VESTA: TIMBRE-FAITHFUL AND LIP-SYNCED JOINT TALKING AUDIO–VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent joint talking audio–video generation produces realistic human videos, yet falls short on both the voice and the lips: the timbre is uncontrolled rather than specified, and the lips stay imperfectly synchronized. Existing timbre-injection methods feed the reference into the token stream, which in a coupled generator leaks time-varying content and unsettles the lips; they supervise both modalities with reconstruction losses alone, and serve every speaker with one fixed pass. We present **VESTA** (**V**ideo-audio **E**xpert-guided **S**ynchronized **T**imbre-faithful **A**vatar), which injects timbre off the time axis and supervises perception directly on the decoded face and voice, through three components. (1) **StatShift** modulates the audio branch's channel statistics (AdaLN) uniformly across time with a reference speaker embedding; carrying no time-varying signal, it controls timbre while empirically preserving the content the lips track. (2) **RewardSteer** adds the perceptual supervision *joint* generators lack: a timbre reward, and for lip-sync a noise-aware critic scored on a face-localized decode of one-step predictions, since both streams are generated and only the face carries the signal. (3) **TimbreTune** is an optional per-sample test-time step for the hard-speaker tail. On the open-source EMTD benchmark, VESTA raises reference-timbre similarity by nearly 50% (from 0.40 to 0.60), a gain ablations attribute to the mechanism rather than the data; it also leads on lip-sync (Sync-C 7.43 vs. 6.55) and stays competitive on PQ, CU and FVD. TimbreTune adds $+6.2\%$, most on the hardest speakers.

1 INTRODUCTION

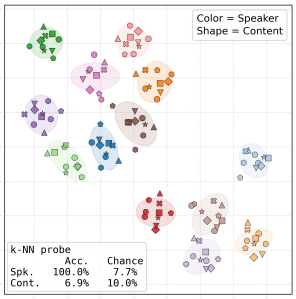
Convincing talking-human video requires agreement on two fronts: the voice timbre fits the person's identity, and the lips follow that voice. Recent joint audio–video generators (SII-OpenMOSS Team, 2026) produce realistic talking humans from a reference portrait and a spoken script, but master neither. The generated timbre is drawn from the model's prior rather than specified, hence uncontrolled; and lip synchronization, to which viewers are especially sensitive, remains imperfect. As a result, a synthesized presenter may *look* like the target person but *sound* like someone else, and its lips may fall out of step with the generated speech. For personalized digital humans and virtual presenters, both a faithful voice and accurate lip synchronization are essential, yet remain unmet.

Several recent methods address this gap in *joint talking audio–video generation with reference timbre* (Fig. 1(a)), differing mainly in *where* the reference is fused: as extra tokens concatenated into the attended sequence (Zhang et al., 2025; Guo et al., 2026; Li et al., 2026b), additively onto the token stream (Qiang et al., 2026), or through a parallel auxiliary branch (Li et al., 2026a). Despite their diversity, these methods share three fundamental limitations. **First**, they inject the reference timbre into the token stream, so this global, content-agnostic attribute shares the activations that encode content, and their separation is left to be *learned*; raw-token injection may moreover leak the reference's words and prosody into the phonetic content the lips must track. The timbre thus stays suboptimal while lip alignment is disturbed as well. **Second**, supervision is limited to latent-space objectives, offering no explicit, targeted signal on the decoded output for either timbre fidelity or lip synchronization, leaving both suboptimal. **Third**, they offer no per-speaker adaptation at inference: each reference speaker is served by a single forward pass through a fixed model, with no mechanism to further refine timbre fidelity for the specific voice.

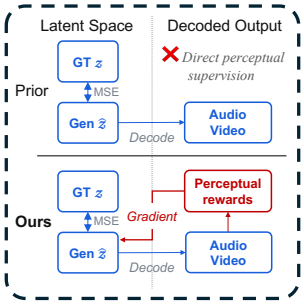
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



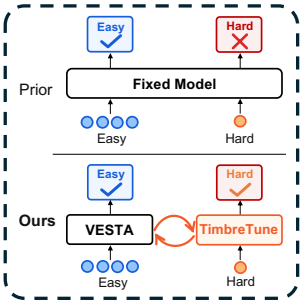
(a) Joint Talking Audio–Video Generation with Reference Timbre



(b) t-SNE of Channel-wise Mean/Std



(c) Latent vs. Perceptual Supervision



(d) Hard-Speaker Adaptation

Figure 1: **(a)** Our joint talking audio–video generation task. **(b)** In our audio VAE, timbre is a channel statistic: t-SNE of the per-channel (μ, σ) of audio VAE latents; each point one utterance, color denotes speaker and shape denotes content. Samples cluster by speaker, not by content. **(c)** Prior work supervises only in latent space, reaching perceptual quality indirectly. **(d)** A fixed model under-serves hard speakers' timbre, which per-sample reward optimization at inference can improve.

