# ULTRADUB: TOWARDS AUTHENTIC DUBBING BY UNIFYING VISUALLY-STEERED FLOW LEARNING AND TRAJECTORY GUIDANCE

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

## ABSTRACT

Visual voice cloning requires intelligible, speaker-consistent speech synchronized with visible articulation. However, sequential multimodal conditioning can disrupt previously established temporal and speaker cues, while imbalanced inference guidance can improve linguistic accuracy at the expense of lip synchronization. In this paper, we propose **UltraDub**, a **U**nifying Visually Steered Flow **l**earning and **tra**jectory Guidance **Dub**bing framework that leverages vision in two ways: as continuous motion for multimodal context aggregation, and as structural rhythm for trajectory rectification. Specifically, we introduce the Motion-guided Dual-context Retrieving (MDR) module, which continually recalibrates linguistic and speaker-style retrieval through shared lip-motion query residuals, utilizing independent time-conditioned gates to regulate their contributions. Furthermore, we propose Rhythm-anchored Trajectory Guidance (RTG), a training-free mechanism that evaluates hierarchical multimodal corrections at a visual-only predictive midpoint, safely strengthening semantic conditioning without sacrificing temporal alignment. Finally, we construct DiverseDub, a multi-scenario benchmark to evaluate video dubbing in the wild. Extensive experiments demonstrate that UltraDub achieves state-of-the-art performance across four datasets. Our anonymous project page is available at: `https://ultradub2026.github.io/`

## 1 INTRODUCTION

Visual Voice Cloning (V2C) aims to generate high-fidelity speech for characters in silent videos, jointly conditioned on a text script and reference speech Chen et al. (2022a). As shown in Fig. 1(a), each condition imposes distinct demands: the script specifies linguistic content, the reference speech conveys speaker characteristics, and the silent video defines articulation rhythm and pauses. Driven by these capabilities, V2C demonstrates immense application potential in film post-production and personalized AIGC. However, achieving authentic dubbing is fundamentally more challenging than traditional Text-To-Speech (TTS) tasks, as it requires models to synthesize intelligible, expressive, and style-consistent speech without sacrificing fine-grained audiovisual alignment.

Existing V2C methods primarily pursue two complementary directions: expressive prosody modeling and acoustic pretraining. The former translates hierarchical visual cues into expressive speech (*e.g.*, M2CI-Dubber Zhao et al. (2025) utilizes facial expressions and HPMDubbing Cong et al. (2023) further leverages scene information to extract emotion). However, these methods learn directly from movie corpora and may suffer from poor data quality and background interference Zhang et al. (2024), worsening clarity and Word Error Rate (WER). The latter direction transfers pronunciation knowledge from clean TTS corpora. For example, ProDubber Zhang et al. (2025b) and InstructDubber Zhang et al. (2025a) use pretrained TTS models Li et al. (2023) to improve clarity. However, they rely on duration predictors coupled with these TTS architectures, which enforce rigid, integer-quantized phoneme-to-frame assignments, ultimately degrading lip-sync accuracy and naturalness Jiang et al. (2025). Crucially, both paradigms rely heavily on Montreal Forced Aligner (MFA) McAuliffe et al. (2017), which is an external forced aligner to predefine temporal boundaries, inherently increasing pipeline complexity and error propagation.



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.

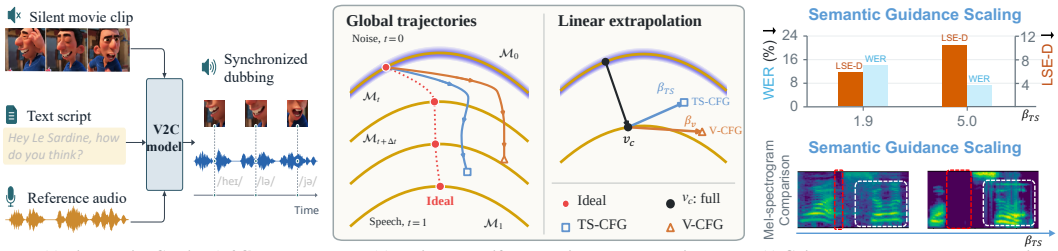



Figure 1: (a) Illustration of V2C task. (b) Conventional CFG relies on linear extrapolation from fully conditioned velocity ($v_c$) via scaling factors $\beta_*$, making it prone to off-manifold trajectory drift. (c) For the multi-conditional V2C task, existing guidance easily triggers imbalance and degradation.

Recent V2C methods explore joint modeling of text, video, and reference speech without explicit MFA supervision, through autoregressive codec prediction Sung-Bin et al. (2025) or non-autoregressive flow matching. By refining the acoustic sequence in parallel, flow matching models avoid dependence on previously generated tokens and improve efficiency. For example, AlignDiT Choi et al. (2025) combines conditional flow matching with auxiliary Connectionist Temporal Classification (CTC) supervision to learn implicit multimodal alignment. Recently, CoSyncDiT Cong et al. (2026) introduces a progressive architecture to sequentially perform speaker adaptation, visual calibration, and text alignment. However, sequential conditioning forces each stage to rely heavily on the current evolving representation, without ensuring that prior cues are preserved. As a result, later semantic updates may gradually deviate from the visual calibration and disrupt established speaker cues, exacerbating a risk of cross-modal interference.

Beyond architectural design, inference poses another challenge for flow-based dubbing methods, which typically rely on classifier-free guidance (CFG). As illustrated in Fig. 1(b), CFG strengthens conditioning through linear extrapolation Wang et al. (2026), adding scaled residuals to the fully conditioned velocity. However, simply amplifying one condition may compromise adherence to others, risking off-manifold drift and artifacts. For example, Fig. 1(c) demonstrates that imbalanced guidance improves clarity but sacrifices synchronization, yielding a sharp surge in LSE-D (↓) Chung & Zisserman (2016). The accompanying mel-spectrograms corroborate this, exhibiting temporal misalignment (red boxes) and spectral distortions (white boxes). These observations reveal a key challenge of authentic dubbing: gains in linguistic accuracy are rendered obsolete if the speech falls out of sync with visual motion. Thus, effective guidance must strengthen conditional adherence while preserving audiovisual temporal alignment.

