AI-Assisted Multilingual Film Dubbing: An Integrated Framework for Speech Translation, Emotion Preservation and Lip Synchronization

Somnath Satpal Jagtap
Dept. of MCA, D. Y. Patil Inst. of MCA and Management
Akurdi, Pune, Maharashtra, India

Vaishnavi Santosh Adhav
Dept. of MCA, D. Y. Patil Inst. of MCA and Management
Akurdi, Pune, Maharashtra, India

Prerna Dipak Jadhav
Dept. of MCA, D. Y. Patil Inst. of MCA and Management
Akurdi, Pune, Maharashtra, India

Under the Guidance of: Mrs. Pranita Manjare
Department of MCA, D. Y. Patil Institute of MCA and Management, Akurdi, Pune, Maharashtra, India

Abstract

Automatic film dubbing aims to replace a source actor's speech with target-language dialogue that preserves semantic meaning, temporal isochrony, vocal timbre, emotional prosody, and visual lip synchronization. However, a systematic audit of primary dubbing literature demonstrates that existing architectures optimize isolated subsets of these five objectives, inducing severe cross-modal degradation when combined directly. Enforcing strict phoneme duration control in sequence-to-sequence translation reduces translation quality by 9.0% to 14.0% relative SacreBLEU on CoVoST-2 (and 17.1% on pause-heavy turns), prompting large language models for viseme-compatible paraphrases degrades ASR-BLEU by 3.00 points, applying post-hoc voice conversion reduces intelligibility by 1.07 ASR-BLEU points, heuristic SSML emphasis lowers TTS naturalness by 0.30 MOS, and high-intensity emotional flow-matching guidance (Sad) inflates word error rate (WER) from 11.32% to 12.12%. Conversely, large-scale empirical analyses of 42,850 professional human dub lines prove that human translators do not sacrifice semantic quality on-screen ($p = 0.349$), maintain moderate speech overlap (0.658), and match exact source-target visemes in only 12.4% of on-screen speech time. Grounded in these findings, this paper presents a critical synthesis of eight foundational dubbing studies and formulates **CineDub AI**, a four-stage, eight-component conceptual research framework. CineDub AI proposes: (1) a continuous mouth-visibility gate $M_{vis}(t) \in [0, 1]$ that interpolates allowable duration budgets $\mathcal{I}_{tgt}(M_{vis})$ and causal duration positional encodings between off-screen narration and close-ups; (2) a band-masked cross-lingual soft lip-energy attention mechanism (X-LPA) that aligns local syllable jaw-opening envelopes without assuming invalid cross-lingual 1-to-1 phoneme-viseme identity; (3) a Gaussian soft-upsampled surrogate SyncNet timing loss that enables differentiable duration refinement; and (4) an articulation-gated Optimal-Transport Conditional Flow Matching (OT-CFM) decoder that combines deterministic consonant masks M_{cons} and upsampled CTC posterior entropy $\tilde{\mathcal{H}}_{CTC}$ with simplex-bounded source emotion priors. Finally, we specify a falsifiable, power-analyzed evaluation protocol distinguishing completed literature analysis from proposed conceptual methods and future empirical validation.

Index Terms—Automatic dubbing, speech-to-speech translation, isochrony, emotion-controllable speech synthesis, speaker identity preservation, lip synchronization, conditional flow matching, multimodal machine learning.

I. Introduction

A. Context and Motivation

Global distribution of film, television, and educational media increasingly requires high-quality localization across diverse linguistic markets [1, 2]. Whereas text subtitling divides viewer visual attention away from actors' facial performances, studio voice dubbing demands an expensive, multi-stage production pipeline encompassing script translation, *rythmo-band* timing adaptation, voice-actor casting, dramatic re-performance, and acoustic post-production mixing [1,2,6]. Automatic dubbing (AD) seeks to automate this workflow by extending Speech-to-Speech Translation (S2ST) to satisfy the strict perceptual constraints of audiovisual media [1, 4, 8].

B. The Five-Way Optimization Conflict in Automatic Dubbing

Unlike conventional S2ST—which optimizes semantic adequacy (BLEU, COMET) and basic intelligibility (WER)—film dubbing is governed by five simultaneous, mutually competing objectives [1, 2, 6–8]:

1. **Semantic Fidelity vs. Temporal Isochrony:** Character-count isometry (IsoMT [15]) correlates poorly with actual speech overlap ($r = 0.279, r^2 = 0.078$, with $> 69\%$ of human dubs violating $\pm 10\%$ character length [2]) and yields zero improvement in Speech Overlap over standard MT (SO = 0.53 vs. 0.53 [5]). Conversely, forcing exact phoneme duration matching in neural machine translation (Txtd2PhnD) boosts SO to 0.92 on CoVoST-2 (0.70 on pause-heavy test101) at the cost of a 14.0% (17.1% on test101) relative drop in SacreBLEU [5], while post-hoc prosodic alignment and spline time-scaling degrade native Ital-

ian listener satisfaction by −10.93 MUSHRA points ($p < 0.01$) [1].

2. **Lip Synchrony vs. Lexical & Visual Integrity:** Prompting a large language model (Claude 3.0 Sonnet) to paraphrase translations for viseme compatibility improves English-to-Spanish SyncNet Lip-Sync Error Distance (LSE-D) from 10.68 to 10.18 but degrades ASR-BLEU by −3.00 points (35.18 → 32.18) [7]. Conversely, editing actor faces via 6-channel conditional generative adversarial networks (cGANs) struggles on bearded faces, worsens LRS2 FID (8.86 vs. 4.44 for Wav2Lip-GAN [3]), and raises severe deepfake likeness and consent concerns [7].

3. **Emotional Expressiveness vs. Phonetic Intelligibility:** Passing speech through text transcripts erases paralinguistic pitch (F_0) and energy contours [2, 8]. Injecting heuristic SSML word-emphasis tags into FastSpeech 2 reduces speech naturalness by −0.30 MOS (3.59 → 3.29) [3], whereas applying high-intensity flow-matching emotional guidance ($\alpha = 3.5$) for low-energy emotions (Sad) blurs phoneme articulation and increases Whisper WER from 11.32% (Neutral) and 10.22% (Happy) to 12.12% [6].

4. **Speaker Identity vs. Articulation Clarity:** Applying separate post-hoc voice conversion (CosyVoice VC) to translated speech increases speaker similarity (SIM) from 0.266 to 0.315 but degrades ASR-BLEU by −1.07 points (24.16 → 23.09) due to formant oversmoothing [8].

5. **Cross-Family Syntactic Reordering:** Seven of the eight primary dubbing studies evaluate exclusively Indo-European Subject-Verb-Object (SVO) languages [1–3, 5–8]. Extending dubbing to Subject-Object-Verb (SOV) Indo-Aryan and Dravidian languages (e.g., Hindi, Marathi, Telugu [9, 10]) inverts clause-level word order, breaking monotonic position-based prosody and viseme mapping.

C. Epistemic Scope and Paper Contributions

In strict adherence to scientific reporting standards, this manuscript is presented as a **systematic literature audit, mathematical formulation, and conceptual framework study**. We explicitly distinguish: (i) empirical findings reported by prior studies [1–16], (ii) analytical conclusions synthesized from our cross-paper audit, (iii) newly proposed conceptual modules and mathematical adaptations in **CineDub AI**, (iv) untested falsifiable hypotheses (H1–H3), and (v) future empirical implementation work. No experimental results or benchmark scores are fabricated for CineDub AI.

II. Background and Related Work

A. Acoustic Prosodic Alignment and Empirical Localization Studies

Federico *et al.* [1] introduced an end-to-end automatic dubbing pipeline on 24 MuST-C English-to-Italian TED Talk clips, augmenting cascaded S2ST with target-length NMT verbosity tokens (short, normal, long), dynamic programming Prosodic Alignment (PA), Mel-spectrogram spline time-scaling, and U-Net foreground-background separation paired with Maximum Likelihood Estimation (MLE) of reverberation time (RT_{60}). Brannon *et al.* [2] subsequently analyzed 319.57 hours of English Prime Video originals across 54 shows (42,850 aligned dialogue lines in En-De and En-Es, with 3,617 lines carrying on/off-screen script annotations). Their study proved that human dubbers prioritize natural speaking tempo (dub speaking-rate standard deviation is significantly lower than source speech: 1.25 vs. 1.47 w/s in Spanish and 1.26 vs. 1.46 w/s in German, $p < 10^{-10}$), maintain moderate temporal overlap (0.658 mean; 0.684 on-screen vs. 0.662 off-screen, $p = 0.003$), do not sacrifice translation adequacy on-screen (COMET-QE $t = 0.936, p = 0.349$), and transfer word-level pitch and energy variations across FastAlign word pairs ($\Delta r = +0.097$ to $+0.125, p < 10^{-6}$).

B. Duration-Controlled NMT and Expressive Video-to-Speech (V2C)

To reduce post-hoc audio stretching, Lakew *et al.* [15] proposed character-level Isometric MT (IsoMT). Addressing character-to-duration mismatch across writing systems, Wu *et al.* [4] (VideoDubber) introduced learnable [P] pause tokens (> 0.3 s), a CNN token duration predictor, and absolute/relative remaining duration positional encodings (PE, $N = 5$) across Es-En, De-En, Zh-En, and En-Zh, paired with vowel-only scaling in zero-shot AdaSpeech 4. Chronopoulou *et al.* [5] (Txtd2PhnD) translated source German text and 100-bin quantized phrase durations directly into interleaved English phonemes and durations trained with Gaussian noise ($\sigma = 0.1$), overriding FastSpeech 2's native duration predictor.