This raises a natural question: how should these three gaps be addressed? We ground our answer in three insights, summarized in Fig. 1(b–d). **For conditioning**, we start from the well-established principle that feature statistics encode style and identity (Huang & Belongie, 2017), and ask *where* a speaker's timbre resides in the audio latent our audio–video generator denoises. Timbre turns out to be a global, time-invariant attribute: from the time-pooled per-channel (μ, σ) of the latent, a simple k-NN probe identifies the speaker with 100% accuracy, while a t-SNE of these same statistics clusters tightly by speaker, irrespective of the spoken content (Fig. 1(b)). Adaptive normalization already reshapes these per-channel statistics to inject the timestep; since timbre likewise resides in these statistics, we route it through the same mechanism rather than adding tokens. **For supervision**, timbre fidelity and lip synchronization are perceptual properties of the decoded output, so existing joint talking audio–video generators, supervising only in latent space, reach them indirectly (Fig. 1(c), top). Yet a pretrained critic can score the decoded output directly, and backpropagation through the decoder has proven effective (Xu et al., 2023; Liu et al., 2025). We bring this recipe to joint audio–video generation of talking humans, instantiating it with two task-specific rewards: reference-timbre similarity and lip–speech synchronization, on top of reconstruction (Fig. 1(c), bottom). **For adaptation**, a fixed model handles typical speakers well but inevitably under-serves challenging ones. Since our reward signal is differentiable and available at inference, these cases admit a further recourse: per-sample optimization toward the reference voice (Fig. 1(d)).

Guided by these insights, we build **VESTA** (**V**ideo-audio **E**xpert-guided **S**ynchronized **T**imbre-faithful **A**vatar), a joint audio–video generator that renders a lip-synchronized talking human from a reference portrait and voice. It comprises three components. **(1) StatShift.** Rather than introducing the reference as additional tokens, we condition the audio branch's normalization on the reference speaker embedding, applying the same channel-wise scale and shift to every audio token. As this modulation is uniform across the sequence, it carries no token-to-token variation in words or prosody, thereby disentangling identity from content *at the injection pathway*. **(2) RewardSteer.** We extend perceptual reward feedback to joint talking audio–video generation. Two rewards act on one-step denoised predictions: a TimbreReward toward the reference speaker and a SyncReward for lip–audio synchronization. The SyncReward uses a noise-aware critic, trained on one-step denoised frames and audio, together with a face-localized partial decode that concentrates the reward on the speaker's face and keeps the video-decoder gradient tractable. **(3) TimbreTune.** At inference, we further adapt the model per sample under the timbre reward, pushing reference-timbre fidelity beyond a single feed-forward pass.

On the open-source EMTD benchmark (Meng et al., 2025), VESTA raises reference-timbre similarity by nearly 50% relative to the best competitor while also leading on lip-sync (Sync-C 7.43 vs. 6.55); TimbreTune adds a further $+6.2\%$.

Our contributions are summarized as follows:

- We propose **StatShift**: since token-level timbre injection in a coupled generator leaks time-varying content and harms lip-sync, we instead modulate the audio latent's channel statistics uniformly across time, matching where timbre resides and *empirically* preserving sync.
- We introduce **RewardSteer**, which supplies the perceptual supervision joint generators lack through a timbre-similarity reward and a lip-sync reward; the latter is redesigned for generated, noisy speech via a noise-aware critic and a face-localized partial decode.
- We add **TimbreTune**, an optional test-time recourse that adapts the model per sample to recover the hard-speaker tail.
- On EMTD, VESTA leads on timbre and lip-sync while staying competitive on generation quality; ablations attribute the timbre gain to the injection mechanism, not the training data.

2 Related Work

2.1 Joint Audio-Video Generation

Recent works have explored jointly generating synchronized audio and video within a unified diffusion framework, fusing the two modalities either within a single tower via shared self-attention (Zhao et al., 2025; Wang et al., 2024a) or across parallel video and audio DiTs via cross-modal attention (SII-OpenMOSS Team, 2026; Low et al., 2025). While these systems produce realistic talking-human videos, the generated voice is drawn from the model's prior and cannot be controlled.

To address this, several methods further condition generation on a reference speech clip that specifies the target timbre. Their injection mechanisms fall into three families: (i) in-context fusion, which exposes the reference as extra tokens via sequence concatenation (Zhang et al., 2025; Guo et al., 2026; Li et al., 2026b); (ii) token-additive fusion, which adds a pooled speaker embedding onto the token stream (Qiang et al., 2026); and (iii) auxiliary-branch fusion, which routes the reference through a parallel network merged back by gated residuals (Li et al., 2026a). Despite their diversity, all inject timbre on the token axis, not in normalization statistics, and supervise only within the diffusion noise space, leaving timbre fidelity and lip synchronization without perceptual supervision.

