OMNICUSTOM: SYNC AUDIO-VIDEO CUSTOMIZATION VIA CONTRASTIVE FLOW MATCHING

Anonymous authors
Paper under double-blind review

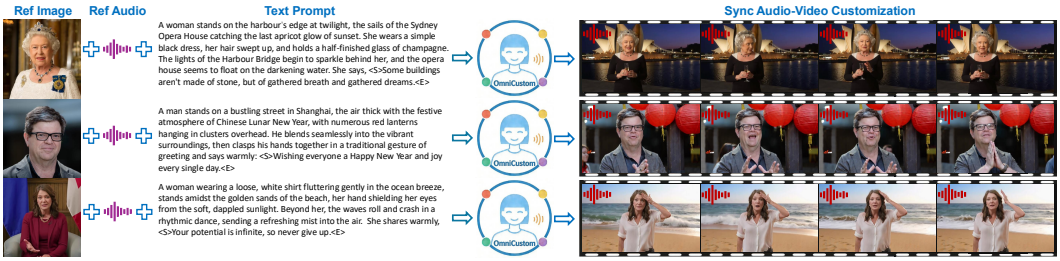


Figure 1: We propose OmniCustom, a novel lightweight framework that achieves state-of-the-art sync audio-video customization. Given a reference image I^r and a reference audio A^r, OmniCustom synchronously generates a video that preserves the identity of I^r and an audio that mimics the timbre of A^r. The speech content can be specified by a text prompt freely.

ABSTRACT

Given a reference image I^r and a reference audio A^r, sync audio-video customization task aims to generate videos that maintain the identity of the reference image while imitating the timbre of the reference audio, with spoken content freely specifiable through user-provided textual prompts. While existing methods have achieved promising progress, a system that provides high-quality preservation of given references while yielding user-friendly inference latency remains a challenge. To deal with this, we propose OmniCustom, a powerful DiT-based sync audio-video customization framework that can synthesize a video following reference image identity, audio timbre, and text prompts all at once in a zero-shot manner. Our framework is built on three designs. First, identity and audio timbre control are achieved by modality-specific LoRA modules that operate through self-attention layers within the joint audio-video generation backbone. Second, we introduce a contrastive learning objective alongside the standard flow matching objective. It uses predicted flows conditioned on reference inputs as positive samples and those without reference conditions as negative samples, enhancing the model's ability to preserve identity and timbre. Third, we train OmniCustom on our constructed large-scale audio-visual human dataset OmniCustom-1M. Extensive experiments demonstrate that OmniCustom achieves state-of-the-art customization performance while providing user-friendly inference runtime.

1 INTRODUCTION

Deep generative models (Goodfellow et al., 2020; Kingma & Dhariwal, 2018; Ho et al., 2020; Liu et al., 2023; Lipman et al., 2023) have demonstrated a remarkable capacity for producing high-quality samples across various data modalities. Leveraging large-scale data and powerful architectures, text-to-video (T2V) (Blattmann et al., 2023; Yang et al., 2025b; Kong et al., 2024; Wan et al., 2025) have enabled the synthesis of vivid visual content from textual descriptions. Building upon these advances, video customization (Jiang et al., 2024; He et al., 2024; Wei et al., 2024; Liu et al., 2025; Jiang et al., 2025) aims at synthesizing identity-preserving videos, which garners increasing attention due to their broad potential for applications ranging from film and advertising to gaming.

Typical video customization techniques can be classified into two categories: tuning-based and tuning-free. Tuning-based methods (Ma et al., 2024; Wang et al., 2024; Wei et al., 2024; Wu et al.,

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.

2024) require fine-tuning the pre-trained model for each new identity during inference, which limits their practical scalability. In contrast, tuning-free methods (He et al., 2024; Huang et al., 2025; Yuan et al., 2025; Jiang et al., 2025) introduces new identities without test-time training. Despite promising performance, their customized outputs yield silent effects. Another line of methods (Yi et al., 2025; Hu et al., 2025; Chen et al., 2025a; Wang et al., 2025) incorporate audio as an additional modality, enabling audio-driven video customization. However, to retain a given timbre, they require generating a driven audio via Text-to-Speech (TTS) (Le et al., 2023) in advance, forming a cumbersome two-stage pipeline. Besides, adopting TTS outputs is incompatible with scenarios containing background sound effects (*e.g.*, beach), which cannot be synthesized by TTS models.