To achieve authentic dubbing, we propose **UltraDub**, a **U**nifying Visually Steered Flow **l**earning and **tra**jectory Guidance **Dub**bing framework that naturally unifies the dual roles of vision in dubbing without degrading conditional fidelity, deploying it as a continuous *motion* to aggregate multimodal context for velocity field estimation, and as a structural *rhythm* to rectify multimodal guidance along the generative trajectory. Specifically, to govern multimodal conditioning during the flow learning phase, we propose the Motion-guided Dual-context Retrieving (MDR) module. As the hidden states evolve, MDR continually recalibrates linguistic and speaker-style retrieval through shared lip-motion query residuals and adaptively modulates their respective contributions via independent time-conditioned gates. With the velocity field established by MDR, we design Rhythm-anchored Trajectory Guidance (RTG), a training-free mechanism devised to mitigate generative drift during inference. It determines a visual-only predictive midpoint as a structural rhythm anchor, where hierarchical multimodal corrections are exclusively evaluated and safely injected into the fully conditioned velocity. Furthermore, we construct DiverseDub, a challenging V2C dubbing benchmark of 2,595 clips across eight diverse categories. Unlike existing limited datasets, it facilitates robust evaluation by incorporating unconstrained real-world scenarios with varied recording environments and speaking styles.

Our main contributions are summarized as follows:

- We propose **UltraDub**, a visually steered flow framework that unifies multimodal context retrieval and trajectory guidance for V2C. Its MDR module anchors linguistic and speaker-style retrieval to shared lip motion through independently gated context readouts.

- We introduce Rhythm-anchored Trajectory Guidance (RTG), a training-free sampling strategy that evaluates hierarchical multimodal corrections at a visual-only predictive midpoint, incorporating visual rhythm into conditional guidance.
- We construct DiverseDub, a challenging benchmark comprising 2,595 clips across eight categories, extending V2C evaluation to unconstrained recording environments and diverse speaking styles.
- Experiments on three public datasets and DiverseDub demonstrate state-of-the-art performance compared with existing methods.

## 2 Related Work

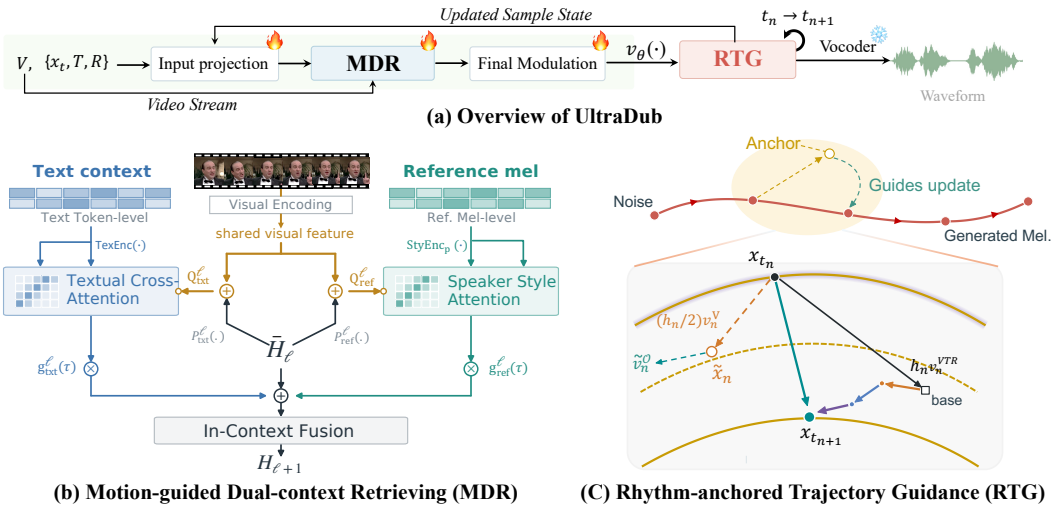
**Visual Voice Cloning.** Visual Voice Cloning (V2C) generates speech for a target video given a script and reference speech, requiring linguistic accuracy, speaker similarity, and lip-synchronization to be satisfied jointly Chen et al. (2022a). Earlier methods explicitly model visual prosody and duration: HPMDubbing and StyleDubber associate lip motion with phoneme timing, whereas ProDubber transfers acoustic knowledge from clean TTS corpora Cong et al. (2023; 2024); Zhang et al. (2025b). More recently, InstructDubber integrates multimodal large language models to provide detailed duration descriptions; however, its audiovisual alignment remains sub-optimal Zhang et al. (2025a). This is because their synchronization heavily relies on forced boundaries or predicted durations, allowing local timing errors to propagate into pronunciation and lip sync. Recent methods replace explicit alignment with joint generation: VoiceCraft-Dub predicts codec tokens autoregressively, while AlignDiT and CoSyncDiT learn implicit alignment with flow-based DiTs Sung-Bin et al. (2025); Choi et al. (2025); Cong et al. (2026). Nevertheless, these flow-based approaches still struggle with cross-modal interference and guidance imbalance. To overcome these limitations, we propose UltraDub, which elegantly unifies the dual roles of vision by employing Motion-guided Dual-context Retrieving (MDR) for robust flow learning, and introducing Rhythm-anchored Trajectory Guidance (RTG) to rectify generative drift during inference, achieving authentic dubbing.

**Classifier-Free Guidance for Flows.** Classifier-free guidance (CFG) strengthens conditional generation by extrapolating the velocity field along the difference between conditional and unconditional predictions Ho & Salimans (2022). In text-to-image diffusion, while this improves prompt adherence, it can distort the sampled distribution and introduce artifacts; thus, CFG++ replaces unconstrained extrapolation with manifold-aware guidance Chung et al. (2025). Flow models inherit this sensitivity under deterministic ODE integration: CFG-Zero* suppresses unreliable early guidance, whereas Rectified-CFG++ anchors conditional corrections to the learned transport path Fan et al. (2025); Saini et al. (2026). However, these methods predominantly address a single semantic condition. In contrast, V2C is more complex, coupling linguistic content, speaker identity, and visual timing within a single trajectory. Amplifying text-speaker guidance may improve pronunciation, but it simultaneously alters the joint velocity governing phoneme onsets, durations, and pauses Choi et al. (2025). Consequently, subsequent residuals are prone to being evaluated along an altered, drifted trajectory. In this work, we evaluate multimodal corrections at a visual rhythm anchor, ensuring that semantic refinements remain bound to the target temporal dynamics.