2.2 Reference-Timbre Control

Voice control is well-established in text-to-speech, through zero-shot cloning and speaker-adaptive fine-tuning (Wang et al., 2024b; Bondaruk & Kubiak, 2025) that injects the speaker either by conditioning normalization on a reference speaker embedding (Chen et al., 2021; Min et al., 2021) or by optimizing a speaker-similarity reward toward the reference voice (Anastassiou et al., 2024). The question in our setting is *where* timbre should enter a joint audio–video generator, in which prior methods inject it as tokens; StatShift instead routes it through normalization statistics, so that the injected timbre shares the audio DiT with timestep modulation and propagates through the cross-modal bridge to the lips. Those speech-synthesis methods, however, are audio-only: they synthesize speech but no accompanying video or lip motion. Our setting instead requires generating the synchronized video and audio together while controlling the reference timbre, so we compare against joint audio–video generators, which produce such clips but, as our experiments show, do not control the reference timbre faithfully.

2.3 Reward-based Training for Diffusion

Reward-based optimization narrows the gap between proxy training losses and perceptual quality in diffusion models. In image generation, it fine-tunes diffusion models against task-specific rewards by backpropagating reward gradients through the generation process (Xu et al., 2023; Clark et al., 2024). Extended to video, such reward gradients are typically computed over the full generated frame (Prabhudesai et al., 2024; Yuan et al., 2024). For lip synchronization, a pretrained SyncNet has served as the differentiable critic (Prajwal et al., 2020; Li et al., 2024); there, however, the audio is a fixed, real recording and the reward only steers the generated lips toward it. Joint talking audio-video generation reshapes this reward in two respects. First, as the audio is itself generated, the reward acts on both generated streams rather than aligning the video to a fixed track, and a critic

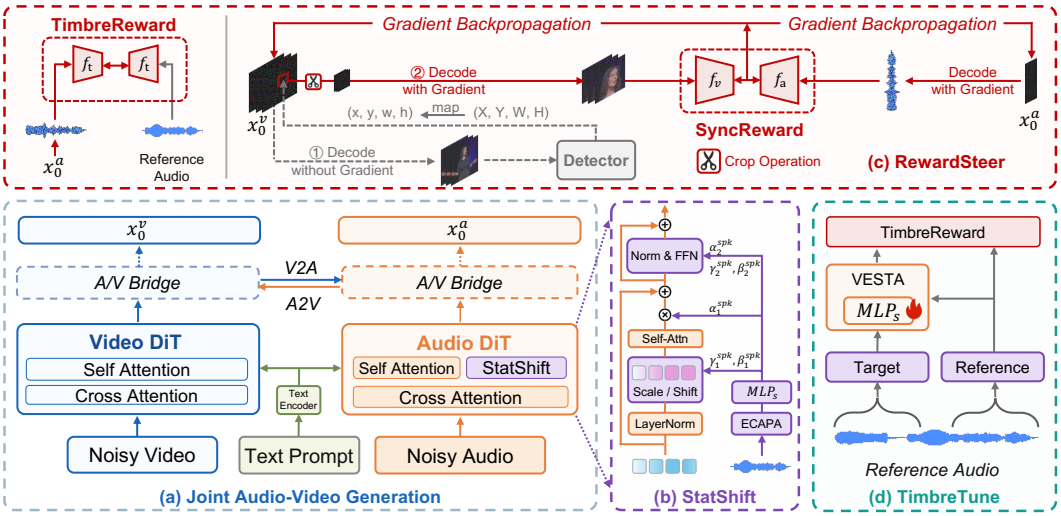


Figure 2: Overview of VESTA. **(a)** Video and audio DiTs denoise via a bidirectional bridge. **(b) StatShift** modulates each audio normalization layer's scale and shift. **(c) RewardSteer** rewards timbre and lip-sync on one-step predictions. **(d) TimbreTune** adapts per sample at inference.

trained on clean audio can no longer be trusted on the resulting noisy speech. Second, unlike these full-frame video rewards, a sync reward must be directional: joint talking generation renders a full human whose torso and background carry no lip-audio evidence, so scoring the entire frame dilutes the signal and, with gradients passing through the VAE decoder, becomes intractable.

3 Method

VESTA (Figure 2) is a dual-branch joint audio–video diffusion model that takes a reference image, a text prompt, and a reference speech clip, and generates a human video with synchronized audio (§3.1). Its three components map one-to-one onto the gaps of §1: StatShift (§3.2) for timbre conditioning, RewardSteer (§3.3) for perceptual supervision during training, and TimbreTune (§3.4) for test-time speaker adaptation.