Very recently, thanks to rapid progress in joint audio-video generation (Low et al., 2025; HaCohen et al., 2026), sync audio-video customization (Guo et al., 2026; MiniMaxAI, 2026) has emerged as a promising research direction. Specifically, as seen in Fig. 1, given a reference image I^r, a reference audio A^r, and a textual prompt, this task aims to synchronously generate a video preserving the identity of I^r and corresponding audio mimicking the timbre of A^r. Here, spoken content can be freely given via textual prompt and contextually relevant background sounds can be produced synchronously. As a representative work, DreamID-Omni (Guo et al., 2026) attempts to achieve multiple functionalities of customization, editing, and annotation within a single network. However, it struggles with high-quality customization due to insufficient dedicated design for reference preservation. Besides, MiniMax-H3 (MiniMaxAI, 2026) is an omni generative system handling multimodal contexts including text, images, video, and audio. Nonetheless, it suffers from high inference overhead, requiring above 15 minutes of runtime on an H100 GPU for a 5-second 720P video. This naturally begs the question of whether we can construct a high-performance sync audio-video customization method with user-friendly inference latency. This paper investigates such feasibility.

We present OmniCustom, an efficient reference-guided Diffusion Transformer (DiT) framework (Peebles & Xie, 2023), via several key innovations. First, conditioned on the joint audio-video generation model OVI (Low et al., 2025), our method introduces separate reference-image and reference-audio branches into the original video and audio streams. Specifically, to maintain efficiency, we incorporate two modality-specific LoRAs (Hu et al., 2022) into the QKV projections of the reference features, which avoids massive computational overhead. Second, our approach is trained with a compound objective that pairs the standard flow-matching loss with an auxiliary contrastive learning objective, which maximizes the dissimilarity between predicted flows from samples with reference conditions and those without, thus enhancing identity and timbre preservation. Third, leveraging these techniques, we train OmniCustom on OmniCustom-1M, a large-scale, high-quality audio-visual human dataset we constructed, comprising one million human portrait videos with rich identity and timbre diversity. Our main contributions are:

- We introduce OmniCustom, a lightweight sync audio-video customization model that uses modality-specific LoRAs to efficiently incorporate reference image and audio branches into the original video and audio streams in OVI, respectively.
- To further improve the fidelity of video identity and audio timbre, we design a contrastive learning objective that uses samples with reference conditions as positive samples and those without as negative samples.
- We construct a large-scale audio-visual human dataset comprising one million examples to train the proposed OmniCustom. Extensive experiments demonstrate that our method yields sate-of-the-art sync audio-video customization performance with user-friendly runtime.

2 Related Works

Joint Audio-Video Generation. Joint audio-visual generation has witnessed rapid advances in recent years. MM-Diffusion (Ruan et al., 2023) is the first attempt, where two subnets for audio and video learn to gradually generate aligned audio-video pairs from Gaussian noise. However, it is trained on a limited data scope (Lee et al., 2022b; Li et al., 2021), leading to insufficient generalization. Then, Seeing-and-Hearing (Xing et al., 2024) adopts ImageBind (Girdhar et al., 2023) as an aligner within the diffusion latent space across different modalities. Nevertheless, it sometimes produces low-quality and temporally misaligned outputs. Recently, Veo 3 (DeepMind, 2025), Sora 2 (OpenAI, 2025), and Seedance 2.0 (ByteDance, 2026) mark new milestones. The open-source OVI (Low et al.,

2025) adopts an audio backbone from MMAudio (Cheng et al., 2025), where audio-video fusion is realized via paired cross-attention between the two modalities.

Conditional Audio Generation. Generative audio modeling is a well-established field covering speech (Shen et al., 2018; Li et al., 2019; Ren et al., 2020; Kim et al., 2020; Popov et al., 2021) and environmental sounds (Huang et al., 2023; Liu et al., 2024; Yang et al., 2023). Text-to-Speech (TTS) (Wang et al., 2023; Zhang et al., 2023; Le et al., 2023) synthesizes speech from input text and mimics speakers from audio prompts. Diffusion-based TTS (Shen et al., 2023; Ju et al., 2024; Chen et al., 2025b; Du et al., 2024) supports parallel processing for fast inference. Another line of work (Peng et al., 2024; Han et al., 2024; Du et al., 2025; Liao et al., 2024) adopts autoregressive architectures, sequentially predicting next tokens to enable zero-shot TTS.

Identity-preserving Video Customization. In recent years, video-customization architectures have shifted from UNet to DiT (Peebles & Xie, 2023). A representative UNet-era method, MagicMe (Ma et al., 2024), trains each personalized ID separately and struggles with zero-shot capabilities. By contrast, ID-Animator (He et al., 2024) achieves personalized outputs from the reference image without extra training. DiT further advances token-based video customization (Yuan et al., 2025; Jiang et al., 2025; Liu et al., 2025). For example, VACE (Jiang et al., 2025) builds an all-in-one video model with a pluggable Context Adapter for cross-task concept injection. Phantom (Liu et al., 2025) enhances joint text-image injection, learning cross-modal alignment from text-image-video triplets.