In monolingual computer vision, script-to-video dubbing (Visual Voice Cloning, V2C) evolved from Neural Dubber [16], HPMDubbing [14], and StyleDubber [13] to Cong *et al.*'s [6] EmoDubber (CVPR 2025). EmoDubber combines Lip-related Prosody Aligning (LPA) using Duration-Level Contrastive Learning (DLCL), Pronunciation Enhancing (PE) via Monotonic Alignment Search (MAS) and a 5-block Audio-Visual Efficient Conformer (AVEC) with CTC loss, Speaker Identity Adapting (SIA), and Flow-based User Emotion Controlling (FUEC) using Optimal-Transport Conditional Flow Matching (OT-CFM) steered by Positive and Negative Guidance (PNGM).

C. Audio-Visual and Textless Speech-to-Speech Translation

Waibel *et al.* [3] (Face-Dubbing++) combined Conformer-LSTM ASR, NMT cross-attention emphasis projection, FastSpeech 2 SSML prosody scaling, VQMIVC voice conversion, and a 6-channel lip-sync cGAN. Avoiding facial pixel manipulation, Goncalves *et al.* [7] (ICASSP 2025) fine-tuned the unit duration predictor of a direct AV-HuBERT S2UT backbone (AV2AV [12]) using a joint SyncNet cosine-similarity loss $\mathcal{L}_{sync}$ and log-MSE duration loss ($\lambda = 10$) on LRS3. In textless S2ST, Choi *et al.* [8] (Dub-S2ST, EMNLP 2025) combined mHuBERT discrete units ($K = 1000$), unit-based speaking-speed adaptation (r_{src}/r_{tgt}), a non-autoregressive Diffusion Transformer (DiT) initialized at exact source unit length L_{src}, and an in-context OT-CFM Mel synthesizer + HiFi-GAN. Complementary external studies have investigated video-guided LLM-TTS duration control (DubWise [9]), cross-lingual emotional TTS (VECL-TTS [10]), and joint voice-isochrony S2ST (TransVIP [11]).

III. Critical Literature Synthesis

Table I synthesizes the architectural capabilities of the eight primary studies across all core dubbing dimensions, while Table II compiles their exact verified quantitative results and failure modes. **Cross-Dataset Interpretation Rule:** Because each study evaluates distinct datasets (MuST-C, CVSS-C, CoVoST-2, Chem, GRID, LRS2/3), language pairs, and ASR/TTS backbones, numerical metrics in Table II must be interpreted strictly relative to each study's internal baselines rather than as direct cross-paper rankings.

IV. Research Gap and Problem Formulation

A. Reassessment of Research Gaps (G1–G4)

Evaluating candidate gaps against both the eight primary studies [1–8] and eight external baselines [9–16] yields three established algorithmic/methodological gaps (G1–G3) and one exploratory cross-family extension (G4-Exploratory), summarized in Table III:

1. **Gap G1 (Cross-Lingual Non-Monotonic Lip-Energy & Automated Emotion Unification):** Cross-lingual S2ST models achieve duration compliance and speaker similarity (VideoDubber [4], Dub-S2ST [8], TransVIP [11]) or unit-level lip sync (Goncalves et al. [7], AV2AV [12]), whereas monolingual V2C models (EmoDubber [6], StyleDubber [13], HPM-Dubbing [14], Neural Dubber [16]) achieve fine-grained lip-prosody alignment and emotional flow matching. However, EmoDubber's MAS and DLCL ($\mathcal{L}_{cl}$) rely on a ground-truth binary phoneme-to-lip alignment matrix $M^{gt}_{lip,pho} \in \{0,1\}^{F\times P}$ extracted via forced alignment on *same-language* audio-video pairs [6]. In cross-lingual dubbing, target phonemes P_{tgt} and source lip frames F_{src} belong to different languages with distinct phoneme lengths and word orders, rendering $M^{gt}_{lip,pho}$ non-existent. Moreover, EmoDubber requires manual user entry of emotion parameters $\{c, \alpha, \beta\}$ [6].

2. **Gap G2 (Visibility-Conditioned and Articulation-Gated Multi-Objective Trade-Off Control):** Enforcing uniform duration, lip-sync, and emotional guidance weights across an utterance degrades semantic adequacy (-9.0% to -17.1% relative BLEU [5]; -3.00 ASR-BLEU [7]) and consonant intelligibility (12.12% WER on Sad [6]). No existing framework conditions translation duration slack and SyncNet refinement on continuous mouth visibility $M_{vis}(t)$ [2] or dampens OT-CFM emotional guidance on consonant transitions.

3. **Gap G3 (Context-Stratified, Non-Circular Multidimensional Evaluation):** Prior evaluations exhibit methodological blind spots: EmoDubber [6] evaluates Emo-IS using the *same* 1D-CNN classifier ψ used for PNGM guidance (circular bias), Goncalves *et al.* [7] apply intrusive PESQ (1.044) to non-aligned translated speech, and Federico *et al.* [1] prove that non-native listeners inflate MUSHRA ratings by $+10.34$ points while missing speaking-rate disfluencies (-10.93 points for natives).

4. **Gap G4-Exploratory (Cross-Family SVO-to-SOV Clause Inversion):** While external TTS studies (DubWise [9], VECL-TTS [10]) evaluate pre-translated Hindi text, end-to-end audio-visual dubbing from SVO English into SOV Indo-Aryan/Dravidian languages (En-Hi, En-Mr) must resolve clause-level word-order inversion when projecting prosodic stress and aligning syllable jaw-opening cycles.

V. Research Questions and Hypotheses

We formulate one overarching research question, three sub-questions (RQ1–RQ3), and three falsifiable hypotheses (H1–H3) whose complete statistical power and rejection criteria are specified in Section VIII:

- **Main RQ:** *How can a conceptual automatic dubbing framework mathematically unify cross-lingual duration-aware translation, automated source emotion and speaker timbre preservation, and audio-visual lip synchronization by conditioning its optimization objectives on time-varying mouth visibility and consonant articulation vulnerability?*

- **RQ1 (Visibility-Modulated Duration & Sync Control):** *To what extent can conditioning NMT target duration budgets $\mathcal{I}_{tgt}(\bar{M}_{vis})$ and Gaussian soft-upsampled SyncNet duration refinement on continuous mouth visibility $M_{vis}(t) \in [0,1]$ improve off-screen semantic adequacy (SacreBLEU, COMET-QE,*

TABLE I
MASTER ARCHITECTURAL AND FUNCTIONAL COMPARISON ACROSS THE EIGHT PRIMARY AUTOMATIC DUBBING STUDIES

Study & Venue Status	Paradigm	Language Pairs	Translation Mode	Duration & Isochrony Control	Emotion & Prosody Transfer	Speaker Identity	Lip-Sync Mechanism	Primary Verified Datasets
1. Federico et al. [1] *IWSLT 2020 (Peer-Rev. Conf.)*	Cascaded S2ST + Acoustic Render	En→It (Indo-Eur.)	Text NMT (3 length-ratio tokens)	Verbosity token + DP PA + Mel spline scaling	Phrase pauses + blind MLE RT_{60} RIR	None (2 fixed Italian TTS voices)	None (phrase audio timing only)	MuST-C En-It (24 TED clips, 14 raters, 657 ratings)
2. Brannon et al. [2] *TACL 2023 (Peer-Rev. Journal)*	Empirical Human Localization Study	En→De, Es (Indo-Eur.)	Human Dubbing vs. Subtitles	Analyzed IoU overlap (0.658) & duration $r = 0.877$	Analyzed line R^2 & FastAlign word $\Delta F_0, \Delta E$ ($p < 10^{-6}$)	Analyzed 9,215 voice actors	Polly viseme co-occurrence (12.4% exact on-screen)	Prime Video corpus (319.6 h src; 42,850 aligned lines; 3,617 on/off tags)
3. Waibel et al. [3] *ICASSPW 2023 & arXiv 2022*	5-Stage Audio-Visual Cascade	En→De (Indo-Eur.)	Text NMT + Attention Stress Proj.	None (unconstrained TTS duration; lip cGAN edits face)	Binary word-emphasis ASR/NMT → FastSpeech 2 SSML	Post-hoc VQMIVC voice conversion	6-ch 256 × 256 lower-face cGAN + SyncNet	LibriSpeech, TED-LIUM, WMT, CSS10, LRS2, 262-clip set (25 spk)
4. Wu et al. [4] *AAAI 2023 (Peer-Rev. Conf.)*	Speech-Aware Length NMT + TTS	Es-En, De-En, Zh-En, En-Zh	Text NMT w/ Token Duration	CNN duration pred. + Abs/Rel Duration PE ($N = 5$) + [P]	Inter-word pause token [P] (> 0.3 s) + vowel scaling	Zero-shot AdaSpeech 4 (not eval. via SIM)	None (utterance & pause duration only)	CVSS (Es/De/Zh-En), MuST-C En-Zh, 1h film set (42 clips, 892 sent.)
5. Chronopoulou et al. [5] *arXiv 2023 (Preprint)*	Joint Text-to-Phoneme & Duration MT	De→En (Indo-Eur.)	Text → Interleaved Phoneme+Dur	100-bin duration tokens + Gaussian noise ($\sigma = 0.1$)	Pause-bounded phrase timing (≥ 300 ms)	Single-speaker FastSpeech 2	None (phoneme duration sum only)	CoVoST-2 De-En (289k train, 15.4k test) + test91 & test101
6. Cong et al. [6] *CVPR 2025 (Peer-Rev. Conf.)*	Monolingual Emotion V2C Synthesis	Monolingual En only (No MT)	None (Takes ground-truth target script)	DLCL ($\mathcal{L}_{cl}$) + MAS locked to video frames F	FUEC (OT-CFM) + manual user PNGM $\{c, \alpha, \beta\}$	SIA + USL + GE2E vector (SECS = 90.62)	Audio-to-lip LPA cross-attention + AVEC-CTC	Chem (6,082 train / 196 test), GRID (33 spk), Dub 3.0, Emobox
7. Goncalves et al. [7] *ICASSP 2025 (arXiv 2024v1)*	Direct Audio-Visual S2UT + SyncNet	En→Es, Pt, It, Fr (Indo-Eur.)	Direct AV-HuBERT → Speech Units	Unit Duration Predictor trained w/ $\mathcal{L}_{dur}$ ($\lambda = 10$)	None (generic unit HiFi-GAN vocoder)	None (deliberately omitted for consent)	Differentiable SyncNet loss $\mathcal{L}_{sync}$ on raw video	LRS3 (4,004 trainval clips, 30 h; 412 test clips, 51 min)
8. Choi et al. [8] *EMNLP 2025 Findings (Peer-Rev.)*	Textless Unit Diffusion + Flow Matching	Fr→En (Indo-Eur.)	Textless mHuBERT → DiT Units	DiT masked length init (L_{src}) + speed ratio r_{src}/r_{tgt}	Implicit Mel context X_{src} in OT-CFM (strips F_0)	In-context OT-CFM + WavLM (SIM = 0.266)	None (audio duration & syllable speed only)	CVSS-C Fr-En (207k train, 14,759 test, 264 h) + LRS3 fine-tuning

BLASER 2.0) while maintaining non-inferior on-screen isochrony (SO, SLC_0.2) and lip synchrony (LSE-D/C)?