3.1 Joint Audio-Video Generation

Notation and preliminaries. A video clip and its speech track are encoded into latents $z_0^v = \mathcal{E}_v(x^v)$ and $z_0^a = \mathcal{E}_a(x^a)$ by a pretrained video VAE $\mathcal{E}_v$ and a pretrained audio VAE $\mathcal{E}_a$, respectively; z_0^v is a spatio-temporal grid of tokens, whereas z_0^a is a one-dimensional sequence along time. Both branches are trained with flow matching (Lipman et al., 2023; Liu et al., 2023): for a timestep $t \in [0, 1]$ and noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the noisy latent is obtained by linear interpolation, $z_t = (1-t)\, z_0 + t\, \epsilon$, and the network is trained to regress the corresponding velocity field $u = \epsilon - z_0$.

Video branch. A reference image and a text prompt, which jointly describes the visual scene and specifies the transcript to be spoken, are encoded by $\mathcal{E}_v$ and a frozen text encoder into the conditions c^{img} and c^{txt}. The video DiT (Peebles & Xie, 2023) predicts the velocity of the video latent and is optimized by

$$\mathcal{L}^v = \mathbb{E}_{t,\, z_0^v,\, \epsilon}\Big[\big\|\hat{v}_\theta^v(z_t^v,\, t,\, c^{\text{img}},\, c^{\text{txt}}) - (\epsilon - z_0^v)\big\|_2^2\Big]. \tag{1}$$

Audio branch. The audio DiT shares the same text condition c^{txt}, whose transcript specifies the speech to be uttered, and additionally receives the reference timbre c^{spk} injected by the proposed StatShift (§3.2). Its objective is analogous:

$$\mathcal{L}^a = \mathbb{E}_{t,\, z_0^a,\, \epsilon}\Big[\big\|\hat{v}_\theta^a(z_t^a,\, t,\, c^{\text{txt}},\, c^{\text{spk}}) - (\epsilon - z_0^a)\big\|_2^2\Big]. \tag{2}$$

Cross-modal bridge. Following SII-OpenMOSS Team (2026), the two branches are coupled by a bidirectional bridge that inserts a pair of cross-attention blocks at every layer, injecting video hidden states into the audio stream and vice versa.

3.2 StatShift

In our joint audio-video generator the audio and video branches are coupled, so disturbing the audio content also disturbs lip alignment; timbre injection must therefore control the voice while keeping that content largely undisturbed. Speaker timbre is carried by the per-channel mean and variance of this audio latent (Fig. 1(b)). StatShift therefore controls timbre by modulating these statistics: it conditions each audio DiT block's normalization on a reference speaker embedding through a single lightweight projection shared across blocks. Applied identically to every token, this modulation injects no time-varying signal, so the spoken content the lips track is left largely undisturbed, in contrast to token-level injection.

Global timbre modulation. Given a reference speech clip, a frozen ECAPA-TDNN (Desplanques et al., 2020) speaker encoder produces a single embedding e^{spk} summarizing the speaker's voice. Each audio DiT block modulates its self-attention and FFN sub-layers via AdaLN, giving six vectors of width D per block (a scale, shift, and gate for each sub-layer). StatShift maps e^{spk} into this same space through a lightweight projection MLP_s:

$$\Delta^{\text{spk}} = \mathrm{MLP}_s(e^{\text{spk}}) \in \mathbb{R}^{6\times D}, \tag{3}$$

that are added to those from the timestep, so that for block ℓ

$$(\gamma_i,\ \beta_i,\ \alpha_i)_{i=1}^{2} = \mathbf{m}^{(\ell)} + \mathrm{MLP}_t(e^t) + \Delta^{\text{spk}}, \tag{4}$$

where $\mathbf{m}^{(\ell)}$ is a per-block learnable base and $\mathrm{MLP}_t(e^t)$ the backbone's existing timestep projection; MLP_s is the only module StatShift adds. For the i-th modulated sub-layer,

$$h_i' = (1+\gamma_i)\odot \mathrm{LayerNorm}(h_i) + \beta_i, \tag{5}$$

$$h_{i+1} = h_i + \alpha_i \odot \mathrm{SubLayer}_i(h_i'). \tag{6}$$

Shared across all time positions, this modulation acts on the global channel statistics and adds no time-varying signal of its own, so its timbre injection stays largely decoupled from content.

Decoupled speaker guidance. We give timbre its own guidance axis rather than folding it into the text guidance. During training, e^{spk} is replaced by a zero vector with probability p, providing an unconditional speaker branch for guidance at inference. We then form a three-branch guidance that decouples text and speaker conditioning in the audio branch, which alone receives the speaker embedding e^{spk}:

$$\hat{v}^a = v_\varnothing + s_{\text{txt}}\,(v_{\text{txt}} - v_\varnothing) + s_{\text{spk}}\,(v_{\text{cond}} - v_{\text{txt}}), \tag{7}$$

where $v_\varnothing$ uses a null prompt and a zero speaker embedding, v_{txt} the text prompt with a zero speaker embedding, and v_{cond} both. The two scales act independently: s_{txt} controls adherence to the text prompt, and s_{spk} the reference-timbre strength.