Very recently, DreamID-Omni (Guo et al., 2026) and MiniMax-H3 (MiniMaxAI, 2026) fulfill sync audio-video customization. Specifically, DreamID-Omni proposes a multi-task framework supporting sync customization, video editing, and audio-driven animation within a single model. MiniMax-H3 suffers from high inference latency: generating a 5-second customized video takes over 15 minutes. Different from them, this paper proposes a dedicated customization framework, which adopts two modality-specific LoRAs for audio-video branch fine-tuning, where contrastive flow matching objectives are designed to enhance the model's ability to preserve identity and timbre.

3 PRELIMINARY

Synchronous Audio-Video Generation. OVI (Low et al., 2025) is a open-source joint audio-video generation model. It adopts a twin backbone with parallel audio and video DiT branches. The video branch is initialized from Wan2.2 5B (Wan et al., 2025), while the structurally identical audio branch is trained from scratch using the pre-trained 1D VAE from MMAudio (Cheng et al., 2025). As illustrated in Fig. 2 (b), OVI's fusion block employs paired cross-attention layers for audio-video fusion, allowing the audio stream to attend to the video stream and vice versa.

Flow-matching Objective. OVI (Low et al., 2025) and Wan2.2 (Wan et al., 2025) apply the flow matching loss (Liu et al., 2023; Lipman et al., 2023) to learn a straight flow trajectory between data and noise. With a Gaussian latent $\boldsymbol{z}_1$, the forward process linearly corrupts the clean latent $\boldsymbol{z}_0$ as:

$$\boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, \tag{1}$$

where t is sampled from a uniform distribution. Then, the backward process learns a velocity field $v_\theta(\cdot)$ that maps samples from the Gaussian distribution to the data distribution. This is formalized as a least squares regression problem, where v_θ is optimized to approximate $\boldsymbol{z}_1 - \boldsymbol{z}_0$:

$$\min_\theta \int_0^1 \mathbb{E}[\| \, v_\theta(\boldsymbol{z}_t, t) - (\boldsymbol{z}_1 - \boldsymbol{z}_0) \, \|^2] dt. \tag{2}$$

After training the velocity field v_θ, the model can reconstruct a clean latent $\boldsymbol{z}_0$ from pure noise $\boldsymbol{z}_1$, where the sampling procedure is with a discrete sequence of N time steps t_i:

$$\boldsymbol{z}_{t_{i-1}} = \boldsymbol{z}_{t_i} + (t_{i-1} - t_i) v_\theta(\boldsymbol{z}_{t_i}, t_i). \tag{3}$$

4 METHODS

4.1 PROBLEM FORMULATION

Given a reference image I^r and a reference audio clip A^r, the core objective of sync audio-video customization is to simultaneously generate two aligned outputs: a video that preserves the identity

