removing approximately 10% of the generated samples. SignDubber is therefore trained only on the resulting high-quality subset. This curated supervision distills the reliable behavior of the synthesis pipeline while excluding many of its individual failures. Second, learning across many text-audio-video samples allows SignDubber to aggregate stable cross-modal correspondences instead of relying on a single stochastic generation. Visual appearance and motion provide recurring supervision for vocal identity and timing, while paired facial expressions and vocal emotions support learning the relation between visible and acoustic affect. Although emotion is not evaluated in this comparison, these shared cross-modal correspondences may contribute to a more stable generation process than that of the raw pipeline.

## A.2 Data Synthesis Details

**Facial-expression prompts.** In the emotional-video branch described in Section 4, we use one of the following emotion-specific prompt clauses to control the signer's facial expression. These clauses describe visible facial configurations rather than speech articulation, which helps the video branch generate expressive signing while suppressing speech-like lip motion.

**Angry.**

> The facial expression is intensely angry and unmistakable, with eyebrows pulled down and inward into a deep scowl, narrowed glaring eyes, flared nostrils, tight cheek muscles, and a rigid clenched jaw. The lips stay firmly compressed in a severe frown as a held facial expression, not as speech articulation.

**Happy.**

> The facial expression is happy, showing a wide, radiant smile that reveals teeth, with natural crinkles forming around the outer eye corners.

**Sad.**

> The facial expression is sad, characterized by heavy, drooping eyelids, lowered eyebrows, and a small frown. The mouth corners point downward.

**Fearful.**

> The facial expression is strongly fearful and unmistakable, with raised inner eyebrows drawn together, very wide eyes showing clear whites above and below the irises, tense cheeks, and a stiff apprehensive mouth. The lips are held slightly open or pulled tight in silent alarm, not moving as if speaking.

**Surprised.**

> The facial expression is surprised, featuring raised eyebrows, widened eyes, and a softly opened mouth with a startled, breath-catching expression.

**IndexTTS2 emotion control and duration matching.** We use the explicit emotion vector of IndexTTS2 (Zhou et al., 2025) to control the emotion of the synthesized speech. For each target emotion, the corresponding coordinate is set to its maximum supported value and every other coordinate is set to zero, so that the audio emotion matches the emotion assigned to the generated video. To retain the speaking rhythm of the reference audio as closely as possible, we generate three consecutive candidates for every target emotion. We then select the candidate whose original, unstretched duration is closest to that of the corresponding emotional video and apply only mild audio time stretching to make the two durations identical.

### A.3 Model Configurations

Table 4 reports the detailed configurations of the two SignDubber backbones. The audio–text joint blocks jointly update the audio and text streams through joint attention, whereas the subsequent fusion blocks retain only the audio stream and fuse it with the frame-aligned visual condition.

Table 4: Architectural configurations of SignDubber-small and SignDubber-large.

| Model | Hidden dim. | Attention heads | Joint blocks | Fusion blocks |
|---|---|---|---|---|
| SignDubber-small | 448 | 7 | 4 | 8 |
| SignDubber-large | 896 | 14 | 7 | 14 |

### A.4 Kimi Prompts Used for the Baselines

As described in Section 5.1, we use Kimi K3 (Team et al., 2026) to infer vocal attributes and emotion from each input video. The Kimi instruction depends on the interface of the downstream baseline. FunCineForge and Qwen3-TTS accept a detailed natural-language style instruction, whereas IndexTTS2 uses a shorter description focused on emotion.

#### Prompt for FunCineForge and Qwen3-TTS

The following instruction is supplied to Kimi when generating the style prompt for FunCineForge (Liu et al., 2026) and Qwen3-TTS (Hu et al., 2026):

> **Role:**
> You are a professional Vocal Profiler. Your task is to translate visual cues from a video into a single, highly descriptive sentence that serves as a style instruction for a TTS model.
> **Task:**

> Analyze the main individual in the video and synthesize their vocal texture, emotional state, and speech rhythm into one seamless English sentence.
>
> **Dimensions for Analysis:**
>
> **Vocal Texture:** Pitch and timbre (e.g., husky, resonant, metallic, airy) based on age and appearance.
>
> **Affective State:** Emotional undertone (e.g., weary authority, bubbling enthusiasm, suppressed tension).
>
> **Prosody:** Rhythm and cadence (e.g., weighted pauses, rapid fluid delivery, erratic stops).
>
> **Strict Constraints:**
>
> **One Sentence Only:** The entire description MUST be a single, grammatically correct sentence.
>
> **Length Limit:** Under 80 words.
>
> **No Metadata:** Do NOT use "In the video," "The person," or "The clip shows."
>
> **Direct Start:** Begin immediately with the subject description (e.g., "A middle-aged man...").
>
> **No Formatting:** Do not use bullet points or bold text in the output.
>
> **Example Outputs (For Context Learning):**
>
> **Example 1 (Authoritative & Fatigued):**
>
> A middle-aged man with a deep, resonant chest voice and grainy texture, showing weary authority through a heavy brow and deliberate blinks, speaks with a slow, weighted cadence and long, thoughtful pauses between phrases.
>
> **Example 2 (Enthusiastic & Bright):**
>
> A young woman possessing a bright, bell-like soprano voice with a clean texture, her face radiating bubbling enthusiasm through animated expressions, delivers speech in a rapid, fluid manner with short, breathy pauses and melodic inflections.
>
> **Example 3 (Gentle & Nostalgic):**
>
> An elderly woman with a soft, thin, and slightly shaky timbre, her serene nostalgia evident in a gentle smile and relaxed features, speaks at a slow, hesitant pace with prolonged pauses for breath and a comforting quality.
>
> **Example 4 (Tense & Anxious):**
>
> A woman in her thirties with a tight, high-pitched, and metallic voice, her suppressed anxiety appearing in darting eyes and a tense jaw, employs an erratic, rushed prosody characterized by sudden stops and clipped, nervous phrasing.
>
> **Example 5 (Laid-back & Monotone):**
>
> A young man with a flat, mid-range voice and smooth nasal texture, his indifference reflected in a relaxed, blank facial expression, maintains a moderately slow and highly consistent pace devoid of dynamic pitch variation or emotional peaks.