- **RQ2 (Articulation-Gated Emotional Flow Matching):** *How can source-extracted speaker embeddings and simplex-bounded emotional prosody priors be injected into a single cross-lingual OT-CFM decoder using a combined consonant and CTC-entropy gate $g_{art}(\tilde{\mathcal{H}}_{CTC}, M_{cons})$ to mitigate both the −1.07 ASR-BLEU loss of post-hoc voice conversion [8] and the consonant WER inflation of high-intensity emotional guidance (12.12% WER on Sad [6])?*

- **RQ3 (Cross-Attention Stress & Soft Lip-Energy Alignment Under SVO-to-SOV & Pause Turns):** *How effectively do multi-head cross-attention stress projection and band-masked soft lip-energy attention (X-LPA) preserve word-level pitch/energy correlations ($\Delta F_0, \Delta E$ [2]), pause-stratified SO (test101 [5]), and native-listener MUSHRA naturalness [1] across multi-pause and SVO-to-SOV (En-Hi/En-Mr) utterances?*

Falsifiable Directional Hypotheses (Untested — Proposed for Empirical Validation):

- **Hypothesis H1:** Conditioning translation length intervals $\mathcal{I}_{tgt}(\bar{M}_{vis})$ (Eq. 2) and soft-upsampled SyncNet timing refinement (Eq. 5) on segment mouth visibility $\bar{M}_{vis}$ is hypothesized to yield a statistically significant ($p_{\text{Holm}} < 0.05$) increase in off-screen ($\bar{M}_{vis} < 0.2$) SacreBLEU and COMET-QE compared to uniform length-forced baselines (Txtd2PhnD [5], ungated AVS2S [7], and Ablation A1), while maintaining non-inferior on-screen ($\bar{M}_{vis} \geq 0.5$) Speech Overlap (ΔSO > -0.02) and SyncNet LSE-D (ΔLSE-D $< +0.20$) under Two One-Sided Tests (TOST).

- **Hypothesis H2:** During high-intensity emotional synthesis (specifically Sad and Angry states), modulating the OT-CFM emotional guidance field by the articulation gate $g_{art}(\tilde{\mathcal{H}}_{CTC}, M_{cons}) = \exp(-\kappa_1 \tilde{\mathcal{H}}_{CTC} - \kappa_2 M_{cons})$ (Eq. 6) is hypothesized to significantly reduce Whisper WER ($p_{\text{Holm}} < 0.05$) compared to un-damped PNGM ($\kappa_1 = \kappa_2 = 0$ [6]) and achieve higher ASR-BLEU than cascaded post-hoc voice con-

TABLE II
VERIFIED QUANTITATIVE RESULTS, STATISTICAL FINDINGS, AND DOCUMENTED FAILURE MODES ACROSS THE EIGHT PRIMARY STUDIES

Study	Verified Primary Quantitative Results (Exact Source PDF Values)	Author-Reported Limitations	Audit-Verified Failure Modes & Design Implications
Federico et al. [1]	**MUSHRA Mixed-Effects ($N = 14$ raters, 24 clips, 657 ratings):** Baseline A intercept: 46.81 (non-Italian) / 47.22 (native Italian). A→B (adding MT+PA): +1.14 (non-Italian, NS) vs. $-\mathbf{10.93}$ ($p < 0.01$, native Italian). B→C (+Audio Rendering): $+\mathbf{10.34}$ ($p < 0.05$, non-Italian) vs. +1.05 (NS, native Italian).	Character-length PA and Mel spline stretching induce speaking-rate disfluencies for native listeners; uses 2 fixed TTS voices and manual transcripts.	Non-native listeners rate acoustic ambience +10.34 pts higher while overlooking speaking-rate distortion (−10.93 pts for natives); mandates native-listener stratification.
Brannon et al. [2]	**Timing (42,850 lines):** Duration $r = 0.877$; mean overlap 0.658 (0.684 on-screen vs. 0.662 off-screen, $p = 0.003$). Char ratio vs. overlap $r = 0.279$ ($r^2 = 0.078$; > 69% violate ±10%). **Rate Std:** Es 1.25 vs. 1.47 w/s, De 1.26 vs. 1.46 w/s ($p < 10^{-10}$). **Visemes:** 1.613× on-screen vs. 1.463× off-screen ($p = 0.017$); exact match 12.4%. **Adequacy (3,617 lines):** On vs. off COMET-QE $t = 0.936$, $p = 0.349$ (NS). **Word Δr ($p < 10^{-6}$):** F_0 +0.120, energy +0.099, rate +0.125.	Proprietary Prime Video corpus (non-releasable); limited to En-De and En-Es; viseme boundaries derived from audio MFA rather than video frames.	Proves character isometry fails ($r^2 = 0.078$), human dubbers preserve steady speaking tempo and semantic adequacy on-screen, and exact viseme matching (12.4%) is a soft constraint.
Waibel et al. [3]	**ASR/MT:** WER = 2.4% (Libri) / 3.9% (TED-LIUM); En-De BLEU = 29.7. **CSS10 TTS MOS ($N = 8$):** Unemphasized naturalness 3.59 ± 0.28, intelligibility 4.69 ± 0.09; adding SSML emphasis yields 3.75 ± 0.28 emphasis perceptibility but **drops naturalness by** −0.30 **MOS to** 3.29 ± 0.16. **cGAN:** LRS2 LSE-D = 6.98, LSE-C = 6.93, FID = 8.86 (vs. Wav2Lip 6.46/7.78/4.44); 262-clip set LSE-D = 8.11 (En) / 7.90 (De).	Heuristic SSML emphasis degrades TTS naturalness; VQMIVC degrades on long utterances; cGAN blurs beards and has worse FID than Wav2Lip; zero duration control.	Altering facial pixels introduces visual artifacts and likeness risks [7] and cannot resolve utterance-level duration mismatches when translations exceed clip length.
Wu et al. [4]	**CVSS/MuST-C (BLEU / SLC_0.2):** Es-En 39.60/45.31%, De-En 36.64/44.83%, Zh-En 13.77/57.03%, En-Zh 13.95/**41.96**% (vs. 24.09% char-baseline [1] & 13.82% Trans.). **1h En-Zh Film Set (42 clips, 8 judges):** BLEU = 7.31, SLC_0.4 = 79.17%, SLC_0.2 = **66.67**% (vs. 27.25% Trans., 38.16% Base.); Quality MOS 3.98, Isochronous 3.41, Naturalness 3.60.	Training/inference duration mismatch ($T_t \neq T_s$ in training); reference BLEU penalizes length-adapted translations; no emotion or visual lip sync.	Even with duration PE, 33.33% of film sentences and 42.97%–58.04% of CVSS/MuST-C sentences violate ±20% duration compliance before TTS vowel scaling.
Chronopoulou et al. [5]	**CoVoST-2 De-En (15,413 test):** StdMT 35.6 BLEU, 0.53 SO; IsoMT [15] 35.3 BLEU, 0.53 SO; Txt2Phn 33.3 BLEU, 0.52 SO; Txtd2PhnD w/o noise **30.6** BLEU (−14.0% rel.), **0.92** SO; w/ 0.1 noise **32.4** BLEU (−9.0% rel.), **0.81** SO. **test91:** 38.6/0.54 → 27.2/0.84 (no noise) or 31.7/0.59 (0.1 noise). **test101:** 28.7/0.32 → 23.8/0.70 (−17.1% rel. BLEU) or 27.2/0.62.	Only 12.5% of training sentences have internal pauses (≥ 300 ms); phoneme-to-word detokenization adds homophone noise; promised human eval omitted in Sec. 6.	Quantifies the exact BLEU penalty (−9.0% to −17.1% relative) of enforcing uniform phoneme duration constraints without visibility-based relaxation.
Cong et al. [6]	**Chem (Set 1.0 / 2.0):** LSE-C = 8.11/8.09, LSE-D = 6.92/6.96, WER = 11.72%/12.81%, SECS = 90.62/85.06, MCD-SL = 5.87/6.22 (w/o PE: WER = 53.36%). **GRID (1.0/2.0):** LSE-D = 6.82/6.89, SECS = 92.22/86.02, WER = 18.53%/19.75%. **Dub 3.0 Zero-Shot:** LSE-D = 6.65, WER = 14.03%, MOS-S = 4.07, MOS-N = 4.05. **Emotion ($\alpha = 3.5$):** Sad raises WER to **12.12**% vs. 11.32% Neutral / 10.22% Happy.	Strictly monolingual English V2C (no translation); requires manual user input $\{c, \alpha, \beta\}$; Sad guidance degrades consonant articulation (WER).	MAS/DLCL requires a same-language ground-truth alignment matrix $M^{gt}_{lip,pho}$ that does not exist cross-lingually; Emo-IS is scored by the same classifier ψ used in PNGM.
Goncalves et al. [7]	**LRS3 (412 clips, 4 pairs):** Reduces avg LSE-D by **9.2**% to **10.67** vs. AV2AV-Speech 11.75 (GT 6.88). Per-pair LSE-D / LSE-C: Es 10.68/2.45, Pt 10.12/3.43, It 10.89/2.97, Fr 10.98/2.46; PESQ = 1.044–1.052, ASR-BLEU = 21.34–35.18. **LLM Viseme Paraphrasing (En-Es):** Improves LSE-D 10.68 → 10.18 **but drops ASR-BLEU by** −3.00 **pts** (35.18 → 32.18) and BLASER (0.797 → 0.766).	Omits speaker voice cloning and emotion; short LRS3 clips (2.32 s avg); no human evaluation; intrusive PESQ (1.04) invalid on translated waveforms.	Proves acoustic duration adjustment ($\mathcal{L}_{sync} + 10\mathcal{L}_{dur}$) preserves ASR-BLEU whereas lexical viseme paraphrasing severely degrades translation adequacy (−3.00 pts).
Choi et al. [8]	**CVSS-C Fr-En (14,759 test):** ASR-BLEU = **24.16** (single: 23.88; w/o speed adapt.: 22.10), BLASER 2.0 = **3.839**, SIM = **0.266** (vs. 0.036 S2UT), DNSMOS = 3.693. **Timing:** DC@0.2 = **100.00**%, SC@0.2 = **71.93**%, S. Corr = **0.614** (vs. 0.355 VideoDubber PE). **MOS ($N = 15$):** Nat. 3.37, Trans. 3.80, Speed 3.98. **CosyVoice VC Ablation:** Raises SIM to 0.315 **but drops ASR-BLEU by** −1.07 **pts** (24.16 → 23.09).	Audio-only S2ST (ignores video and lip sync); mHuBERT quantization strips pitch contours and explicit emotion; tested on Fr-En sentence-level clips only.	Proves in-context OT-CFM speaker conditioning avoids the −1.07 ASR-BLEU penalty of cascaded VC, while highlighting the need for visual and emotional conditioning.