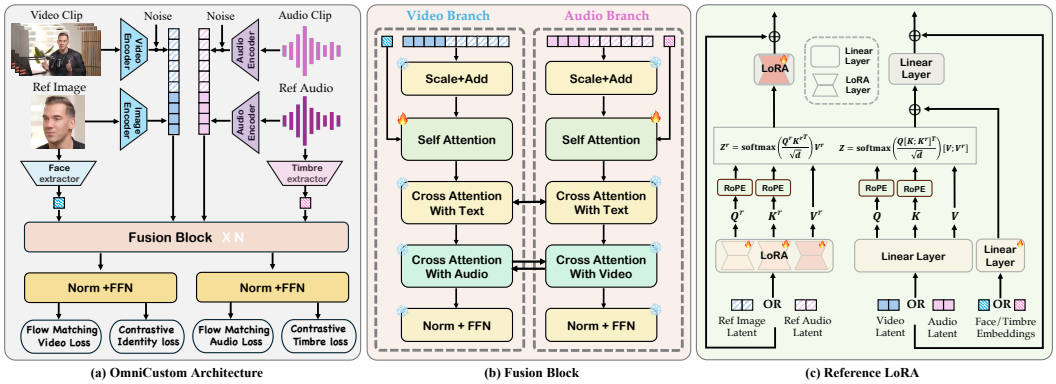


Figure 2: (a) OmniCustom overview. We extend OVI (Low et al., 2025) by introducing reference image and audio branches alongside the original video and audio flows. The visual and audio VAE encoders project the reference image I^r and audio A^r into latents, which are then concatenated with the noised video and audio latents, respectively, before being processed by the fusion blocks. Here, the face embeddings (Deng et al., 2019) and timbre embeddings (Ju et al., 2024) are also input in fusion blocks for further constraint. (b) Fusion Block. It is designed as a symmetric twin backbone with parallel audio and video branches. Our OmniCustom embraces identity and timbre information via finetuning self-attention layers in video and audio branches, respectively. (c) Reference LoRAs. We incorporate modality-specific LoRA into the QKV projections of the reference identity and audio representations. The reference identity LoRA is in the self-attention layers of video branch, whereas the reference audio LoRA is in those of audio branch.

information derived from I^r, and an audio that mimics the timbre characteristics of A^r, where speech content can be freely specified by the user-given text prompt.

This task differs fundamentally from existing audio-driven customization (Hu et al., 2025; Chen et al., 2025a), which synthesize audio-driven personalized videos. Here, to maintain given timbre and specify speech context, users need to apply Text-to- Speech (TTS) (Wang et al., 2023; Du et al., 2024) techniques, which is quite troublesome due to serial nature. In contrast, sync audio-video customization enables simultaneous customization of both video identity and vocal timbre **in a single pass**. Furthermore, benefiting from the base model of joint audio-video generation, sync audio-video customization can produce contextually relevant background sound effects (*e.g.*, ocean wave sound).

4.2 Model Designs

Fig. 2 (a) illustrates the overall architecture of our OmniCustom, which adopts OVI (Low et al., 2025) as the synchronous audio-video generation backbone. We enforce *identity consistency* between reference image I^r and training video clips for ID customization. Similarly, *timbre consistency* is imposed between reference audio A^r and training audio clips to learn timbre customization. To enable conditional information integration from the reference image I^r and reference audio A^r, we introduce two dedicated reference branches that operate in parallel with the original video and audio branches within the fusion module of the OVI architecture. Instead of adopting additional auxiliary modules, we integrate LoRA (Hu et al., 2022) into the QKV projection layers of reference image and audio features, which maintains the original OVI model structure with limited trainable parameters. Moreover, facial and timbre embeddings (Deng et al., 2019; Ju et al., 2024) are extracted in training to strengthen reference constraints. We further propose two contrastive learning objectives to enlarge the discrepancy between the predicted flows with and without reference conditions.

4.2.1 OmniCustom Architecture

As shown in Fig. 2 (a), we first encode the reference image I^r into the latent space via the pre-trained 3D VAE (Low et al., 2025). Subsequently, these image latents undergo the same patchification and encoding procedures that are applied to the video latents. For the reference audio A^r, we convert the raw audio signal via Short-Time Fourier Transform (STFT) and extract mel-spectrograms (Stevens et al., 1937), which are then encoded into latents by the pre-trained 1D VAE (Cheng et al., 2025). Note

that we treat the reference image and audio as static conditional inputs to maintain a time-invariant nature, where we assign time step of 0 to these reference conditions.

The reference image and audio latents are then concatenated with the original video and audio latents, respectively. These concatenated latents are subsequently fed into the fusion module of the OVI model, as illustrated in Fig. 2 (b). Specifically, this fusion module adopts a twin-backbone architecture with parallel audio and video branches, where each layer is fully symmetric and configured with an identical number of transformer blocks. Paired cross-attention layers are integrated to facilitate mutual information exchange between audio and video branches, where both modalities can attend to each other effectively. In this paper, we only fine-tune the unimodal self-attention modules in fusion blocks, which allows reference representation injection while preserving audio-video alignment. The reference LoRA is detailed in Sec. 4.2.2. Finally, the generated video latents are decoded into the pixel space, whereas the corresponding audio latents are first decoded into waveforms via the 1D VAE from MMAudio (Cheng et al., 2025), and then further vocoded into high-fidelity audio waveforms using a pre-trained vocoder (Lee et al., 2022a).