PROMPT FOR INDEXTTS2

The following instruction is supplied to Kimi when generating the emotion prompt for IndexTTS2 (Zhou et al., 2025).

> Describe the emotion of the person in one short sentence to serve as the emotion prompt for a speech-synthesis model. Do not use wording such as "in the video." For example: "A contemplative, reserved, and slightly confused narrative state, with a calm and thoughtful tone"; or "A relaxed, cheerful, and interested narrative state, with an easygoing and engaging tone."

### A.5 IMPLEMENTATION AND EVALUATION DETAILS

**Pretrained models and checkpoints.** For reproducibility, we report the exact pretrained checkpoints and model interfaces used in our data construction and baseline evaluation. For pose-conditioned generation with LTX-2, we use the official `LTX-2.3-22b-IC-LoRA-Union-Control` checkpoint released by Lightricks[1]. The pose-control branch takes the DWPose sequence extracted from the source sign-language video as its structural condition. For Qwen3-TTS, we use the `Qwen3-TTS-12Hz-1.7B-VoiceDesign`

[1] `https://huggingface.co/Lightricks/LTX-2.3-22b-IC-LoRA-Union-Control`

checkpoint[2]. For IndexTTS2, we use the official `IndexTTS-2` release[3]. The public release used in our experiments does not provide the duration-control functionality described by the paper, and we therefore do not impose an explicit target duration during IndexTTS2 inference. For Qwen3.5-Omni-Plus, we use the official API provided by the model developer.

**Audio boundary detection.** The audio boundary used in Bnd-Err is determined by thresholding short-time RMS energy. Let $r_i$ denote the RMS energy of the $i$-th audio frame, converted to dBFS. We use an activity threshold of

$$\tau_a = -40 \text{ dBFS}. \tag{4}$$

The effective speech onset $t_s^a$ and offset $t_e^a$ are defined as the timestamps of the first and last audio frames, respectively, whose RMS energy exceeds $\tau_a$.

**Video motion energy and boundary detection.** We extract human keypoints with DWPose at up to 25 fps and retain keypoints whose confidence scores, as returned by DWPose, are greater than or equal to $\tau_c = 0.3$. For each pair of adjacent valid frames, we compute motion separately over three keypoint groups: upper-body joints, facial landmarks, and hand landmarks. For a group $\mathcal{G}$, its motion energy at frame $t$ is

$$e_t^{\mathcal{G}} = \frac{\sum_{k \in \mathcal{G}_t} w_{t,k} \left\| \mathbf{p}_{t,k} - \mathbf{p}_{t-1,k} \right\|_2 / s_t}{\sum_{k \in \mathcal{G}_t} w_{t,k}}, \qquad w_{t,k} = \min(c_{t-1,k}, c_{t,k}), \tag{5}$$

where $\mathbf{p}_{t,k}$ and $c_{t,k}$ denote the 2D location and confidence of keypoint $k$, respectively. $\mathcal{G}_t$ contains keypoints whose confidence is at least $\tau_c$ in both frames. The normalization scale $s_t$ is the mean person scale of the two frames, where person scale is defined primarily as the distance between the shoulder midpoint and hip midpoint; shoulder width and body extent are used as fallbacks when the required torso keypoints are unavailable.

The frame-wise motion energy is the mean over all valid groups,

$$m_t = \frac{1}{|\mathcal{V}_t|} \sum_{\mathcal{G} \in \mathcal{V}_t} e_t^{\mathcal{G}}, \tag{6}$$

where $\mathcal{V}_t$ is the set of groups containing valid keypoints. Motion is computed only when two valid pose observations are temporally adjacent, allowing a maximum gap of $1.75$ times the expected sampling interval.

We classify a frame as active when $m_t \geq \tau_v$, with $\tau_v = 0.05$. Because $m_t$ represents displacement over the interval preceding frame $t$, the video-motion onset and offset are defined as

$$t_s^v = \max(0, t_{i_s} - \Delta t), \qquad t_e^v = t_{i_e}, \tag{7}$$

where $i_s$ and $i_e$ are the first and last active frames, respectively, and $\Delta t$ is the median interval between sampled frames.

[2] `https://huggingface.co/Qwen/Qwen3-TTS-12Hz-1.7B-VoiceDesign`
[3] `https://huggingface.co/IndexTeam/IndexTTS-2`