version [8], while maintaining non-inferior out-of-domain emotion intensity (Emo-IS within $\delta_{\text{IS}} = 0.03$ via held-out emotion2vec-plus-large) and speaker similarity (WavLM-SIM within $\delta_{\text{SIM}} = 0.02$).

Table III
Audit and Validation Matrix of Retained Research Gaps Against Primary ([1]– [8]) and External ([9]– [16]) Literature

Gap ID & Status	Closest Prior Work & What It Solves	Why Direct Combination Fails (Verified Technical Bottleneck)	Counterexamples Checked	Precise Difference & Required Evidence
G1: Cross-Lingual Lip-Energy & Auto-Emotion Unification *(Supported Gap)*	Cross-lingual duration S2ST [4, 8, 11]; AV unit sync [7, 12]; monolingual V2C [6, 13, 14, 16].	EmoDubber's [6] MAS/DLCL requires a same-language MFA binary matrix $M^{gt}_{lip,pho}$ invalid cross-lingually; PNGM requires manual user input $\{c, \alpha, \beta\}$.	AV2AV [12] (edits face, no emotion); TransVIP [11] (no lip sync); VECL-TTS [10] (TTS only).	Band-masked soft lip-energy attention (X-LPA, Eq. 4) + simplex-bounded auto emotion priors (Eq. 1); test via Ablation A2.
G2: Visibility- & Articulation-Gated Control *(Supported Gap)*	Human on/off-screen study [2]; global duration noise σ [5]; uniform $\mathcal{L}_{sync}$ [7] & PNGM [6].	Uniform duration forcing drops BLEU by 9–17.1% [5]; viseme paraphrasing drops ASR-BLEU −3.00 pts [7]; Sad guidance raises WER to 12.12% [6].	None of [1]– [16] gate duration slack on $M_{vis}(t)$ or dampen OT-CFM emotion on consonants.	Visibility budget $\mathcal{I}_{tgt}(\bar{M}_{vis})$ (Eq. 2) + articulation gate $g_{art}(\tilde{\mathcal{H}}_{CTC}, M_{cons})$ (Eq. 6); test via Ablations A1 & A4.
G3: Non-Circular Stratified Evaluation *(Supported Gap)*	Native vs. non-native MUSHRA [1]; on/off-screen COMET-QE [2]; test91/101 [5].	Circular Emo-IS evaluation using guidance classifier ψ [6]; invalid intrusive PESQ (1.044 [7]); unstratified listener cohorts [3, 6].	Prior studies evaluate isolated axes without joint visibility, pause, and native-listener stratification.	Held-out emotion2vec-plus-large + DNSMOS + lme4 native MUSHRA across stratified benchmarks.
G4-Exploratory: SVO-to-SOV Adaptation *(Exploratory Scope)*	VideoDubber [4] (En-Zh SVO-to-SVO); DubWise [9] & VECL-TTS [10] (En-Hi TTS).	SVO-to-SOV clause inversion reverses verb/object positions relative to source lip gestures, breaking monotonic prosody mapping.	DubWise [9] and VECL-TTS [10] synthesize from pre-translated Hindi text rather than S2ST.	Cross-attention stress projection $m^*(i) = \arg\max_m \alpha_{im}$ + local X-LPA envelope matching (H3).

- **Hypothesis H3:** On pause-heavy turns (CoVoST-2 test101 [5]) and SVO-to-SOV pairs (En-Hi, En-Mr), projecting word stress Z-scores via NMT cross-attention alignments and aligning local syllable energy envelopes via band-masked X-LPA (Eq. 4) is hypothesized to achieve significantly higher FastAlign word-pair $\Delta F_0, \Delta E$ Pearson correlation r (Fisher's z-test $p_{\text{Holm}} < 0.05$), higher test101 SO, and higher native-listener MUSHRA scores than monotonic position-mapped baselines.

VI. Proposed CineDub AI Conceptual Framework

A. Architectural Overview and Path A vs. Path B Compatibility

CineDub AI is organized into four stages comprising eight components (Fig. 1). A critical architectural audit confirms that text-supervised phoneme translation and textless discrete-unit translation require distinct models, vocabularies, and training corpora. Accordingly, we define **Path A (Text-Supervised Phoneme Configuration)** as the primary architecture of CineDub AI, and specify **Path B (Textless Discrete-Unit Configuration)** as a modular alternative for unwritten languages that replaces the text NMT and IPA phoneme vocabulary in Components 3 and 5 with mHuBERT units ($K = 1000$) and a masked Diffusion Transformer (DiT) [8] trained separately.

B. Component-by-Component Specification (Stages I–IV)

- **Component 1 (Acoustic Scene Decomposition — Stage I, Established [1, 5]):** Separates input audio A_s (16 kHz) via U-Net / Hybrid Demucs masking into clean foreground speech S_s and background track B_s, segments phrases around internal pauses ($\geq$ 300 ms) via Silero VAD and MFA into intervals $T_s = \{(t_{\text{start},k}, t_{\text{end},k})\}$, and estimates blind reverberation time RT_{60} via MLE to synthesize room impulse response R_s [1].

- **Component 2 (Lip-ROI & Continuous Visibility Gate — Stage I, Adapted [2, 6, 7]):** Detects 68-point 3D facial landmarks via RetinaFace/FAN to crop a 96×96 mouth ROI at 25 FPS, extracting visual features $E_{lip} \in \mathbb{R}^{F \times d_v}$ via frozen visual AV-HuBERT [7] and normalized inner-lip opening area $a^{lip}_f \in [0, 1]$ [2]. Computes frame-level mouth visibility $M^{(f)}_{vis} \in [0, 1]$ from face detection confidence and head yaw ($|\theta_{\text{yaw}}| \leq 60°$), yielding segment-mean visibility $\bar{M}_{vis} = \frac{1}{F_k} \sum_{f \in T_k} M^{(f)}_{vis}$.

- **Component 3 (Linguistic & Prosodic Stress Front-End — Stage II, Established [2, 3, 8]):** In **Path A (Primary)**, a Conformer-LSTM ASR ensemble [3] transcribes S_s into punctuated tokens $X_s = (x_1, \ldots, x_M)$ and computes word-level prosodic stress Z-scores $E_w = (z_1, \ldots, z_M)$ from line-normalized log-F_0 and log-energy [2]. In **Path B (Modular Alternative)**, mHuBERT + k-means ($K = 1000$) extracts deduplicated units U_s and implicit speaking speed $r_{src} = \hat{L}_{src}/L_{src}$ [8].

- **Component 4 (Speaker Timbre & Simplex-Bounded Emotion Priors — Stage II, Adapted/Novel [2, 6, 8]):** Extracts unit-normalized speaker embedding $\text{spk}_{src} \in \mathbb{R}^{d_s}$ via WavLM-Large [8] / GE2E [6] and C-class emotion probabilities $l_\psi(S_s) = [l_0, \ldots, l_{C-1}]^\top \in \Delta^{C-1}$ ($C = 7$) via the pre-trained 6-layer 1D-CNN Emobox classifier ψ [6], computing bounded guidance scales $(\hat{\alpha}, \hat{\beta})$ via Eq. (1).