## 3 Method

### 3.1 Overview

Given a silent video $V$, a script $T$, and a reference audio clip $R$, we learn a conditional velocity field $v_\theta(x_t, t; \mathcal{C})$, with $\mathcal{C} = (V, T, R)$, for noise-to-speech generation. During training, for a clean mel sequence $x_1$ and Gaussian noise $x_0 \sim \mathcal{N}(0, I)$, the intermediate state at flow time $t \in [0, 1]$ is $x_t = (1 - t)x_0 + tx_1$. Unlike previous methods, the proposed UltraDub introduces two novel designs, Motion-guided Dual-context Retrieving (MDR) and Rhythm-anchored Trajectory Guidance (RTG), as illustrated in Fig. 2. First, the input projection initializes prior representations from $x_t$, $R$, and $T$. Then, the proposed MDR module (Fig. 2(b)) absorbs the input prior and dynamic lip-motion features extracted from $V$. Specifically, MDR enables the model to recalibrate the content and speaker-style context retrieval to a shared visual timeline while retaining independent control over each context's contribution by time-conditioned gates. Evolving through MDR, the output result




**(a) Overview of UltraDub**

**(b) Motion-guided Dual-context Retrieving (MDR)**

**(C) Rhythm-anchored Trajectory Guidance (RTG)**

Figure 2: Overview of UltraDub, a **U**nifying Visually Steered Flow **l**earning and **tra**jectory Guidance **Dub**bing framework. The visually steered flow learning phase consists of an input projection, the core Motion-guided Dual-context Retrieving (MDR) layers, and a final modulation to predict the target velocity field $v_\theta$. With the velocity field established, Rhythm-anchored Trajectory Guidance (RTG) acts as a training-free inference mechanism. It constructs a visual-only midpoint as a rhythm anchor to evaluate multimodal corrections, continuously rectifying the base transport velocity.

then undergoes final modulation to predict the target velocity. RTG (Fig. 2(c)) guides multimodal corrections toward consistency with the motion rhythm in a training-free manner. For an integration step of size $h_n = t_{n+1} - t_n$ in flow time, RTG determines the rhythm-related velocity by constructing an auxiliary midpoint anchor. Multimodal corrections are evaluated at this rhythm anchor to adjust the base transport direction defined by fully conditioned velocity. The resulting update is $x_{t_{n+1}} = \Phi^{h_n}_{\mathrm{RTG}}[v_\theta](x_{t_n}, t_n; \mathcal{C})$, where $\Phi^{h_n}_{\mathrm{RTG}}[v_\theta]$ denotes one RTG integration step of size $h_n$ using the learned velocity network $v_\theta$. Each updated state is fed back into the network until $t = 1$. Finally, the generated mel-spectrogram is extracted and decoded into a waveform by a pretrained vocoder.

### 3.2 Input Projection without Early Visual Fusion

Unlike previous methods Choi et al. (2025); Cong et al. (2026) that fuse visual features into initial projection early, the input projection takes only the current noisy state $x_t$, the reference audio $R$, and the script $T$:

$$p_t = \mathcal{F}_{\mathrm{in}}\left([x_t \,\|\, \mathbf{B} \odot \mathrm{MelEnc}(R) \,\|\, \mathrm{TexEnc}(T; \langle \mathrm{pad} \rangle)]\right) \in \mathbb{R}^{L \times D}, \tag{1}$$

where $\mathrm{MelEnc}(\cdot)$ extracts the mel-spectrogram of $R$, $\mathbf{B}$ is a binary mask that retains reference frames Chen et al. (2025), and $\odot$ denotes element-wise multiplication. The text encoder $\mathrm{TexEnc}(\cdot)$ comprises token embeddings and ConvNeXt V2 blocks Woo et al. (2023), and uses the padding symbol $\langle \mathrm{pad} \rangle$ to extend the text sequence to the mel-sequence length $L$. Here, $\|$ denotes channel-wise concatenation, and $\mathcal{F}_{\mathrm{in}}(\cdot)$ is a learnable input projection with convolutional positional encoding. The resulting $p_t$ serves as the initial prior without visual features.

### 3.3 Motion-guided Dual-context Retrieving

**Visual Encoding.** To ensure a fair comparison, we adopt the same visual extractor pipeline $\mathrm{ViEnc}(\cdot)$ Choi et al. (2025) with an AV-HuBERT-based Shi et al. (2022) backbone to encode $V$:

$$\mathbf{z}_{\mathrm{lip}} = (1 - \mathbf{B}) \odot \mathrm{Up}\big(\mathrm{ViEnc}(V)\big) \in \mathbb{R}^{L \times D}, \tag{2}$$

where $\mathrm{Up}(\cdot)$ comprises two transposed-convolution upsampling layers followed by a Conformer encoder to model fine-grained temporal dynamics, while $1 - \mathbf{B}$ retains the target region. As a result, $\mathbf{z}_{\mathrm{lip}}$ captures lip-motion patterns and provides visual rhythmic cues for subsequent retrieval.

**Dual-Memory Retrieval.** Rather than relying on early audio-visual fusion or staged progressive conditioning, MDR establishes a stable dual-memory retrieval mechanism that continually recalibrates linguistic content and speaker-style context aligning with shared visual timing as the DiT hidden states evolve. Let the textual and speaker-style memories be $\mathbf{M}_{\text{txt}} = \text{TexEnc}(T)$ and $\mathbf{M}_{\text{ref}} = \text{StyEnc}_{\text{p}}(\mathbf{B} \odot \text{MelEnc}(R))$, where $\mathbf{M}_{\text{txt}}$ is obtained from the same text encoder used for input projection but retains the text-token resolution. To construct $\mathbf{M}_{\text{ref}}$, $\text{StyEnc}_{\text{p}}(\cdot)$ packs the valid reference frames according to $\mathbf{B}$ and encodes them into a dedicated speaker-style memory using two lightweight convolutional layers with LayerNorm and SiLU activation. At layer $\ell$, let $\bar{\mathbf{H}}_\ell = \text{AdaLN}_\ell(\mathbf{H}_\ell, \boldsymbol{\tau})$, where $\mathbf{H}_0$=$p_t$ and $\boldsymbol{\tau}$ embeds the flow time $t$. Two independently parameterized cross-attention branches retrieve linguistic content and speaker-style cues. For $c \in \{\text{txt}, \text{ref}\}$,