3.3 RewardSteer

Reconstruction losses alone leave both timbre fidelity and lip synchronization without direct supervision. Following reward feedback learning (Xu et al., 2023), we introduce **RewardSteer** (Fig. 2(c)): a timbre reward toward the reference speaker and a lip-sync reward that tightens mouth–speech alignment. Both are defined on the one-step denoised prediction rather than on the noisy latent. Given z_t and the predicted velocity $\hat{v}_\theta$, flow matching admits the closed-form estimate

$$\hat{z}_0 = z_t - t\,\hat{v}_\theta. \tag{8}$$

Each reward is evaluated on $\hat{z}_0$ after decoding, and its gradient flows back through the frozen VAE decoder and Eq. 8 into the denoising model. Since a one-step estimate is only reliable once enough noise has been removed, each reward is gated by a timestep threshold.

TimbreReward. The audio estimate $\hat{z}_0^a$ is decoded into a waveform with gradients retained, and both this waveform and the reference clip are encoded by a frozen speaker verification network into e_{gen} and e_{ref}. The reward drives up their cosine similarity,

$$\mathcal{L}_{\text{tim}} = 1 - \sigma\big(\tau\,(s - m)\big), \quad s = \cos(e_{\text{gen}},\, e_{\text{ref}}), \tag{9}$$

where σ is the sigmoid function and τ, m control its sharpness and margin, as in standard speaker verification objectives (Wan et al., 2018); the gradient thus concentrates around the margin and saturates once the two voices are already close. This reward is applied whenever $t \le T_{\text{tim}}$. Two properties are worth noting: e_{ref} comes from the conditioning reference clip rather than the ground-truth target speech, and the verification network here is trained independently of the one used for evaluation (§4.1) and shares no weights with it; the reward thus neither leaks target information nor directly optimizes the reported metric.

SyncReward. A pretrained SyncNet's usual operating point does not hold here: it assumes a clean, real audio track, whereas both modalities are generated and still noisy when the reward fires, making its score and gradient unreliable. A further question is where to evaluate the reward: only the face bears the synchronization signal while torso and background carry none, so scoring the whole frame is undirected. Moreover, the reward's gradient must pass through the VAE decoder, where decoding the full clip is intractable.

We address the first by pretraining a pair of visual and audio encoders f_v and f_a on one-step denoised video frames paired with one-step denoised audio, so that both carry the residual noise the reward will encounter; the two encoders are trained with an InfoNCE objective (van den Oord et al., 2018), hard-negative mining, and data augmentation.

We address the second by confining the differentiable decode to the face. Since the face is not at a fixed location and may move across frames, we locate it dynamically. Per training step we sample a single contiguous segment of K latent frames. With probability 0.5 we choose it uniformly at random; otherwise we draw it from the top-scoring segments by audio spectral flux (Bello et al., 2005). We then decode this segment in two passes. The first pass decodes the segment *without* gradients, so that a face detector can locate the speaker. The detected box is expanded and mapped back to latent coordinates, and the second pass re-decodes only that region *with* gradients. This focuses both the reward and its gradient on the only region that bears the supervisory signal; as a by-product, the retained computation graph is bounded to a small face-centered block rather than the full clip, since only the second pass carries gradients. Pseudocode of this face-aware partial decode is provided in the supplementary material.

The decoded region is resized to a fixed resolution, giving a short face clip $\hat{V}$. The audio estimate $\hat{z}_0^a$, whose decoder is far cheaper, is decoded in full, and the mel-spectrogram segment aligned with the same time span is taken as $\hat{A}$. The reward is the agreement between the two encoders' embeddings,

$$r_{\text{sync}} = \cos\big(f_v(\hat{V}),\, f_a(\hat{A})\big), \quad \mathcal{L}_{\text{sync}} = 1 - r_{\text{sync}}. \tag{10}$$

The overall training objective is

$$\mathcal{L} = \lambda_v \mathcal{L}^v + \lambda_a \mathcal{L}^a + \lambda_{\text{tim}} \mathcal{L}_{\text{tim}} + \lambda_{\text{sync}} \mathcal{L}_{\text{sync}}, \tag{11}$$

where the weights balance the flow-matching and reward terms; each reward is zero outside its gate ($t \le T_{\text{tim}}$ and $t \le T_{\text{sync}}$, respectively).

3.4 TIMBRETUNE

To recover the tail of hard speakers a fixed model under-serves, we propose **TimbreTune** (Fig. 2(d)), a reward-guided test-time training strategy that adapts the model per reference using the differentiable timbre reward of §3.3.