- **Component 5 (Visibility-Slack Duration NMT / Unit DiT — Stage III, Novel Adaptation**

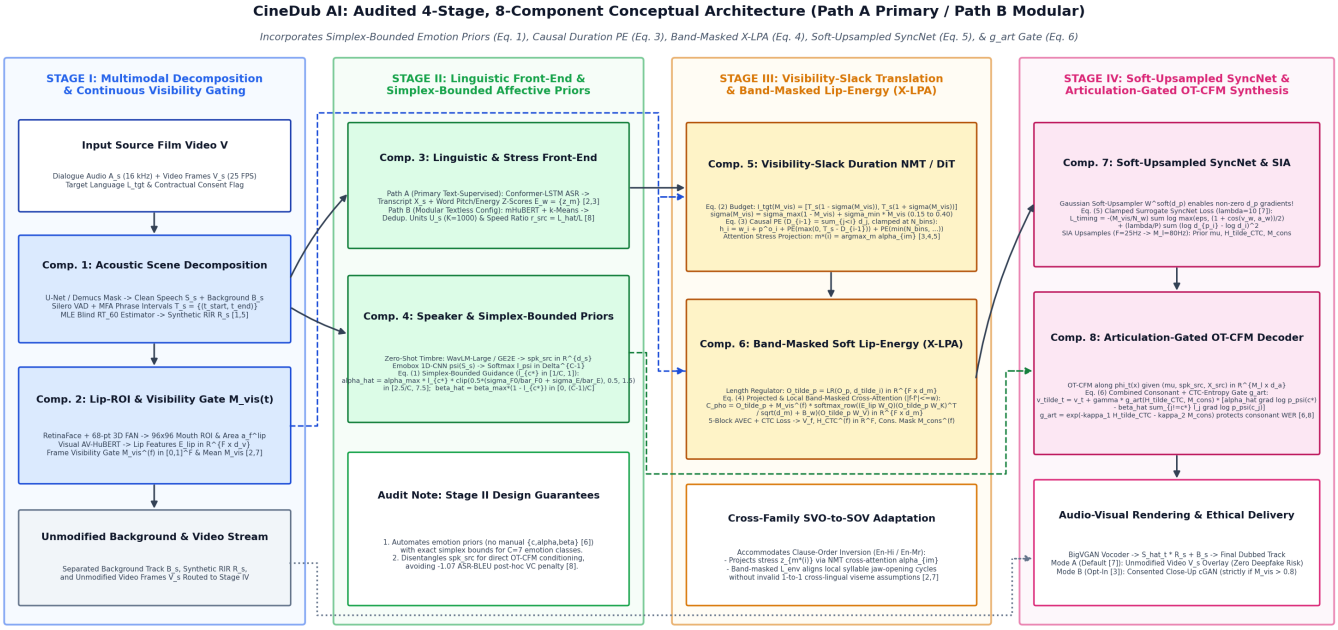


Fig. 1. Audited four-stage, eight-component **CineDub AI** conceptual architecture. Inter-stage connections are routed along dedicated channels, distinguishing the primary text-supervised phoneme path (Path A) from the alternative modular textless unit configuration (Path B) and illustrating Equations (1)–(6).

[3–5, 8]): In **Path A**, translates X_s into target phonemes $P_{tgt} = (p_1, \ldots, p_P)$ and discrete phoneme durations $d_i \in \mathbb{Z}^+$ (including inter-word pause tokens [P]) using visibility-relaxed duration intervals $\mathcal{I}_{tgt}(M_{vis})$ (Eq. 2) and causal duration positional encodings (Eq. 3), while projecting source stress z_m to target phoneme i via multi-head cross-attention $m^*(i) = \arg\max_m \frac{1}{H}\sum_{h=1}^{H} \alpha_{im}^{(h)}$ [3]. In **Path B**, DiT translates U_s into target units U_t initialized at length $L_{tgt} \in \mathcal{I}_{tgt}(\bar{M}_{vis})$ with pace scaling r_{src}/r_{tgt} [8].

- **Component 6 (Band-Masked Cross-Lingual Soft Lip-Energy Attention X-LPA — Stage III, Novel Proposed Module):** Expands target phoneme embeddings $O_p \in \mathbb{R}^{P\times d_m}$ to video frame rate F via Length Regulation $\tilde{O}_p = \mathrm{LR}(O_p, \tilde{d}_i) \in \mathbb{R}^{F\times d_m}$, applies local band-masked cross-attention with E_{lip} (Eq. 4), and processes the result through a 5-block AVEC Conformer under CTC loss $\mathcal{L}_{CTC}$ [6] and mouth-aperture energy envelope loss $\mathcal{L}_{env}$, outputting $V_f \in \mathbb{R}^{F\times d_m}$, frame CTC entropy $\mathcal{H}_{CTC}^{(f)} \in \mathbb{R}^F$, and deterministic phoneme-class consonant mask $M_{cons}^{(f)} \in \{0,1\}^F$.
- **Component 7 (Gaussian Soft-Upsampled SyncNet Timing & SIA — Stage IV, Adapted [6,7]):** Fine-tunes predicted durations d_{p_i} using a frozen single-pass surrogate Mel renderer $\hat{A}_t^{surr}(W^{soft}(d_p))$ with Differentiable Gaussian Soft-Upsampling under Eq. (5), then upsamples V_f, $\mathcal{H}_{CTC}^{(f)}$, and $M_{cons}^{(f)}$ from video rate F (25 Hz) to Mel length M_l ($\sim$ 80 Hz) via SIA/USL [6] to form the acoustic prior $\mu \in \mathbb{R}^{M_l\times d_a}$ and Mel-aligned gates $\tilde{\mathcal{H}}_{CTC}, M_{cons} \in \mathbb{R}^{M_l\times 1}$.
- **Component 8 (Articulation-Gated OT-CFM Decoder & Mode-Selectable Delivery — Stage IV, Novel Integration [1, 3, 6–8]):** Generates target Mel-spectrogram $\hat{M} \in \mathbb{R}^{M_l\times d_a}$ via OT-CFM ODE integration steered by the articulation-damped vector field $\tilde{v}_{t,c^*}$ (Eq. 6), vocodes $\hat{M}$ to waveform $\hat{S}_t$ via BigVGAN [6, 8], convolves with synthetic RIR R_s, and mixes with background B_s [1]. Visual output defaults to **Mode A (Unmodified Video Overlay [7])**, with **Mode B (Consented Close-Up cGAN [3])** restricted to verified actor-consent close-ups ($\bar{M}_{vis} > 0.8$).

VII. Mathematical Formulation

A. Simplex-Bounded Emotion Guidance Scales

Let $l_\psi(S_s) = [l_0, \ldots, l_{C-1}]^\top \in \Delta^{C-1}$ ($\sum_{j=0}^{C-1} l_j = 1$, $C = 7$ [6]) denote the source emotion probability vector from classifier ψ, and let $c^* = \arg\max_j l_j$. By the simplex pigeonhole principle, $l_{c^*} \in [1/C, 1]$ and $\sum_{j\neq c^*} l_j = 1 - l_{c^*} \in [0, (C-1)/C]$. To automate $\{c^*, \hat{\alpha}, \hat{\beta}\}$ using dimensionless pitch and energy standard deviation ratios $(\sigma_{F_0}/\bar{\sigma}_{F_0}, \sigma_E/\bar{\sigma}_E)$ [2], we propose the heuristic scaling:

$$\begin{aligned} \hat{\alpha} &= \alpha_{\max}\, l_{c^*}\, \mathrm{clip}\left(\frac{1}{2}\left(\frac{\sigma_{F_0}}{\bar{\sigma}_{F_0}} + \frac{\sigma_E}{\bar{\sigma}_E}\right), 0.5, 1.5\right) \\ &\in \left[0.5\alpha_{\max}/C,\ 1.5\alpha_{\max}\right], \\ \hat{\beta} &= \beta_{\max}(1 - l_{c^*}) \in \left[0,\ \beta_{\max}(C-1)/C\right] \end{aligned} \tag{1}$$

Setting unvalidated baseline ceilings $\alpha_{\max} = 5.0, \beta_{\max} = 1.0$ (within the $[0, 9]$ exploration range of [6]) yields $\hat{\alpha} \in$

$[2.5/C, 7.5]$ and $\hat{\beta} \in [0, (C-1)/C]$.

B. Visibility-Modulated Duration Interval and Causal Duration Positional Encoding

Given phrase-averaged mouth visibility $\bar{M}_{vis} \in [0,1]$ and source phrase duration $T_s > 0$ (in acoustic frames), the proposed visibility-modulated relative duration slack $\sigma(\bar{M}_{vis})$ and allowable target duration interval $\mathcal{I}_{tgt}(\bar{M}_{vis})$ are defined as:

$$\begin{aligned} \sigma(\bar{M}_{vis}) &= \sigma_{\max}(1-\bar{M}_{vis}) + \sigma_{\min}\bar{M}_{vis} \ \in [\sigma_{\min}, \sigma_{\max}], \\ \mathcal{I}_{tgt}(\bar{M}_{vis}) &= \Big[T_s\big(1-\sigma(\bar{M}_{vis})\big),\ T_s\big(1+\sigma(\bar{M}_{vis})\big)\Big] \end{aligned} \tag{2}$$

where $\sigma_{\min} = 0.15$ and $\sigma_{\max} = 0.40$ are proposed design tolerances anchored to the $p = 0.2, 0.4$ compliance thresholds of $[4, 8]$. To ensure strict causality at autoregressive decoding step i (before d_i is predicted) and prevent out-of-bounds bin indices when candidate durations use positive slack up to $\tilde{T}_{tgt}(\bar{M}_{vis}) = T_s(1+\sigma(\bar{M}_{vis}))$, let $D_{i-1} = \sum_{j=1}^{i-1} d_j$ ($D_0 = 0$) denote the causal cumulative prefix duration. Adapting VideoDubber [4], the decoder input representation $h_i \in \mathbb{R}^{d_m}$ is:

$$\begin{aligned} h_i = w_i + p_i^o + \mathrm{PE}\big(\max(0, \lfloor T_s - D_{i-1} \rfloor)\big) \\ + \mathrm{PE}\left(\min\left(N_{bins}, \left\lfloor \frac{D_{i-1}}{\tilde{T}_{tgt}(\bar{M}_{vis})} \cdot N_{bins} \right\rfloor\right)\right), \\ \mathrm{PE}_{(pos,2k)} = \sin\big(pos/10000^{2k/d_m}\big), \\ \mathrm{PE}_{(pos,2k+1)} = \cos\big(pos/10000^{2k/d_m}\big) \end{aligned} \tag{3}$$

where $N_{bins} = 100$ (extending the $N = 5$ ratio buckets of [4] to the 100-bin resolution of [5]).