$$\mathbf{r}_c^\ell = \text{MHAttn}_c^\ell\big(\mathbf{Q}_c^\ell,\, \mathcal{K}_c^\ell(\mathbf{M}_c),\, \mathcal{V}_c^\ell(\mathbf{M}_c)\big)\,, \quad \mathbf{Q}_c^\ell = \mathcal{P}_c^\ell(\bar{\mathbf{H}}_\ell) + \mathbf{z}_{\text{lip}}, \tag{3}$$

where $\mathcal{P}_c^\ell$, $\mathcal{K}_c^\ell$, and $\mathcal{V}_c^\ell$ are branch-specific learnable projections. Crucially, both queries are driven by the shared visual residual, enabling continuous recalibration of the two retrieval pathways as DiT hidden states evolve, and the outputs are integrated through independent time-conditioned gates $\mathbf{g}_c^\ell(\boldsymbol{\tau}) \odot \mathbf{r}_c^\ell$. Each gate $\mathbf{g}_c^\ell(\cdot)$ is an independently parameterized, zero-initialized affine projection of $\text{SiLU}(\boldsymbol{\tau})$. This couples content and style retrieval to a common visual timeline while preserving separate control over their contributions.

**In-Context Fusion and Final Modulation.** The in-context fusion module consists of a bidirectional self-attention and a feed-forward network, yielding contextually refined representations for the next MDR layer. After the final MDR layer, the final modulation module applies a linear projection with a time-adaptive layer to predict the target velocity field.

### 3.4 Rhythm-anchored Trajectory Guidance

In V2C, the conditioning modalities impose distinct demands: text specifies linguistic content, reference speech conveys voice characteristics, and video provides the timing. However, stronger content or voice conditioning does not necessarily preserve visual rhythm. Instead of normal extrapolation, RTG addresses this tension by introducing rhythm anchoring and hierarchical multimodal corrections during training-free sampling.

**Rhythm Anchor.** Let $\mathcal{O} \in \{\emptyset, V, VT, VTR\}$ index the retained conditions, with $VTR$ denoting full conditioning $\mathcal{C}$, and define $v_n^{\mathcal{O}} := v_\theta(x_{t_n}, t_n; \mathcal{O})$; in RTG, $v_n^V$ places the auxiliary anchor, whereas $v_n^{VTR}$ supplies the base transport direction. For an integration step $h_n := t_{n+1} - t_n > 0$, we evaluate guidance at a predicted midpoint Saini et al. (2025), with its location determined specifically by the visual-only field:

$$t_n^m := t_n + \frac{h_n}{2}, \qquad \widetilde{x}_n \leftarrow x_{t_n} + \frac{h_n}{2} v_n^V, \tag{4}$$

where text and reference audio inputs are dropped for this prediction, so explicit content and voice guidance do not determine the anchor displacement. The resulting half-step provides a visually directed look-ahead state at which multimodal refinement is assessed, making target motion the reference for subsequent corrections.

**Hierarchical Multimodal Corrections.** At the anchor, we re-evaluate the same network under each condition configuration:

$$\widetilde{v}_n^{\mathcal{O}} \leftarrow v_\theta(\widetilde{x}_n, t_n^m; \mathcal{O}). \tag{5}$$

We organize these predictions into the hierarchy $\emptyset \to V \to VT \to VTR$: visual conditioning establishes the initial contrast, text refines content given vision, and reference acoustics further refine voice given both vision and text. This ordering makes content guidance vision-conditioned and voice guidance content-aware, rather than treating the modalities as independent corrections:

$$\Delta_n^V := \widetilde{v}_n^V - \widetilde{v}_n^\emptyset, \quad \Delta_n^T := \widetilde{v}_n^{VT} - \widetilde{v}_n^V, \quad \Delta_n^R := \widetilde{v}_n^{VTR} - \widetilde{v}_n^{VT}. \tag{6}$$

Evaluating all three contrasts at the same visual anchor provides a common rhythm reference while allowing their strengths to be controlled separately.

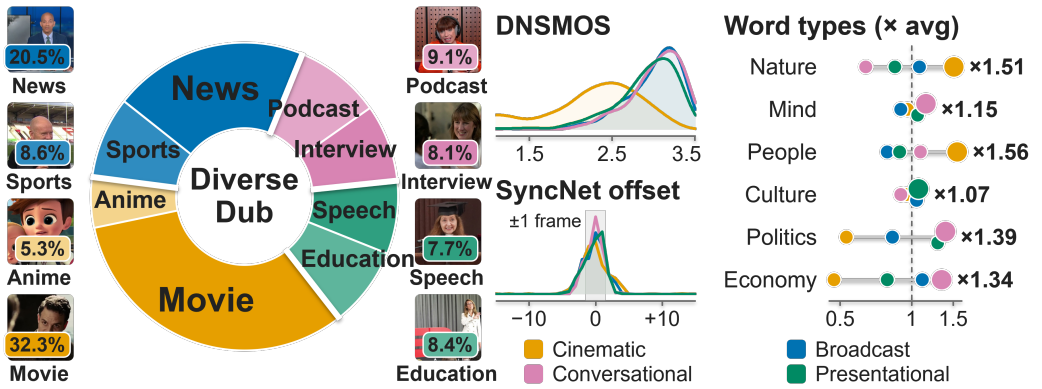



Figure 3: Statistical overview of **DiverseDub**. (a) Left: distribution of clips over the eight dubbing categories, with a representative frame for each category. (b) Middle top: per-domain distributions of clip-level DNSMOS quality. (c) Middle bottom: per-domain distributions of SyncNet audio-visual offsets. (d) Right: relative frequency of word types among the nouns and adjectives of each domain, normalized by the corpus average.