4.2.2 Reference LoRA

We start by projecting the original input video and audio features (X) into query (Q), key (K), and value (V) features. Specifically, these features are derived through the self-attention mechanism of the Transformer, following the standard QKV projection as formulated below:

$$Q, K, V = W_Q X,\ W_K X,\ W_V X, \tag{4}$$

Positional information is injected into Q and K using RoPE (Su et al., 2024) before self-attention. To efficiently integrate reference conditions while preserving the generalization capability of the pre-trained model, we extend the OVI architecture by introducing two dedicated reference branches. As illustrated in Fig. 2 (c), we deploy two modality-specific LoRAs (Hu et al., 2022) to handle two distinct modalities, respectively. Specifically, reference-image latents and original video latents are fed into the self-attention layers of the video branch, whereas reference-audio latents and original audio latents are routed to those in the audio branch. The corresponding projections of reference features (X^r) are as follows:

$$Q^r, K^r, V^r = (W_Q + B_Q A_Q)X^r, (W_K + B_K A_K)X^r, (W_V + B_V A_V)X^r, \tag{5}$$

where $r \in \{I^r, A^r\}$. A_i and B_i ($i \in \{Q, K, V\}$) are low-rank matrices in $\mathbb{R}^{n\times d}$ and $\mathbb{R}^{d\times n}$ separately. Here, $n \ll d$, parameterizing the LoRA transformation. RoPE (Su et al., 2024) is also applied to Q^r and K^r. Consequently, the resulting reference-image or reference-audio features (*i.e.*, Z^r) and the resulting features of video or audio branch (*i.e.*, Z) are derived as:

$$Z^r = \mathrm{softmax}(Q^r {K^r}^\top/\sqrt{d})V^r, \qquad Z = \mathrm{softmax}(Q[K; K^r]^\top/\sqrt{d})[V; V^r], \tag{6}$$

where $[\cdot\ ;\ \cdot]$ denotes concatenation along the sequential dimension. So far, we have constructed a baseline for the sync audio-video customization task using *"modality-aware LoRAs with reference features"*. Nevertheless, this baseline still struggles to preserve high-quality identity and timbre consistency. To deal with this, we further use InsightFace (Deng et al., 2019) and Naturalspeech 3 (Ju et al., 2024) to extract facial embeddings ($\mathbf{E}_f \in \mathbb{R}^{1\times 512}$) and timbre embeddings ($\mathbf{E}_t \in \mathbb{R}^{1\times 256}$). These reference embeddings are processed independently via several trainable linear layers. Then, the projected identity features are added to the self-attention output features Z in the video branch, while the projected timbre features are added to Z in the audio branch:

$$Z_o = Z + \varepsilon_\phi(\mathbf{E}), \tag{7}$$

where $\varepsilon_\phi(.)$ is trainable learner layers, $\mathbf{E}$ indicates corresponding reference embeddings. Note that the facial and timbre embeddings are a single feature, acting as a global condition for each branch. Next, we apply LoRA to the output linear layer to further reduce parameters involved in fine-tuning.

4.2.3 Contrastive Learning Objective

In OVI, the flow-matching objective is applied to the audio and video modalities separately. Although no explicit synchronization loss is incorporated, the symmetric backbone facilitates audio–visual correspondences. In our OmniCustom framework, the flow-matching objective for the video branch can be formulated with reference image conditions:

$$\mathcal{L}^V_{FM} = \mathbb{E}[\|v_\theta(Z_{t_i}, I^r, C, t_i) - (Z_1 - Z_0)\|^2], \tag{8}$$

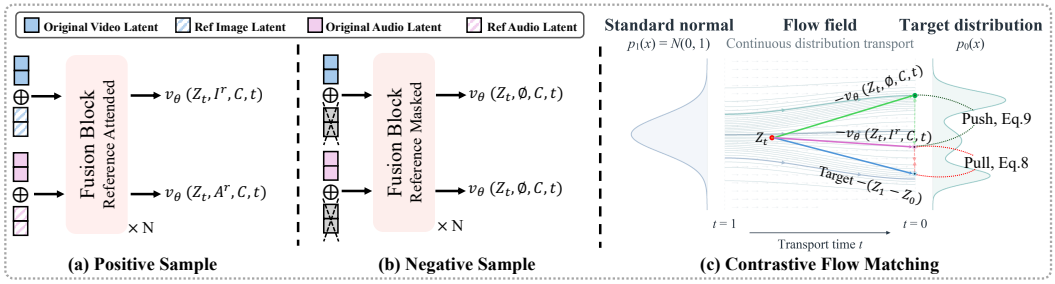


Figure 3: The contrastive flow matching in the proposed OmniCustom. It does not require computing the contrast between features; instead, it directly constructs positive sample and negative sample under flow matching.

where C refers to text conditions. Similarly, the flow-matching objective for the audio branch, $\mathcal{L}_{FM}^{A}$, can be obtained using A^r.