C. Band-Masked Cross-Lingual Soft Lip-Energy Attention (X-LPA)

Let $\tilde{O}_p = \mathrm{LR}(O_p, \tilde{d}_i) \in \mathbb{R}^{F \times d_m}$ denote target phoneme+stress embeddings expanded to video frame length F, and let $E_{lip} \in \mathbb{R}^{F \times d_v}$ denote visual lip features. To match tensor dimensions ($d_v \to d_m$) and restrict cross-modal attention to a local temporal neighborhood $|f - f'| \le w$ ($w = 5$ frames $= \pm 200$ ms), we introduce learnable projections $W_Q \in \mathbb{R}^{d_v \times d_m}$, $W_K, W_V \in \mathbb{R}^{d_m \times d_m}$ and a band mask $(B_w)_{f,f'} = 0$ if $|f - f'| \le w$ and $-\infty$ otherwise:

$$\begin{aligned} A_{lip} &= \mathrm{softmax}_{\mathrm{row}}\left(\frac{(E_{lip}W_Q)(\tilde{O}_pW_K)^\top}{\sqrt{d_m}} + B_w\right), \\ C_{pho} &= \tilde{O}_p + M_{vis}^{(f)} \odot \big(A_{lip}\tilde{O}_pW_V\big) \end{aligned} \tag{4}$$

where $M_{vis}^{(f)} \in [0,1]^{F \times 1}$ is broadcast across d_m channels via $\odot$. Note that X-LPA aligns local syllable energy and jaw-opening envelopes ($\mathcal{L}_{env} = \frac{1}{F}\sum_{f=1}^{F} M_{vis}^{(f)}(\hat{e}_f - a_f^{lip})^2$) rather than cross-lingual phonetic identity.

D. Gaussian Soft-Upsampled Surrogate SyncNet Refinement

Because integer rounding in standard Length Regulation has zero derivative almost everywhere, fine-tuning duration predictions $d_p = (d_{p_1}, \ldots, d_{p_P}) \in \mathbb{R}_{>0}^P$ requires a differentiable duration-to-frame mapping. Inside the frozen single-pass surrogate Mel renderer $\hat{A}_t^{surr}$, we employ Differentiable Gaussian Soft-Upsampling with phoneme center positions $c_i = \sum_{j=1}^{i-1} d_{p_j} + \frac{1}{2}d_{p_i}$ and bandwidth σ_g:

$$\begin{aligned} W_{m,i}^{soft}(d_p) &= \frac{\exp\big(-(m-c_i)^2/(2\sigma_g^2)\big)}{\sum_{k=1}^{P}\exp\big(-(m-c_k)^2/(2\sigma_g^2)\big)}, \\ s_w(d_p) &= \max\left(\epsilon,\ \frac{1+\cos\big(v_w(V_s),\ a_w(\hat{A}_t^{surr}(W^{soft}(d_p)))\big)}{2}\right), \\ \mathcal{L}_{timing}(d_p) &= -\frac{\bar{M}_{vis}}{N_w}\sum_{w=1}^{N_w}\log s_w(d_p) \\ &\quad + \frac{\lambda}{P}\sum_{i=1}^{P}(\log d_{p_i} - \log d_i)^2 \end{aligned} \tag{5}$$

where $\epsilon = 10^{-6}$ guarantees a strictly positive argument inside $\log(\cdot)$ and $\lambda = 10$ regularizes log-durations toward initial predictions d_i [7].

E. Consonant- and CTC-Entropy-Damped OT-CFM Vector Field

Let $\phi_t(x) = (1-(1-\sigma_{min})t)x_0 + tM \in \mathbb{R}^{M_l \times d_a}$ denote the OT-CFM linear interpolation path [6, 8]. Because CTC frame posteriors are blank-dominated and peaky (making it an unvalidated assumption that $\tilde{\mathcal{H}}_{CTC} = \mathrm{Upsample}_{F \to M_l}(\mathcal{H}_{CTC}^{(f)}) \in [0, \ln|\mathcal{V}_{pho}|]^{M_l \times 1}$ alone captures all consonant spans), we combine $\tilde{\mathcal{H}}_{CTC}$ with the deterministic binary consonant mask $M_{cons} \in \{0,1\}^{M_l \times 1}$ (where $M_{cons}(m) = 1$ if Mel frame m maps to a consonant $p_i \in \mathcal{V}_{cons}$ under Length Regulation):

$$\begin{aligned} \tilde{v}_{t,c^*} &= v_t\big(\phi_t(x) \mid \mu, \mathrm{spk}_{src}, X_{src}; \theta\big) \\ &\quad + \gamma\, g_{art} \odot \Big(\hat{\alpha}\nabla_{\phi_t}\log p_\psi(c^* \mid \phi_t) \\ &\quad - \hat{\beta}\textstyle\sum_{j \neq c^*} l_j \nabla_{\phi_t}\log p_\psi(c_j \mid \phi_t)\Big), \\ g_{art} &= \exp\big(-\kappa_1\tilde{\mathcal{H}}_{CTC} - \kappa_2 M_{cons}\big) \in \big[|\mathcal{V}_{pho}|^{-\kappa_1}e^{-\kappa_2}, 1\big]^{M_l \times 1} \end{aligned} \tag{6}$$

where $\kappa_1, \kappa_2 \ge 0$ are damping hyperparameters to be ablated independently and jointly.

VIII. Research Methodology and Expected Evaluation Protocol

A. Dataset Licensing, Access Restrictions, and Leakage Controls

As shown in Fig. 2, no empirical experiments are claimed as completed in this conceptual study. Future empirical validation is designed across four complementary benchmarks with verified licensing and speaker-disjoint splits:

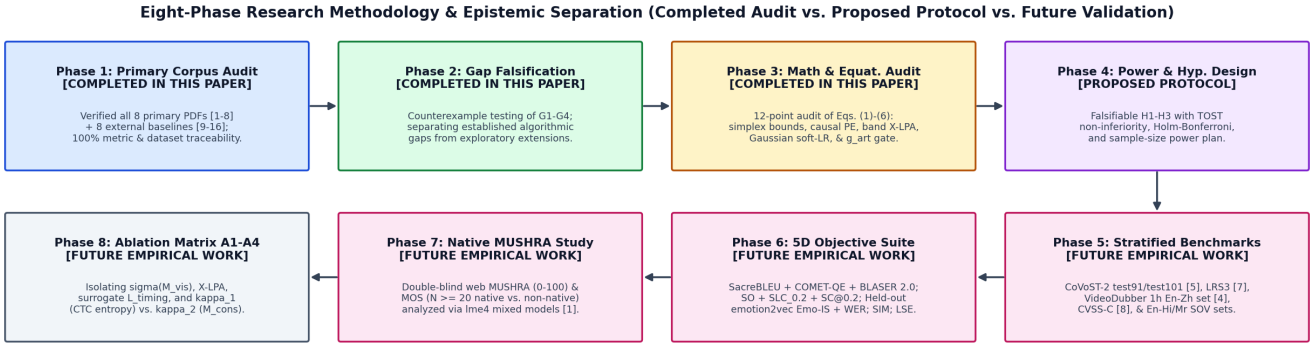


Fig. 2. Eight-phase methodological and validation workflow explicitly separating completed literature/mathematical audit phases (Phases 1–3), proposed statistical protocol design (Phase 4), and future empirical benchmarking (Phases 5–8).

1. **Pause-Stratified Isochrony Set (CoVoST-2 De-En + test91/test101 [5]):** CoVoST-2 annotations carry a CC0-1.0 license, and the on-camera test91 ($N = 91$, zero pauses) and test101 ($N = 101$, ≥ 1 pause ≥ 300 ms) scripts/code are released under Apache-2.0 with zero speaker overlap against training data.

2. **Cross-Family & S2ST Set (CVSS-C Fr-En [8], MuST-C En-Zh, & 1h Film Set [4]):** CVSS-C ($14,759$ test pairs) is licensed under CC-BY-NC-4.0 and MuST-C under CC-BY-NC-ND-4.0. Note that the 42-clip (892-sentence) En-Zh film test set [4] comprises copyrighted movie excerpts used strictly for zero-shot academic evaluation, while the Heroes corpus [2] is proprietary and non-releasable.

3. **Audio-Visual & Emotional Sets (LRS3 [7], Chem, GRID, & Dub 3.0 [6]):** LRS3 (412 test clips across En $\rightarrow$ Es, Pt, It, Fr, stratified into off-screen $\bar{M}_{vis} < 0.2$ and on-screen $\bar{M}_{vis} \geq 0.5$) and LRS2 require institutional academic EULAs (non-commercial research only). Under Dub 3.0 [6], unseen Chem speaker references are paired with GRID videos to prevent speaker leakage.

4. **Exploratory SVO-to-SOV Set (En-Hi / En-Mr [9, 10]):** Evaluates clause-order inversion across ≥ 200 parallel Indo-Aryan utterances.

B. Baselines, Ablation Matrix, and Statistical Power Plan

Five external baselines—(B1) **Federico et al.** [1], (B2) **Face-Dubbing++** [3], (B3) **VideoDubber + AdaSpeech 4** [4] / **Txtd2PhnD** [5], (B4) **Direct AVS2S + SyncNet** [7], and (B5) **Dub-S2ST** [8] + undamped **EmoDubber FUEC** [6] (Table V)—will be compared alongside four internal ablations: (A1) uniform duration forcing ($\sigma(\bar{M}_{vis}) \equiv \sigma_{\min}$); (A2) removing X-LPA ($M_{vis}^{(f)} \equiv 0$); (A3) removing surrogate SyncNet refinement ($\bar{M}_{vis} = 0$ in Eq. 5); and (A4) ablating the articulation gate ($\kappa_1 = 0, \kappa_2 = 0$ vs. $\kappa_1 > 0, \kappa_2 = 0$ vs. $\kappa_1 = 0, \kappa_2 > 0$ in Eq. 6).