**Guided Flow Update.** RTG combines the current fully conditioned velocity with the above anchored corrections to introduce rhythm guidance:

$$\widehat{v}_n \leftarrow v_n^{VTR} + \alpha(t_n)\left(s_V \Delta_n^V + s_T \Delta_n^T + s_R \Delta_n^R\right), \tag{7}$$

where $s_V, s_T, s_R$ scale the conditional increments, and $\alpha(t_n) = \lambda_{\max} t_n^{\gamma}$ with $\lambda_{\max}, \gamma \geq 0$; for $\gamma = 0$, $\alpha(t_n) = \lambda_{\max}$ throughout. The complete prediction–correction procedure defines the one-step RTG integration operator:

$$\Phi_{\mathrm{RTG}}^{h_n}[v_\theta](x_{t_n}, t_n; \mathcal{C}) := x_{t_n} + h_n \widehat{v}_n. \tag{8}$$

This operator advances the acoustic state by $h_n$ in flow time using the learned velocity network $v_\theta$. Starting from $x_0 \sim \mathcal{N}(0, I)$, we repeatedly apply $x_{t_{n+1}} \leftarrow \Phi_{\mathrm{RTG}}^{h_n}[v_\theta](x_{t_n}, t_n; \mathcal{C})$ until $t = 1$. The generated target mel is then extracted and decoded into speech by a fixed pretrained vocoder.

## 4 DiverseDub Benchmark

To support a comprehensive evaluation under authentic dubbing conditions, we construct **DiverseDub**, an English video dubbing benchmark comprising 2,595 single-shot clips (4.79 h) across eight dubbing categories. In contrast to existing benchmarks dominated by a particular source domain or production format, DiverseDub covers diverse speaking styles, visual compositions, recording environments, and editorial contexts. Fig. 3 summarizes the benchmark from four complementary perspectives: the category distribution characterizes its scenario coverage, the clip-level DNSMOS and SyncNet distributions describe the speech quality, the SyncNet offsets describe the synchronization of the retained samples, respectively, the word-type distribution reflects its lexical diversity. Further details regarding DiverseDub will be provided in the Appendix.

## 5 Experiments

**Datasets.** We evaluate our framework on three public datasets and our newly introduced real-world benchmark, DiverseDub. **LRS3** Afouras et al. (2018) contains TED and TEDx speech, with approximately 119,000 utterances from 5,090 speakers in its pretraining split. For evaluation, we use the LRS3-Cross test set, following the official split of AlignDiT Choi et al. (2025). **GRID** Cooke et al. (2006) contains laboratory recordings with frontal faces and a constrained vocabulary. The test set comprises 3,283 samples. **CelebV-Dub** Sung-Bin et al. (2025) is a recently released public dataset for dubbing. However, its evaluation capacity is highly limited in scale and diversity, containing only 213 test samples. **DiverseDub** extends the evaluation scope to diverse, unconstrained real-world scenarios to assess generalization beyond existing benchmarks (Sec. 4). Across all evaluations, we adhere to the cross-utterance (*i.e.*, Setting2) protocol: the reference audio is always

Table 1: Results on the DiverseDub benchmark. Arrows indicate the preferred direction; the best result among dubbing methods is shown in bold.

| Method | WER (%) ↓ | SPKSIM (%) ↑ | EmoSIM (%) ↑ | LSE-C ↑ | LSE-D ↓ |
|---|---|---|---|---|---|
| GT | 1.07 | 69.92 | 100.00 | 6.64 | 7.83 |
| ProDubber | 13.86 | 25.17 | 69.29 | 2.40 | 11.87 |
| VoiceCraft-Dub | 84.68 | 41.26 | 72.90 | 4.89 | 9.60 |
| AlignDiT | 20.61 | 49.78 | 80.09 | 6.33 | 8.15 |
| InstructDubber | **8.38** | 23.87 | 67.72 | 1.85 | 12.47 |
| CoSyncDiT | 12.01 | 56.32 | 81.72 | 6.88 | 7.86 |
| **UltraDub (Ours)** | 9.60 | **61.15** | **82.93** | **7.13** | **7.56** |

Table 2: Results on the CelebVDub benchmark. Arrows indicate the preferred direction; the best result among dubbing methods is shown in bold.

| Method | WER (%) ↓ | SPKSIM (%) ↑ | EmoSIM (%) ↑ | LSE-C ↑ | LSE-D ↓ |
|---|---|---|---|---|---|
| GT | 4.15 | 66.13 | 100.00 | 6.88 | 7.39 |
| ProDubber | 9.71 | 22.32 | 65.97 | 2.44 | 11.94 |
| VoiceCraft-Dub | 41.19 | 34.65 | 72.89 | 5.46 | 8.83 |
| AlignDiT | 25.86 | 46.25 | 74.89 | 6.29 | 8.01 |
| InstructDubber | **6.47** | 24.82 | 72.43 | 2.43 | 11.96 |
| CoSyncDiT | 12.52 | 49.05 | 78.93 | 6.86 | 7.56 |
| **UltraDub (Ours)** | 9.21 | **56.91** | **80.91** | **7.38** | **7.09** |

sourced from a different utterance of the target speaker. This rigorous setting prevents information leakage and maximally tests the model's robustness for authentic dubbing.

**Implementation Details.** UltraDub operates on mel spectrograms extracted from 16-kHz audio and visual features from 25 FPS video stream. The noisy state, masked reference mel, and 512-dimensional padded text features form a 712-channel input that is projected to width 1024 without visual fusion. Four ConvNeXt V2 blocks produce the 512-dimensional text memory, while two kernel-3 convolution-LayerNorm-SiLU blocks encode packed reference frames into a 256-dimensional memory. Each MDR layer performs two parallel 16-head cross-attention reads (64 dimensions per head) from the text and reference memories. In both branches, the shared lip-motion residuals guide the branch-specific query projections, while keys, values, and attention normalization remain separate. The retrieved features are integrated through the layer's two 1024-dimensional time-conditioned gates, followed by self-attention and an FFN. Training masks 70–100% of valid mel frames. With 32 sampling steps and a sway coefficient of $-1$, the flow-time nodes are $t_n = 1 - \cos(\pi n/64)$, $n = 0, \ldots, 32$. The rhythm anchor velocity constructs the midpoint state $\widetilde{x}_n = x_{t_n} + \frac{h_n}{2} v_n^V$, while the fully conditioned $v_n^{VTR}$ supplies the base velocity. At $t_n^m = (t_n + t_{n+1})/2$, four condition branches are evaluated at this rhythm anchor to obtain $\Delta_n^V$, $\Delta_n^T$, and $\Delta_n^R$. We set $\alpha(t_n) = \lambda_{\max} t_n^\gamma$ with $(\lambda_{\max}, \gamma) = (1, 0)$; hence the common multiplier is one at every step, while the three corrections retain their distinct scales.