Since a reference clip contains multiple segments of the same speaker, we construct several (reference, target) pairs from it. For each pair, we sample timesteps from the inference schedule, compute the one-step denoised audio, and optimize the timbre reward of Eq. 9 with respect to the model parameters. Only the LoRA (Hu et al., 2022) parameters of the audio DiT's speaker projection MLP_s are updated, so adaptation refines timbre without retraining the coupled video branch. After adaptation, the model runs standard inference to co-generate the final audio–video output.

Table 1: Quantitative comparison on the open-source EMTD test set (110 English samples). ↑/↓: higher/lower is better; **bold** marks the best per column.

Method	Audio Quality			Video Quality		Audio–Video Consistency		
	WER↓	PQ↑	CU↑	FID↓	FVD↓	SIM↑	Sync-C↑	Sync-D↓
MOVA	0.058	6.78	6.38	49.8	140.8	0.099	6.17	8.75
Ovi	**0.044**	5.74	5.82	45.8	156.7	0.115	6.55	8.52
DreamID-Omni	0.075	5.42	5.42	90.8	279.6	0.382	4.93	9.71
OmniCustom	0.059	5.37	5.44	89.1	219.4	0.365	5.37	9.23
UniAVGen	0.090	6.45	5.96	**43.4**	153.9	0.402	4.83	9.99
VESTA (ours)	0.051	**6.89**	**6.50**	47.4	**132.9**	**0.600**	**7.43**	**7.98**

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

VESTA is trained in two stages with AdamW at learning rate 1×10^{-4}. **Stage 1** builds joint audio–video generation: the audio DiT is pretrained on the English subset of Emilia (He et al., 2024), the video DiT initialized from Wan 2.2 I2V A14B (Wan Team et al., 2025), and the two are coupled via a bidirectional cross-modal bridge (SII-OpenMOSS Team, 2026) and jointly trained on an internally collected talking-human dataset for 100K steps (batch size 192). **Stage 2** adds StatShift and RewardSteer and continues with the full objective (Eq. 11) for 50K steps (batch size 96).

The speaker embedding is extracted from a frozen ECAPA-TDNN encoder and dropped with probability 0.2 (text prompt: 0.1) to enable the CFG branches of Eq. 7; inference uses 30 steps. The TimbreReward uses an ECAPA-TDNN verification network, disjoint from the WavLM-large (Chen et al., 2022) model that scores our SIM metric. All four loss weights are 1; remaining reward hyperparameters ($\tau, m, T_{\text{tim}}, T_{\text{sync}}, K$), CFG scales, and SyncNet detector settings are in the supplementary material. Unless explicitly stated in §4.3.3, all experiments run in a single feed-forward pass without TimbreTune, isolating StatShift and RewardSteer for fair comparison.

4.2 COMPARISON WITH PRIOR METHODS

Compared methods. We compare VESTA against five recent joint audio–video generators, grouped by whether they condition on a reference timbre. *Without timbre conditioning*: MOVA (SII-OpenMOSS Team, 2026) and Ovi (Low et al., 2025), both dual-tower models coupled by bidirectional cross-attention. *With timbre conditioning*: UniAVGen (Zhang et al., 2025), which concatenates reference-audio mel latents into the audio input sequence; DreamID-Omni (Guo et al., 2026), which concatenates reference-audio VAE latents along the sequence dimension; OmniCustom (Li et al., 2026b), which attends to reference-audio tokens via LoRA-adapted self-attention.

Evaluation benchmark. We evaluate on the open-source EMTD test set (Meng et al., 2025), all 110 English samples, fully disjoint from our training data. Each sample provides a reference image, a video caption, a reference audio, and a driving speech content; the driving content is cyclically shifted so it never matches its reference-audio transcript, preventing content leakage.

Metrics. We report metrics along three axes. *Audio quality*: Word Error Rate (WER) from Whisper-large-v3 (Radford et al., 2023), plus Production Quality (PQ) and Content Usefulness (CU) from AudioBox-Aesthetics (Tjandra et al., 2025). *Video quality*: Fréchet Inception Distance (FID) (Heusel et al., 2017) and Fréchet Video Distance (FVD) (Unterthiner et al., 2018). *Audio–video consistency*: SyncNet (Chung & Zisserman, 2017) confidence and distance (Sync-C↑/Sync-D↓) for lip-sync, and a Speaker Similarity (SIM) score following seed-tts-eval (Anastassiou et al., 2024)—cosine similarity between WavLM (Chen et al., 2022) speaker embeddings of generated speech and reference audio—for reference-timbre fidelity.