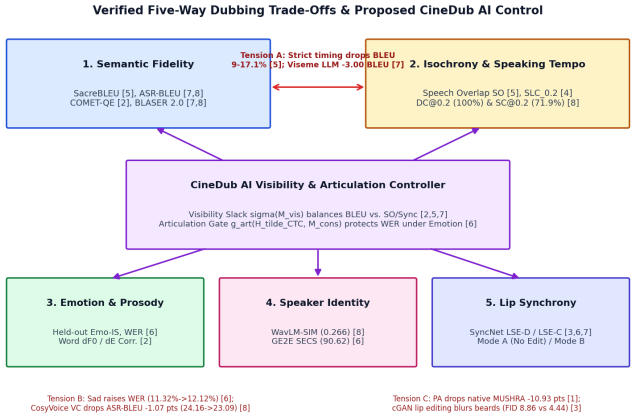


Fig. 3. Verified five-way dubbing trade-offs and proposed Pareto control via visibility slack $\sigma(M_{vis})$ and articulation gate $g_{art}(\tilde{\mathcal{H}}_{CTC}, M_{cons})$.

Statistical Testing & Power Analysis: Objective comparisons will use paired bootstrap resampling ($10,000$ iterations, 95% BCa CIs) for SacreBLEU/ASR-BLEU, two-sided paired Wilcoxon signed-rank tests for COMET-QE, SO, WER, Emo-IS, SIM, and LSE-D, Fisher's z-test for word-pair $\Delta F_0/\Delta E$ correlations [2], and Two One-Sided Tests (TOST) for non-inferiority claims, with Holm-Bonferroni step-down correction ($\alpha_{\text{FWER}} = 0.05$) and Cohen's d_z / rank-biserial r_{rb} effect sizes. At $1 - \beta = 0.80$ power and $\alpha = 0.05$, detecting $d_z = 0.25$ requires $N \geq 128$ paired segments (satisfied by the 412-clip LRS3 and 892-sentence film sets). Human perceptual evaluation will recruit $N \geq 20$ **native target-language listeners** and $N \geq 20$ **non-native listeners** rating ≥ 24 clips on a double-blind web MUSHRA interface (0–100 scale, inter-rater reliability via intraclass correlation ICC$(2, k)$) and 5-point MOS scales, analyzed via lme4 linear mixed-effects models (Score Condition * NativeGroup + (1 + Condition | Listener) + (1 | Clip)) [1].

TABLE IV
MAPPING OF RESEARCH QUESTIONS AND HYPOTHESES TO EVALUATION METRICS, STATISTICAL TESTS, AND VERIFIED MITIGATIONS

RQ & Hyp.	Evaluation Dimension	Objective Automatic Metrics & Statistical Models	Human Perceptual Protocol & Reliability	Audit-Verified Metric Limitations & Mitigation Strategy
RQ1, H1	**1. Semantic Fidelity & Adequacy**	Reference SacreBLEU [5] & ASR-BLEU (Whisper-V3 [6] / wav2vec 2.0 [8]); Reference-free COMET-QE [2] & BLASER 2.0 [7, 8]. Paired bootstrap & Wilcoxon + Holm-Bonferroni.	Bilingual adequacy & fluency MOS (1–5 [4, 8]) stratified by on-screen ($\bar{M}_{vis} \geq 0.5$) vs. off-screen ($\bar{M}_{vis} < 0.2$) [2].	*Limitation:* Reference BLEU penalizes valid length-adapted paraphrases [4], and phoneme detokenization adds homophone noise [5]. *Mitigation:* Pair ASR-BLEU with reference-free COMET-QE [2] and BLASER 2.0 [8].
RQ1, RQ3, H1, H3	**2. Isochrony, Pause Timing & Tempo**	Speech Overlap (SO) $= 1 - \lvert T_s - T_t \rvert / T_s$ [5]; SLC_p/DC@p ($p = 0.2, 0.4$) [4, 8]; Syllable SC@0.2 & S. Corr [8]; phrase IoU [2]. TOST non-inferiority ($\delta_{SO} = 0.02$).	Temporal & pause synchronicity MOS (1–5 [4, 8]) on single-sentence test91 and pause-heavy test101 [5].	*Limitation:* Utterance-level DC@0.2 and SLC_0.2 mask internal pause misalignment [5], and duration compliance does not prove lip sync. *Mitigation:* Report phrase IoU [2], S. Corr [8], and LSE-D/C separately.
RQ2, RQ3, H2, H3	**3. Emotion, Stress & Intelligibility**	Independent held-out emotion2vec-plus-large Emo-IS & macro-F1; FastAlign word-pair $\Delta F_0, \Delta E$ Pearson r (Fisher's z-test [2]); Whisper WER [6] & DNSMOS [8].	Emotion similarity, emphasis perceptibility [3], and naturalness MUSHRA (0–100 [1]) & MOS-N [6]; ICC$(2, k)$ reliability.	*Limitation:* Scoring Emo-IS with the guidance classifier ψ [6] causes circular in-domain bias; intrusive PESQ (1.044 [7]) fails on translated speech. *Mitigation:* Mandate held-out emotion2vec-plus-large + DNSMOS [8].
RQ2, H2	**4. Speaker Identity & Acoustic Scene**	WavLM-Large SIM [8] and GE2E SECS [6] cosine similarity; MCD [6]; TOST non-inferiority ($\delta_{SIM} = 0.02$).	Zero-shot speaker similarity MOS-S (1–5 [6]) + MUSHRA realism with/without RIR & background mix [1].	*Limitation:* Separate VC inflates SIM (0.315) while degrading ASR-BLEU (−1.07 pts [8]). *Mitigation:* Jointly evaluate (SIM, WER, ASR-BLEU) on speaker-disjoint Dub 3.0 [6].
RQ1, RQ3, H1	**5. Audio-Visual Lip Synchrony & Ethics**	SyncNet LSE-D (↓) and LSE-C (↑) [3, 6, 7] stratified by $\bar{M}_{vis}$ (TOST $\delta_{LSE\text{-}D} = 0.20$); FID (↓) [3] when optional Mode B is active.	Native ($N \geq 20$) vs. non-native ($N \geq 20$) listener video MUSHRA analyzed via lme4 mixed-effects models [1].	*Limitation:* Non-native listeners inflate MUSHRA by +10.34 pts while missing PA disfluencies (−10.93 pts for natives [1]). *Mitigation:* Mandate native target-language listener stratification [1].

TABLE V
ARCHITECTURAL COMPARISON OF REPRESENTATIVE EXISTING SYSTEMS AND THE PROPOSED CINEDUB AI CONCEPTUAL FRAMEWORK

System / Study	Cross-Lingual	Duration / Speed Control	Emotion & Emphasis	Speaker Timbre Cloning	Lip-Sync Mechanism	Verified Principal Limitation
Federico et al. [1] (IWSLT 2020)	Yes (En-It)	Verbosity token + PA + Mel spline	Phrase pauses + MLE RT_{60} RIR	None (2 fixed TTS voices)	None (audio timing only)	Post-hoc PA stretching degrades native Italian naturalness by −10.93 MUSHRA points ($p < 0.01$).
Face-Dubbing++ [3] (Waibel et al., 2022/23)	Yes (En-De)	None (unconstrained TTS length)	Binary ASR-MT-SSML word stress	Post-hoc VQMIVC conversion	6-channel lower-face cGAN	Zero duration control; SSML drops naturalness (−0.30 MOS); cGAN blurs beards (FID 8.86 vs. 4.44).
VideoDubber [4] / Txtd2PhnD [5] (2023)	Yes (4 pairs / De-En)	Abs/Rel PE ($N = 5$) + [P] / 100-bin Phn	Word/phrase pauses only	AdaSpeech 4 [4] / None [5]	None (audio duration only)	No lip-sync or emotion transfer; 33.3–58.0% of clips violate SLC_0.2 [4]; strict timing drops BLEU 9–17.1% [5].
EmoDubber [6] (Cong et al., CVPR 2025)	**No** (Mono-lingual En)	Frame-locked DLCL + MAS	FUEC + PNGM (manual user input)	SIA + USL (SECS = 90.62)	Audio-to-lip LPA (LSE-D = 6.92)	Cannot translate across languages; requires same-language MFA matrix $M^{gt}_{lip,pho}$; Sad raises WER to 12.12%.
Goncalves et al. [7] (ICASSP 2025)	Yes (4 pairs)	Unit Duration Predictor ($\mathcal{L}_{dur}$)	None	None (generic unit vocoder)	SyncNet $\mathcal{L}_{sync}$ on unmodified video	Omits voice cloning and emotion; LLM viseme paraphrasing drops ASR-BLEU by −3.00 pts (35.18 → 32.18).
Dub-S2ST [8] (Choi et al., EMNLP 2025)	Yes (Fr-En)	DiT masked length init (DC@0.2 = 100%)	Implicit Mel context X_{src} only	In-context OT-CFM (SIM = 0.266)	None (audio-only S2ST)	Strips pitch during mHuBERT quantization; ignores visual lip movements; separate VC drops ASR-BLEU −1.07 pts.
CineDub AI *(Proposed Framework)*	**Yes** (Path A / Path B)	**Visibility-slack** $\mathcal{I}_{tgt}(\bar{M}_{vis})$ + **Causal PE**	**Bounded** $\hat{\alpha}, \hat{\beta}$ + E_w + g_{art} **gate**	**Integrated SIA + OT-CFM (no VC)**	**Band X-LPA + Soft-LR** $\mathcal{L}_{timing}$	*Conceptual framework*; requires empirical multi-corpus training, hyperparameter tuning, and human validation.