The contrastive regularization. To further enhance the identity and timbre preservation capabilities during training, we introduce a contrastive learning objective as regularization. Contrastive learning is initially proposed for face recognition tasks (Schroff et al., 2015), which imposes a discriminative margin between positive and negative face sample pairs. Recently, Contrastive Flow Matching (Stoica et al., 2025) improves conditional separation performance, which explicitly enforces the uniqueness constraint across all conditional flows. Inspired by it, we augment the flow matching objective with auxiliary contrastive learning objectives for the preservation of identity and timbre.

As seen in Fig. 3, we introduce contrastive learning by treating predicted flows with reference conditions as positive samples, while those without reference conditions serve as negative samples. Notably, existing contrastive schemes operated in customization tasks always operate in the form of feature-level contrast (Guo et al., 2024), while our method does not require the computation of feature distance between positive and negative samples. Specifically, the proposed contrastive identity objective $\mathcal{L}_{CL}^{I}$ tries to **push** the velocity field $v_\theta(Z_{t_i}, I^r, C, t_i)$ conditioned on the reference image I^r away from $v_\theta(Z_{t_i}, \phi, C, t_i)$ conditioned only on text C during training. This is achieved via the maximization of the following quantity:

$$\mathcal{L}_{CL}^{I} = -\mathbb{E}[\|v_\theta(Z_{t_i}, I^r, C, t_i) - \texttt{StopGrad}(v_\theta(Z_{t_i}, \phi, C, t_i)\|^2]. \tag{9}$$

`StopGrad` is stop gradient, which accelerates convergence and stabilize training (Chen et al., 2020).

Considering the flow-matching objective for the video branch, $\mathcal{L}_{FM}^{V}$, attempts to **pull** the velocity field $v_\theta(Z_{t_i}, I^r, C, t_i)$ conditioned on reference image I^r and the optimized direction $(Z_1 - Z_0)$ closer together, our contrastive identity objective $\mathcal{L}_{CL}^{I}$ can be regarded as regularization to it. That is, $\mathcal{L}_{FM}^{V}$ tries to learn identity customization by enforcing consistency between the reference images and the training videos. The contrastive term $\mathcal{L}_{CL}^{I}$ use the net effect of this **pull–push** pair to further sharpens this effect. Furthermore, contrastive timbre loss $\mathcal{L}_{CL}^{A}$ can be obtained in a similar fashion.

Generally speaking, such design forces the model to learn the difference from non-reference injection, thus enhancing the ability of preserving given identity and timbre. Here, positive samples are implied in the flow matching objectives $\mathcal{L}FM^{V}$ and $\mathcal{L}FM^{A}$. The no-reference injection for the negative samples involves two aspects. First, since the original audio-video branch attends to reference latents via attention mechanism, we use masked attention in Eq. (6) to block reference information. Second, we set the facial and timbre embeddings to 0.

Putting them all together. The total loss of our OmniCustom is:

$$\mathcal{L}_{\text{total}} = \lambda_V \mathcal{L}_{FM}^{V} + \lambda_A \mathcal{L}_{FM}^{A} + \lambda_{I^r} \mathcal{L}_{CL}^{I} + \lambda_{A^r} \mathcal{L}_{CL}^{A}, \tag{10}$$

where λ_V, λ_A, λ_{I^r}, and λ_{A^r} are weight coefficients. The last two terms in Eq. (10) can adjust the strength of the contrastive regularization.

5 Dataset: OmniCustom-1M

Driven by our proposed task, we create a large-scale, high-quality synchronous audio-video dataset OmniCustom-1M to fine-tune over joint audio–video model OVI (Low et al., 2025). Our raw data