Quantitative comparison. Table 1 summarizes the results across three axes. *Audio–video consistency*: VESTA raises SIM to 0.600, a nearly 50% relative gain over the strongest timbre-conditioned competitor (0.402; mechanism-driven, see §4.3.1); the two unconditional baselines sit near chance while the timbre-conditioned ones plateau at 0.37–0.40, transferring only a coarse identity. VESTA also delivers the strongest lip-sync (Sync-C 7.43), ahead of the next-best Ovi (6.55) and far above



Figure 3: Qualitative comparison of VESTA against MOVA (SII-OpenMOSS Team, 2026) and UniAVGen (Zhang et al., 2025) on two samples. The *Input* row provides the reference image, the reference audio, and the target text with annotated phonemes. Each subsequent row presents one method's generated frames at the three annotated instants, along with its per-sample speaker similarity (Sim) and lip-sync confidence (Sync-C).

every timbre-conditioned baseline (Sync-C ≤ 5.37). *Video quality*: VESTA maintains competitive visual fidelity (FID 47.4, FVD 132.9), close to the strongest baseline UniAVGen on FID, best on FVD, and clearly ahead of the other timbre-conditioned methods on both. *Audio quality*: VESTA scores the best PQ (6.89) and CU (6.50), on par with unconditional MOVA and clearly ahead of every timbre-conditioned baseline, showing that timbre control does not cost speech quality.

Qualitative comparison. Figure 3 contrasts VESTA with MOVA and UniAVGen on two samples at three phoneme instants, each demanding a distinct mouth shape. MOVA, lacking timbre conditioning, produces a mismatched voice with only moderate lip-sync. UniAVGen recovers the reference timbre only coarsely and has the weakest lip-sync of the three: its lips remain open at /m/ and unrounded at /oU/ and /w/. VESTA leads on both axes by a clear margin, showing the intended mouth shape at each annotated phoneme instant while preserving the reference voice.

Human evaluation. Beyond automatic metrics, we conduct a double-blind pairwise GSB study on EMTD. Thirteen independent raters compared VESTA to MOVA and UniAVGen along two dimensions (lip-sync and timbre similarity), marking each pair as *Win* (VESTA better), *Tie*, or *Lose* (baseline better). The study contributes 1040 pairwise judgements in total, i.e. 260 pooled votes per (dimension, baseline) pair. Table 2 shows VESTA is preferred over both baselines on both dimensions, with GSB scores of $+0.32$ to $+0.71$ across all four comparisons.

Table 2: Pairwise GSB human evaluation. GSB=(Win−Lose)/N.

Setup	Win/Tie/Lose (%)	GSB
Sync vs. MOVA	55.6 / 20.6 / 23.7	+0.32
Sync vs. UniAV	75.4 / 9.0 / 15.6	+0.60
SIM vs. MOVA	82.7 / 5.8 / 11.5	+0.71
SIM vs. UniAV	72.3 / 3.8 / 23.8	+0.49

4.3 Ablation Studies

For efficiency, unless otherwise specified, all ablation variants are warm-started from the same Stage-1 checkpoint, trained on an English subset under identical settings, and evaluated on EMTD.

4.3.1 Timbre Injection Mechanism.

To test whether statistics modulation beats token-level conditioning for the global, channel-wise timbre attribute, we compare: (1) *No injection*, where no speaker embedding is provided; (2) *Self-attention*, which concatenates reference audio tokens to the target audio sequence; (3) *Cross-attention*, which routes the reference tokens through the conditioning context; (4) *Addition*, which sums a pooled speaker embedding onto the audio hidden states at each block; (5) *StatShift* (ours), which modulates the layer normalization of each audio DiT block with the speaker embedding.

As shown in Table 3, all four conditioning methods improve over the no-injection baseline (0.080), confirming that the speaker embedding carries useful timbre information regardless of the injection pathway. StatShift nonetheless achieves the highest similarity at 0.522, exceeding self-attention by 0.205, cross-attention by 0.168, and addition by 0.138; since these token-injection variants share our backbone, data, and rewards yet only match the token-injection competitors of Table 1 (SIM

Table 3: Ablation of timbre injection.

Injection method	SIM ↑	Sync-C ↑	Sync-D ↓
No injection	0.080	7.29	8.09
Self-attention	0.317	6.85	8.35
Cross-attention	0.354	6.91	8.31
Addition	0.384	7.21	8.13
StatShift	**0.522**	**7.34**	**8.06**

Table 4: Ablation of reward objectives; Timbre/Sync columns toggle the two rewards.

Timbre	Sync	SIM ↑	Sync-C ↑	Sync-D ↓
✗	✗	0.462	6.11	8.78
✗	✓	0.469	7.21	8.12
✓	✗	0.519	6.16	8.75
✓	✓	**0.522**	**7.34**	**8.06**

0.37–0.40), this margin reflects the injection mechanism rather than the training set. This aligns with Fig. 1(b): timbre resides in channel-wise statistics, precisely what AdaLN modulates. On lip-sync, the token-level methods (self- and cross-attention) fall below the no-injection reference while StatShift matches it, indicating that token injection tends to cost synchronization through content leakage (§3.2), which StatShift's uniform modulation avoids, as does additive injection.

4.3.2 Reward Objectives.

We evaluate the contribution of our two rewards, a speaker-similarity reward for timbre and a SyncNet reward for lip-sync, by toggling each on or off with all other settings fixed.