IX. DISCUSSION

A. Theoretical Resolution of Multi-Objective Dubbing Conflicts

The primary analytical insight emerging from our audit of Papers [1–8] is that automatic dubbing quality collapses

when soft, context-dependent perceptual cues are enforced as rigid, utterance-wide invariants (Fig. 3). Brannon *et al.* [2] demonstrated that professional human localization preserves semantic adequacy (COMET-QE $p = 0.349$) and steady speaking tempo ($p < 10^{-10}$) first, treats phrase overlap moderately (0.658), and satisfies exact viseme identity in only 12.4% of on-screen speech time. By conditioning the allowable target duration interval $\mathcal{I}_{tgt}(\bar{M}_{vis})$ (Eq. 2), band-masked X-LPA attention (Eq. 4), and soft-upsampled SyncNet duration refinement (Eq. 5) on continuous mouth visibility $M_{vis}(t)$, CineDub AI is designed to relax length and viseme constraints during off-screen narration or occluded shots—where uniform duration forcing otherwise costs 9.0%–17.1% relative BLEU [5]—while tightening acoustic-visual alignment during close-up shots. Similarly, our formulation hypothesizes that combining a deterministic consonant mask M_{cons} with upsampled CTC entropy $\tilde{\mathcal{H}}_{CTC}$ inside g_{art} (Eq. 6) can selectively attenuate emotional vector-field perturbations during obstruent/consonant articulation while permitting full emotional pitch/energy expression on vowel nuclei.

B. Cross-Family Linguistic Challenges (SVO-to-SOV Adaptation)

Extending automatic dubbing from European SVO languages to Indo-Aryan (En-Hi, En-Mr) and Dravidian (En-Te) SOV languages [9, 10] introduces clause-level word-order inversion and polysyllabic rate expansion. In CineDub AI, projecting word-level stress Z-scores via NMT multi-head cross-attention alignments ($m^*(i) = \arg\max_m \alpha_{im}$ [3]) and restricting X-LPA to a $\pm w$-frame local band mask B_w (Eq. 4) are designed to anchor prosodic emphasis to semantically aligned target words and align local jaw-opening energy cycles without assuming monotonic source-to-target phoneme order.

X. Limitations and Threats to Validity

As a conceptual and methodological study, CineDub AI is subject to five explicit limitations and threats to validity:

1. **Absence of Empirical Training Results:** CineDub AI has not yet been trained or empirically benchmarked; all proposed gains remain falsifiable hypotheses (H1–H3) awaiting experimental testing.

2. **Surrogate Renderer Approximation Gap:** Fine-tuning durations d_{p_i} via a lightweight Gaussian soft-upsampled surrogate renderer $\hat{A}_t^{surr}$ (Eq. 5) assumes its frame-level energy timing closely matches the final multi-step OT-CFM + BigVGAN waveform.

3. **Unvalidated CTC-Entropy & Guidance Hyperparameters:** Because CTC posteriors are peaky and blank-dominated, whether $\tilde{\mathcal{H}}_{CTC}$ adds protective value beyond the deterministic consonant mask M_{cons} in Eq. (6), as well as the optimal values of $(\alpha_{\max}, \beta_{\max}, \kappa_1, \kappa_2)$, must be established via Ablation A4.

4. **Upstream Segmentation & Overlap Error Propagation:** Errors in Demucs separation, Silero VAD pause detection, or MFA phone alignment degrade downstream timing, particularly on overlapping multi-speaker film dialogue (> 30% overlap excluded in [2]).

5. **Domain & Licensing Constraints:** Clean sentence-level corpora (CoVoST-2, CVSS-C) differ acoustically from dramatic cinema (Heroes [2] is proprietary, and LRS2/LRS3 carry non-commercial academic EULAs).

XI. Ethical Considerations

Generative audiovisual dubbing directly implicates actor likeness rights, vocal biometric cloning, and deepfake misuse [7]. Manipulating an actor's facial pixels (Mode B [3, 12]) or cloning their vocal identity (spk_{src} [6, 8]) without explicit contractual consent poses severe legal and ethical risks. CineDub AI enforces **Mode A (Integrity-Preserving Unmodified Video Overlay [7])** as its default operating mode, requires cryptographic provenance watermarking and explicit consent flags before enabling zero-shot voice cloning or Mode B facial synthesis, and restricts intended use to authorized media localization.

XII. Conclusion and Future Work

This paper presented a comprehensive technical and quantitative audit of eight primary automatic dubbing studies [1–8] and eight complementary baselines [9–16], resolving literature contradictions and formulating the **CineDub AI** conceptual framework. By mathematically coupling simplex-bounded emotion priors (Eq. 1), visibility-modulated duration intervals and causal positional encodings (Eqs. 2–3), band-masked soft lip-energy attention (X-LPA, Eq. 4), Gaussian soft-upsampled SyncNet duration refinement (Eq. 5), and consonant/CTC-entropy-damped OT-CFM synthesis (Eq. 6), CineDub AI provides a rigorous, falsifiable blueprint for multi-objective film dubbing. Future work will implement Path A and Path B in PyTorch/Fairseq, execute Ablations A1–A4 across CoVoST-2 test91/test101, LRS3, CVSS-C, and En-Hi/En-Mr benchmarks, and conduct IRB-approved native-listener MUSHRA evaluations to empirically test Hypotheses H1–H3.

Acknowledgment

The authors express their sincere gratitude to their Research Guide, **Mrs. Pranita Manjare** (Department of MCA, D. Y. Patil Institute of MCA and Management, Akurdi, Pune, Maharashtra, India), for her academic supervision, methodological guidance, and critical review throughout this research.

References

[1] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and H. Sawaf, "From speech-to-speech translation to automatic dubbing," in *Proc. 17th Int. Conf. Spoken Language Translation (IWSLT)*, Online: ACL, Jul. 2020, pp. 257–264, doi: 10.18653/v1/2020.iwslt-1.31.

[2] W. Brannon, Y. Virkar, and B. Thompson, "Dubbing in practice: A large scale study of human localization with insights for automatic dubbing," *Trans. Assoc. Comput. Linguistics (TACL)*, vol. 11, pp. 419–435, 2023, doi: 10.1162/tacl_a_00551.

[3] A. Waibel, M. Behr, F. I. Eyiokur, D. Yaman, T.-N. Nguyen, C. Mullov, M. A. Demirtas, A. Kantarci, S. Constantin, and H. K. Ekenel, "Face-Dubbing++: Lip-synchronous, voice preserving translation of videos," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSPW59220.2023.10193719. (Extended version: *arXiv:2206.04523*, Jun. 2022).

[4] Y. Wu, J. Guo, X. Tan, C. Zhang, B. Li, R. Song, L. He, S. Zhao, A. Menezes, and J. Bian, "VideoDubber: Machine translation with speech-aware length control for video dubbing," in *Proc. 37th AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 11, Jun. 2023, pp. 13772–13779, doi: 10.1609/aaai.v37i11.26613.

[5] A. Chronopoulou, B. Thompson, P. Mathur, Y. Virkar, S. M. Lakew, and M. Federico, "Jointly optimizing translations and speech timing to improve isochrony in automatic dubbing," *arXiv preprint arXiv:2302.12979*, Feb. 2023.

[6] G. Cong, J. Pan, L. Li, Y. Qi, Y. Peng, A. van den Hengel, J. Yang, and Q. Huang, "EmoDubber: Towards high quality and emotion controllable movie dubbing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 15863–15873, doi: 10.1109/CVPR52734.2025.01479.

[7] L. Goncalves, P. Mathur, X. Niu, B. Houston, C. Lavania, S. Vishnubhotla, L. Sun, and A. Ferritto, "Improving lip-synchrony in direct audio-visual speech-to-speech translation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. (Uploaded source analyzed: *arXiv:2412.16530v1*, Dec. 2024).

[8] J. Choi, J. Kim, and J. S. Chung, "Dub-S2ST: Textless speech-to-speech translation for seamless dubbing," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Nov. 2025, pp. 9341–9353, doi: 10.18653/v1/2025.findings-emnlp.524.

[9] N. Sahipjohn, A. Gudmalwar, N. Shah, P. Wasnik, and R. R. Shah, "DubWise: Video-guided speech duration control in multimodal LLM-based text-to-speech for dubbing," in *Proc. Interspeech 2024*, Sep. 2024, pp. 2960–2964, doi: 10.21437/Interspeech.2024-1700.

[10] A. Gudmalwar, N. Shah, S. Akarsh, P. Wasnik, and R. R. Shah, "VECL-TTS: Voice identity and emotional style controllable cross-lingual text-to-speech," in *Proc. Interspeech 2024*, Sep. 2024, pp. 2955–2959, doi: 10.21437/Interspeech.2024-1695.

[11] C. Le, Y. Qian, D. Wang, L. Zhou, S. Liu, X. Wang, M. Yousefi, Y. Qian, J. Li, S. Zhao, and M. Zeng, "TransVIP: Speech to speech translation system with voice and isochrony preservation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024.

[12] J. Y. Choi, S. J. Park, M. Kim, and Y. M. Ro, "AV2AV: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27325–27335, doi: 10.1109/CVPR52733.2024.02580.

[13] G. Cong, Y. Qi, L. Li, A. Beheshti, Z. Zhang, A. van den Hengel, M.-H. Yang, C. Yan, and Q. Huang, "StyleDubber: Towards multi-scale style learning for movie dubbing," in *Findings of the Association for Computational Linguistics: ACL 2024*, Aug. 2024, pp. 6767–6779, doi: 10.18653/v1/2024.findings-acl.404.

[14] G. Cong, L. Li, Y. Qi, Z.-J. Zha, Q. Wu, W. Wang, B. Jiang, M.-H. Yang, and Q. Huang, "Learning to dub movies via hierarchical prosody models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14687–14697, doi: 10.1109/CVPR52729.2023.01411.

[15] S. M. Lakew, Y. Virkar, P. Mathur, and M. Federico, "Isometric MT: Neural machine translation for automatic dubbing," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, May 2022, pp. 6242–6246, doi: 10.1109/ICASSP43922.2022.9747205.

[16] C. Hu, Q. Tian, T. Li, W. Yuping, Y. Wang, and H. Zhao, "Neural Dubber: Dubbing for videos according to scripts," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, Dec. 2021, pp. 16582–16595.