**Baselines.** We compare our method against five recent strong baselines: ProDubber Zhang et al. (2025b) (CVPR'25), VoiceCraft-Dub Sung-Bin et al. (2025) (ICCV'25), AlignDiT Choi et al. (2025) (MM'25), InstructDubber Zhang et al. (2025a) (AAAI'26), and CoSyncDiT Cong et al. (2026) (ECCV'25). To ensure a fair comparison, all baseline methods are evaluated using their officially released implementations and pre-trained weights.

**Metrics.** We evaluate content accuracy via Word Error Rate (WER↓) Radford et al. (2023), speaker identity via Speaker Embedding Similarity (SPKSIM↑) Chen et al. (2022b), emotional consistency via Emotion Embedding Similarity (EmoSIM↑) Ma et al. (2024), and lip-synchronization via SyncNet Chung & Zisserman (2016) confidence (LSE-C↑) and distance (LSE-D↓).

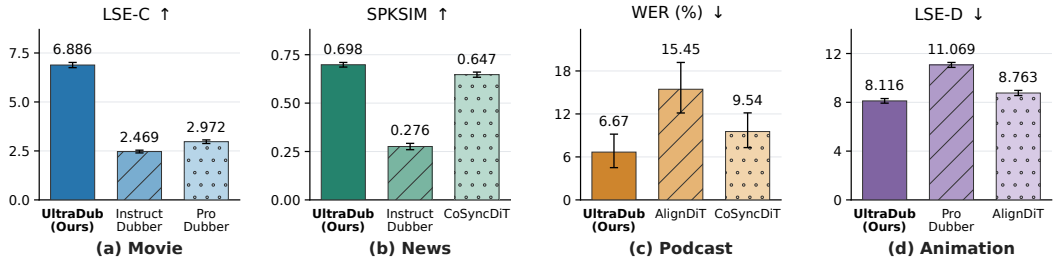



Figure 4: Category-wise comparisons on DiverseDub. Error bars indicate 95% confidence intervals.

Table 3: Comparison with dubbing methods on LRS3 benchmark. Best results among dubbing methods are shown in bold. Note that EmoSIM is not evaluated here, as LRS3 predominantly features neutral TED speeches lacking significant emotional variance.

| Methods | WER (%) ↓ | SPKSIM (%) ↑ | LSE-C ↑ | LSE-D ↓ |
|---|---|---|---|---|
| GT | 2.33 | 56.18 | 7.80 | 6.85 |
| ProDubber | 3.41 | 24.27 | 4.27 | 10.12 |
| VoiceCraftDub | 14.85 | 35.17 | 6.49 | 8.08 |
| AlignDiT | 1.40 | 51.46 | 7.91 | 6.78 |
| InstructDubber | 2.39 | 24.16 | 3.29 | 11.12 |
| CoSyncDiT | 2.09 | 47.53 | 7.83 | 6.88 |
| **UltraDub (Ours)** | **1.37** | **52.05** | **8.23** | **6.50** |

**Results on the DiverseDub Benchmark.** As shown in Tab. 1, UltraDub achieves a strong balance across all evaluation metrics on the unconstrained DiverseDub benchmark. While InstructDubber obtains the lowest WER of 8.38%, it exhibits notable trade-offs in speaker similarity and synchronization, with an SPKSIM of 23.87% and an LSE-D of 12.47. In contrast, UltraDub maintains competitive intelligibility with a WER of 9.60%, alongside stronger identity preservation and lip synchronization. Furthermore, compared with CoSyncDiT's sequential conditioning, UltraDub improves SPKSIM from 56.32% to 61.15% and EmoSIM from 81.72% to 82.93%. These gains are consistent with MDR's design of coordinating linguistic and speaker-style retrieval through shared visual cues. UltraDub also achieves the best synchronization, with an LSE-C of 7.13 and an LSE-D of 7.56, supporting the use of visual timing to organize both context retrieval and RTG-based trajectory correction.

**Category-wise Robustness on DiverseDub.** To further assess performance across production formats, we examine four distinct categories in Fig. 4. In movies and animation, UltraDub demonstrates superior synchronization, achieving an LSE-C of 6.886 and an LSE-D of 8.116, respectively, outperforming the displayed baselines, including the MFA-dependent ProDubber. For news, UltraDub achieves an SPKSIM of 0.698, surpassing CoSyncDiT's 0.647 and InstructDubber's 0.276. For podcasts, it achieves the lowest WER among the displayed methods at 6.67%, compared with 15.45% for AlignDiT and 9.54% for CoSyncDiT. These category-level results demonstrate UltraDub's strengths in synchronization, speaker preservation, and intelligibility across diverse dubbing scenarios.

**Results on the CelebVDub Benchmark.** As shown in Tab. 2, our method achieves the best speaker similarity, emotion similarity, and lip synchronization among the compared dubbing methods. Compared with CoSyncDiT, it increases SPKSIM from 49.05% to 56.91% and EmoSIM from 78.93% to 80.91%, while improving LSE-C from 6.86 to 7.38, reducing LSE-D from 7.56 to 7.09, and lowering WER from 12.52% to 9.21%. Although InstructDubber yields lower WER, its substantially weaker lip synchronization underscores the need to coordinate linguistic accuracy with visual timing. These results support UltraDub's visually steered design, which anchors context retrieval and trajectory correction to lip motion, jointly improving alignment and voice fidelity while retaining competitive intelligibility.

**Results on the LRS3 Benchmark.** As shown in Tab. 3, UltraDub achieves state-of-the-art performance across all four evaluated metrics among the compared dubbing methods. Compared with AlignDiT, the strongest baseline on all metrics, UltraDub reduces WER from 1.40% to 1.37% and

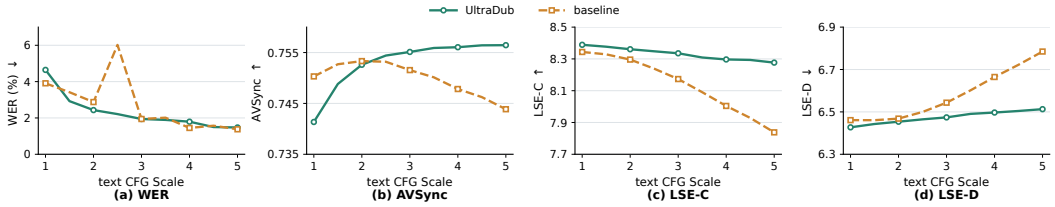



Figure 5: Preserving Audiovisual Alignment under Increasing Text Guidance Strength.