As shown in Table 4, each reward clearly improves its target metric. The timbre reward raises SIM from 0.462 to 0.519 ($+12\%$ relative), confirming that explicit speaker-similarity supervision is essential for faithful timbre reproduction. The sync reward improves Sync-C from 6.11 to 7.21 and Sync-D from 8.78 to 8.12, demonstrating that direct lip-sync supervision effectively tightens mouth–speech alignment. Crucially, the two rewards do not interfere with each other: the timbre reward's gain on SIM is stable regardless of whether the sync reward is active ($+0.057$ vs. $+0.053$), and the sync reward has negligible effect on SIM ($+0.007$ and $+0.003$). The full combination achieves the best performance across all three metrics, confirming that the two objectives can be jointly optimized. We attribute this training-time independence to StatShift's decoupled injection: timbre is routed through channel statistics rather than the content-carrying tokens the lips track, so the two rewards act on largely orthogonal factors.

4.3.3 TimbreTune.

To evaluate the optional TimbreTune adaptation (§3.4), we rank samples by their baseline SIM (before adaptation) and report the cumulative average SIM, before and after adaptation, for the bottom-$X\%$.

Table 5: TimbreTune improvement stratified by sample difficulty (bottom-$X\%$ of baseline SIM).

Bottom	Before	After	Δ	Δ%
10%	0.415	0.523	+0.108	+26.0%
30%	0.479	0.550	+0.071	+14.8%
50%	0.520	0.589	+0.069	+13.3%
70%	0.553	0.605	+0.052	+9.4%
90%	0.584	0.625	+0.041	+7.0%
100%	0.600	0.637	+0.037	+6.2%

As shown in Table 5, TimbreTune improves SIM across all difficulty strata, with the gain on average increasing for harder cases: the bottom 10% improves by $+0.108$ ($+26.0\%$ relative), while the overall average improves by $+0.037$ ($+6.2\%$). This trend suggests that challenging samples have more room for per-sample optimization to recover, while the improvement on the overall average confirms the benefit holds at the dataset level, not only on the hard tail. As an optional post-hoc step, TimbreTune serves as a per-sample recourse where higher timbre fidelity is desired at a small lip-sync trade-off (see supplementary); further ablations on TimbreTune are provided in the supplementary material.

5 Conclusion

We presented VESTA, a joint audio–video generation framework for faithful reference-timbre control. It combines three components: StatShift, an AdaLN-based timbre injection that modulates the audio latent's channel statistics; RewardSteer, differentiable timbre and lip-sync rewards on one-step diffusion predictions; and TimbreTune, a per-sample test-time adaptation under the timbre reward. On the EMTD benchmark, VESTA raises reference-timbre similarity by nearly 50% relative to the best existing method while achieving the strongest lip synchronization.

AI USE STATEMENT

In this work, we used generative AI tools for two tasks in the recommended-disclosure category: to aid or polish writing, and for retrieval and discovery. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Philip Anastassiou, Jiawei Chen, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

Łukasz Bondaruk and Jakub Kubiak. LoRP-TTS: Low-rank personalized text-to-speech. *arXiv preprint arXiv:2502.07562*, 2025.

Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Adaspeech: Adaptive text to speech for custom voice. *arXiv preprint arXiv:2103.00993*, 2021.

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 Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, pp. 251–263, 2017.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proc. Interspeech*, 2020.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. Latentsync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024.

Liyang Li, Wen Wang, Canyu Zhao, Tianjian Feng, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Mmcontrol: Unified multi-modal control for joint audio-video generation. *arXiv preprint arXiv:2604.19679*, 2026a.

Maomao Li, Zhen Li, Kaipeng Zhang, Guosheng Yin, Zhifeng Li, and Dong Xu. Omnicustom: Sync audio-video customization via joint audio-video generation model. *arXiv preprint arXiv:2602.12304*, 2026b.

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

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. *arXiv preprint arXiv:2106.03153*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and C V Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2020.

Chunyu Qiang, Jun Wang, Xiaopeng Wang, Kang Yin, and Yuxin Guo. Mm-sonate: Multimodal controllable audio-video generation with zero-shot voice cloning. *arXiv preprint arXiv:2601.01568*, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023.

SII-OpenMOSS Team. Mova: Towards scalable and synchronized video–audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Wan Team, Ang Wang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. In *NeurIPS Workshop on Audio Imagination*, 2024a.

Wenbin Wang, Yang Song, and Sanjay Jha. USAT: A universal speaker-adaptive text-to-speech approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 15903–15935, 2023.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025.

Lei Zhao, Linfeng Feng, Dongxu Ge, Rujin Chen, Fangqiu Yi, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Uniform: A unified multi-task diffusion transformer for audio-video generation. *arXiv preprint arXiv:2502.03897*, 2025.