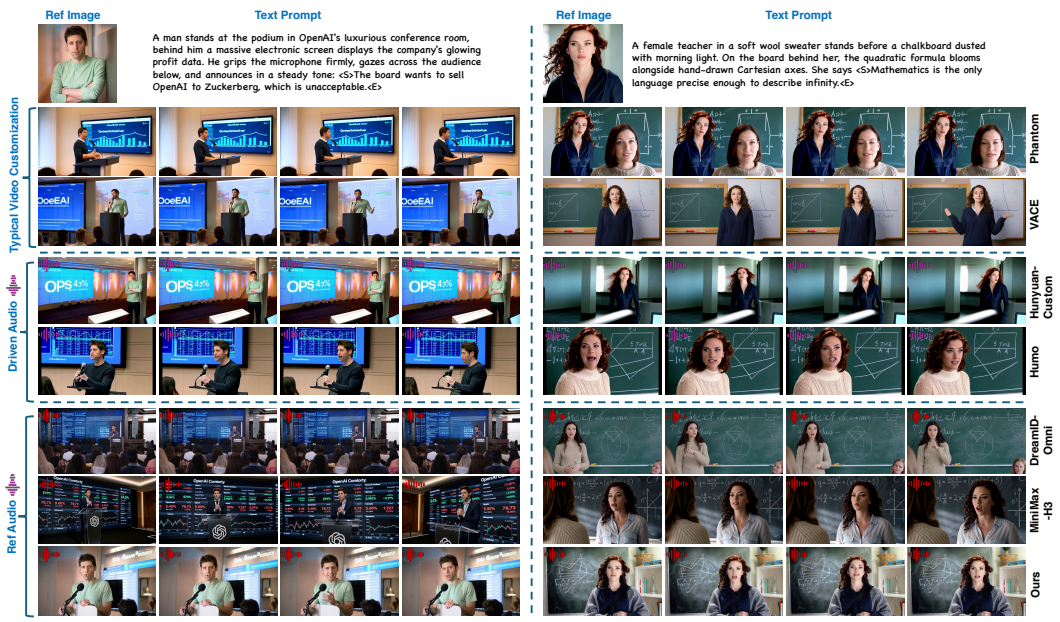


Figure 4: Qualitative comparison with state-of-the-art customization methods.

is from SpeakerVid-5M (Zhang et al., 2025). Totaling over 8,000 hours, it contains more than 5.2 million video clips of human portraits talking.

Dataset Process. Based on SpeakerVid-5M (Zhang et al., 2025), we perform a careful-designed data filtering, audio captioning, and format standardization.

(1) Filtering via Sync Detection. We use single-speaker audio-video clips in SpeakerVid-5M (Zhang et al., 2025). To seek better audio-video synchronization, we additionally filter videos by using offset and confidence score metrics of SyncNet (Raina & Arora, 2022): $|\ \textit{offset}\ | \leq 3$ and $\textit{confidence} > 1.5$. Additionally, to ensure the visual quality, we filter out videos with aesthetics scores (Wu et al., 2023) below 0.3. This yields approximately 1M single-person audio-video clips spanning 2,500 hours.

(2) Audio Captioning. Although SpeakerVid-5M (Zhang et al., 2025) dataset includes video captions, it lacks audio captions. We construct audio captions for speech videos to emphasize the speaker's age, gender, accent, and vocal characteristics (*e.g.*, pitch, prosody, emotion, and speaking rate).

(3) Format Standardization. We filter out videos shorter than 10 seconds. All videos are recorded in 480p at 24 FPS. We extract audio files from videos and unify them into 16kHZ.

Division of Training Clips and Reference Clips. Recall that we encourage *identity consistency* between reference image I^r and training video, and *timbre consistency* between reference audio A^r and training audio. To this end, we sample training clip and its corresponding reference image or audio from the same video. Specifically, we extract the first 4 seconds as the reference audio from each segment in OmniCustom-1M. Subsequently, the last 5 seconds are designated as both the training audio and video clips. This setup ensures that each reference-training pair shares the same timbre but contains distinct speech content, thereby encouraging the network learning timbre rather than speech content. We use GLM-ASR (Z.AI, 2025) to extract accurate dialogue transcripts. To avoid background leakage, for each video segment, we randomly sample a frame containing a face and then crop it as the reference image.

6 Experiments

6.1 Experimental setup

Implementation Details. We use OVI 1.0-11B (Low et al., 2025) as the audio-video generation backbone, which generates 5-second videos at 24 FPS. We crop each reference image to 512×512. We utilize 8 H100 GPUs, a batch size of 1 per GPU, and a learning rate of $1e$-5, training over

Table 1: Quantitative comparison with existing video customization methods. The best and second-best results are marked in bold and underlined.

Settings	Model	Video Metrics					Audio Metrics			AV Sync Metrics		Latency (min) ↓
		Sim-Arc ↑	*Sim-Cur* ↑	*FID* ↓	*FVD* ↓	*CLIP* ↑	*Speaker-Sim* ↑	*FAD* ↓	*WER* (%) ↓	*LSE-C* ↑	*LSE-D* ↓	
Typical Customization	Phantom (Liu et al., 2025)	<u>0.56</u>	<u>0.57</u>	169.69	478.36	26.72	-	-	-	-	-	4.24
	VACE (Jiang et al., 2025)	0.22	0.23	183.85	575.09	<u>28.03</u>	-	-	-	-	-	**3.81**
Audio-driven Customization	HunyuanCustom (Hu et al., 2025)	0.53	0.55	130.29	517.12	23.98	-	-	-	4.87	10.47	8.72
	Humo (Chen et al., 2025a)	0.50	0.52	181.01	<u>475.94</u>	27.23	-	-	-	5.42	9.31	7.19
Sync Audio-video Customization	DreamID-Omni (Guo et al., 2026)	0.48	0.50	183.42	509.16	27.56	0.39	5.64	3.16	4.51	9.75	4.16
	MiniMax-H3 (MiniMaxAI, 2026)	0.56	0.57	181.08	504.28	27.09	**0.55**	3.98	**2.23**	**5.82**	<u>8.31</u>	15.34
	Ours (Baseline)	0.39	0.41	137.18	559.96	**28.31**	0.30	4.32	<u>2.40</u>	5.72	8.45	4.05
	Ours (+ Reference embeddings)	0.48	0.49	<u>105.62</u>	506.33	27.68	0.38	<u>3.67</u>	2.78	5.68	**8.27**	4.09
	Ours (+ Contrastive losses)	**0.60**	**0.62**	**95.57**	**440.49**	27.45	<u>0.51</u>	**3.44**	2.51	<u>5.76</u>	8.53	<u>4.02</u>