Table 4: Comparison with dubbing methods on the GRID benchmark. EmoSIM is not evaluated because GRID consists primarily of neutral emotion.

| Methods | WER (%) ↓ | SPKSIM (%) ↑ | LSE-C ↑ | LSE-D ↓ |
|---|---|---|---|---|
| GT | 13.14 | 63.69 | 7.14 | 6.84 |
| ProDubber | 44.29 | 34.91 | 3.71 | 11.08 |
| VoiceCraftDub | 64.47 | 33.37 | 5.13 | 8.50 |
| AlignDiT | 19.47 | 55.06 | 7.17 | 6.84 |
| InstructDubber | 27.67 | 36.30 | 3.33 | 11.32 |
| CoSyncDiT | 19.64 | 53.45 | 7.21 | 6.86 |
| **UltraDub (Ours)** | **14.53** | **58.01** | **7.60** | **6.44** |

improves SPKSIM from 51.46% to 52.05%. The gains are more pronounced in synchronization: LSE-C increases from 7.91 to 8.23, while LSE-D decreases from 6.78 to 6.50. These results show that UltraDub improves audiovisual alignment alongside linguistic accuracy and speaker similarity, supporting our approach of using visual motion to coordinate multimodal retrieval and guide the rhythm anchored trajectory.

**Results on the GRID Benchmark.** As shown in Tab. 4, UltraDub outperforms all compared dubbing methods across the four metrics. It reduces WER from AlignDiT's 19.47% to 14.53%, approaching the ground-truth WER of 14.00%, while achieving the highest SPKSIM of 58.01%. It also delivers the best synchronization, with an LSE-C of 7.60 and an LSE-D of 6.44, demonstrating joint gains in intelligibility, speaker preservation, and alignment.

**Robustness to guidance strength.** Fig. 5 examines linguistic accuracy and audiovisual synchronization as the text CFG scale increases. We additionally report AVSync Choi et al. (2025), which measures the cosine similarity between AV-HuBERT features of the same video paired with synthesized and ground-truth speech; higher scores indicate better synchronization. Increasing the scale from 1 to 5 reduces UltraDub's WER from 4.65% to 1.48% while improving AVSync from 0.7414 to 0.7565. By contrast, the baseline exhibits greater WER fluctuations, and its AVSync declines after peaking near scale 2. Although both methods show decreasing LSE-C and increasing LSE-D, UltraDub degrades more gradually and consistently outperforms the baseline on both metrics. At scale 5, UltraDub yields only slightly higher WER, 1.48% versus 1.38%, while achieving better AVSync, 0.7565 versus 0.7439, higher LSE-C, 8.277 versus 7.838, and lower LSE-D, 6.513 versus 6.785. These results support the rationale behind rhythm-anchored guidance: evaluating multimodal corrections at a shared visual anchor helps accommodate stronger text conditioning while limiting synchronization degradation.

## 6 Conclusion

In this paper, we propose UltraDub to coordinate multimodal context retrieval through MDR and anchor training-free trajectory guidance through RTG. Together, these components establish visual timing as a shared reference for flow learning and sampling. We also introduce DiverseDub, a benchmark spanning eight categories for evaluating dubbing in unconstrained real-world scenarios. Experiments on four benchmarks demonstrate improvements in intelligibility, speaker preservation, and lip synchronization. Further analysis shows that UltraDub limits alignment degradation under stronger guidance, supporting more balanced multimodal conditioning during synthesis.

## REFERENCES


Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018.

Brandon Castellano. PySceneDetect: Video cut detection and analysis tool. `https://github.com/Breakthrough/PySceneDetect`, 2026. Version 0.7.1.

Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2C: visual voice cloning. In *CVPR*, pp. 21210–21219, 2022a.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518, 2022b.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.313.

Jeongsoo Choi, Ji-Hoon Kim, Kim Sung-Bin, Tae-Hyun Oh, and Joon Son Chung. Aligndit: Multimodal aligned diffusion transformer for synchronized speech generation. In *ACM MM*, pp. 10758–10767, 2025.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *ICLR*, volume 2025, pp. 30824–30850, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV 2016 Workshops*, pp. 251–263, 2016.

Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to dub movies via hierarchical prosody models. In *CVPR*, pp. 14687–14697, 2023.

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. *arXiv preprint arXiv:2402.12636*, 2024.

Gaoxiang Cong, Liang Li, Jiaxin Ye, Zhedong Zhang, Hongming Shan, Yuankai Qi, and Qingming Huang. Cosyncdit: Cognitive synchronous diffusion transformer for movie dubbing. *arXiv preprint arXiv:2604.12292*, 2026.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, et al. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*, 2025.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *NeurIPS*, 2023.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of ACL*, pp. 15747–15760, 2024.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pp. 498–502, 2017.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pp. 28492–28518, 2023.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. The UCREL semantic analysis system. In *Proceedings of the Workshop on Beyond Named Entity Recognition: Semantic Labelling for NLP Tasks (LREC)*, pp. 7–12, 2004.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP*, pp. 6493–6497. IEEE, 2021.

Shreshth Saini, Shashank Gupta, and Alan Bovik. Rectified CFG++ for flow based models. In *NeurIPS*, pp. 149034–149074, 2025.

Shreshth Saini, Shashank Gupta, and Alan Bovik. Rectified cfg++ for flow based models. *NeurIPS*, 38:149034–149074, 2026.

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

Kim Sung-Bin, Jeongsoo Choi, Puyuan Peng, Joon Son Chung, Tae-Hyun Oh, and David Harwath. Voicecraft-dub: Automated video dubbing with neural codec language models. *arXiv preprint arXiv:2504.02386*, 2025.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3927–3935, 2021.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. Cfg-ctrl: Control-based classifier-free diffusion guidance. *arXiv preprint arXiv:2603.03281*, 2026.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, pp. 16133–16142, 2023.

Zhedong Zhang, Liang Li, Gaoxiang Cong, YIN Haibing, Yuhan Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In *ACM MM*, 2024.

Zhedong Zhang, Liang Li, Gaoxiang Cong, Chunshan Liu, Yuhan Gao, Xiaowan Wang, Tao Gu, and Yuankai Qi. Instructdubber: Instruction-based alignment for zero-shot movie dubbing. *arXiv preprint arXiv:2512.17154*, 2025a.

Zhedong Zhang, Liang Li, Chenggang Yan, Chunshan Liu, Anton van den Hengel, and Yuankai Qi. Prosody-enhanced acoustic pre-training and acoustic-disentangled prosody adapting for movie dubbing. *arXiv preprint arXiv:2503.12042*, 2025b.

Yuan Zhao, Rui Liu, and Gaoxiang Cong. Towards expressive video dubbing with multiscale multimodal context interaction. In *ICASSP*, pp. 1–5, 2025.

# A APPENDIX

## AI USE STATEMENT

In this work, we used generative AI tools to implement methods and clean the dataset. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, assist in writing proofs, propose or refine hypotheses, design or provide feedback on research methodologies or experiments, implement methods, support qualitative and thematic data analysis, or interpret results. Generating synthetic data sets, formulating mathematical claims, or providing critical ingredients for proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## A.1 BENCHMARK DETAILS AND CONSTRUCTION

**Data Sources and Cleaning.** We collect English audiovisual content from YouTube using category-specific queries and curated movie and animation playlists designed to cover diverse production formats and real-world dubbing conditions. Each source video is first segmented into individual shots using pySceneDetect Castellano (2026). We then detect and track visible faces, identify the active speaker by associating face tracks with the corresponding speech using TalkNet Tao et al. (2021), and crop the resulting tracks into speaker-centered clips. Segments containing shot transitions, multiple active speakers, overlapping speech, within-clip speaker changes, or unreliable face–speech associations are removed.

**Dubbing Categories** We assign each clip a single primary category according to the production format or dominant editorial context of its source video, and group the categories into four domains: *Cinematic* (*Movie*, *Anime*), *Broadcast* (*News*, *Sports*), *Conversational* (*Interview*, *Podcast*), and *Presentational* (*Speech*, *Education*). These categories cover scripted and animated productions, broadcast and event coverage, conversational programs, and public and educational speech. In particular, *Movie* includes live-action films and film-related media, while *Sports* primarily contains condensed sports and event coverage. *Podcast* also includes moderated multi-participant programs and documentary-style discussions, although each retained clip contains only one active speaker. The resulting distribution, shown in Fig. 3(a), provides broad coverage without allowing a single content domain to dominate the benchmark.

**Construction Pipeline** Following structural cleaning, we perform evaluation-oriented curation to ensure both broad scenario coverage and reliable ground-truth annotations. Candidates are selected across all eight categories while avoiding excessive concentration in any individual source domain. We retain only clips with reliable audio–visual synchronization and adequate DNSMOS Reddy et al. (2021) speech quality. We additionally compute SyncNet Chung & Zisserman (2016) statistics for every retained clip and report their complete per-domain distributions in Fig. 3, rather than reporting only aggregate averages. The mean DNSMOS is 2.73, with domain means ranging from 2.31 to 3.03, and 63.7% of clips have offsets within $\pm 1$ frame.

For transcript construction, Whisper large-v3 Radford et al. (2023) first produces an initial transcription. Segments with low ASR confidence, measured by the negative log-likelihood of the transcription given the speech, are discarded, and incomplete sentences at segment boundaries are trimmed. Each resulting clip is re-transcribed and retained only if its WER falls below a strict threshold. An LLM agent then cross-checks the transcriptions and corrects evident recognition errors, followed by manual verification. Samples with unresolved transcript ambiguities, inconsistent speaker identity, synchronization failures, or uncertain textual annotations are discarded. The word-type distribution in Fig. 3(d) is computed from the final verified transcripts using USAS Rayson et al. (2004) semantic fields and provides a complementary view of the linguistic content covered by DiverseDub.

## A.2 MEL-SPECTROGRAMS COMPARISION

Fig. 6 presents qualitative mel-spectrogram comparisons, with red and white boxes highlighting alignment intervals (onsets, offsets, pauses) and local phonetic details, respectively. Across all

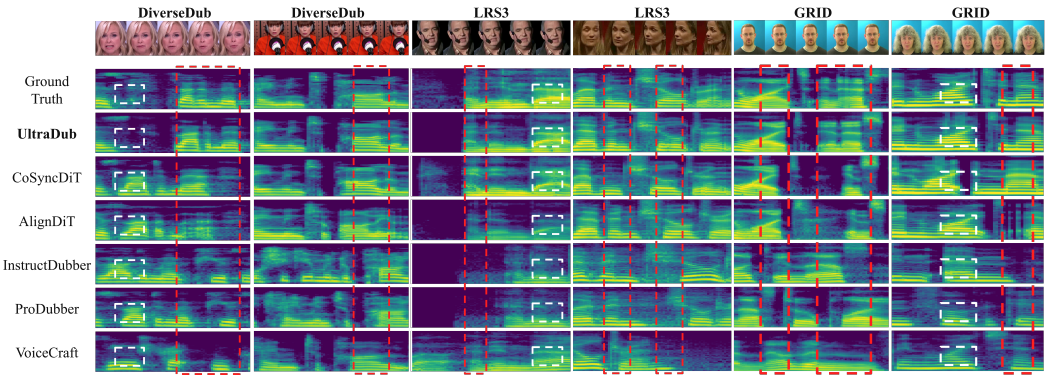



Figure 6: The mel-spectrograms of ground truth and synthesized results by different models in DiverseDub, LRS3 and GRID. The red and white rectangles highlight the details of the audio-visual alignment of the SOTA model.

benchmarks, UltraDub produces temporal boundaries and alignment trajectories that most closely resemble the ground truth, precisely synchronizing vocal events with lip movements. Although AlignDiT and CoSyncDiT generally capture the coarse timing of speech, they both exhibit local boundary shifts, delayed onsets, and inaccurate syllable offsets within the highlighted regions. The discrepancies are far more catastrophic for InstructDubber, ProDubber, and VoiceCraft, which trigger speech events at completely incorrect time steps and lose temporal alignment altogether.