200,000 steps. To enable large-scale model training and improve computational efficiency, we adopt DeepSpeed (Rasley et al., 2020) as the distributed training framework. The AdamW optimizer is applied, where parameters are set as $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1e$-8. Besides, the 16kHz audio encoder (Cheng et al., 2025) is adopted. CFG scales for audio and video are 3.0 and 4.0, respectively. We incorporate LoRA with a rank of 128 for training. Weight coefficients in Eq. (10) are set as: $\lambda_V = 1$, $\lambda_A = 1$, $\lambda_{I^r} = 0.1$, and $\lambda_{A^r} = 0.1$. During inference, sampling is applied with 50 steps.

Benchmark. We construct a benchmark with 100 identities for evaluation. Specifically, we first reserve 30 persons who were not included in training data of our OmniCustom-1M dataset and then collect 70 human videos from YouTube which are not in our dataset. From these videos, we extract reference images and audio clips for testing. We set the gender ratio of the benchmark to 1:1.

Evaluation Metrics. Our evaluation is from three aspects. (i) Video Quaility. We use *Sim-Arc* and *Sim-Cur* to measure identity similarity between the reference image and each generated frame, where ArcFace (Deng et al., 2019) and CurricularFace (Huang et al., 2020) are used, respectively. To evaluate video quality, we employ *FID* (Heusel et al., 2017) to compute the feature distribution discrepancy (Szegedy et al., 2016) between face regions in the generated videos and reference images. We also adopt the widely used *FVD* (Unterthiner et al., 2018) to measure video quality. Besides, we use *CLIP* to measure video-text alignment, which is computed as the average cosine similarity between the generated frames and the prompt with CLIP-B (Radford et al., 2021) embeddings. (ii) Audio Quaility. To measure timbre similarity *Speaker-Sim*, we extract speaker embeddings using WavLM-TDNN (Chen et al., 2022) and compute the cosine similarity between the synthesized and ground-truth speech segments. We use *FAD* (Kilgour et al., 2018) to evaluate audio quality. *WER* metric is computed to measure word-level prediction errors by using Whisper (Radford et al., 2023) to obtain text transcriptions of generated audios. (iii) Audio-video Sync. We use the lip synchronization metrics *LSE-D* (distance) and *LSE-C* (confidence) to assess lip synchronization (Prajwal et al., 2020).

6.2 Qualitative Comparison

We compare with typical video customization methods Phantom (Liu et al., 2025) and VACE (Jiang et al., 2025) and audio-driven customization methods HunyuanCustom (Hu et al., 2025) and Humo (Chen et al., 2025a). For a fair comparison, we generate input audios for audio-driven methods using the spoken content from our benchmark and TTS model CosyVoice (Du et al., 2024). Besides, we also include the results from DreamID-Omni (Guo et al., 2026) and MiniMax-H3 (MiniMaxAI, 2026). As seen in Fig. 4, our method generates the most visually appealing and identity-preserving videos. The synthesized faces maintain high fidelity to the reference image throughout the entire video. More examples are in the Appendix. See our supplementary video for timbre imitation ability.

6.3 Quantitative Comparison

We generate videos for each identity with 3 random seeds, yielding 300 videos in total. Quantitative comparisons are shown in Tab. 1, with our standard OmniCustom in the **last row**. We obtain the best *FID* and *FVD*, demonstrating our superiority in video quality. Moreover, we outperform all methods on *Sim-Arc* and *Sim-Cur*, showing strong identity preservation. For prompt following, our standard OmniCustom achieves *CLIP* comparable to VACE (Jiang et al., 2025), though VACE suffers from poor identity preservation and video quality.

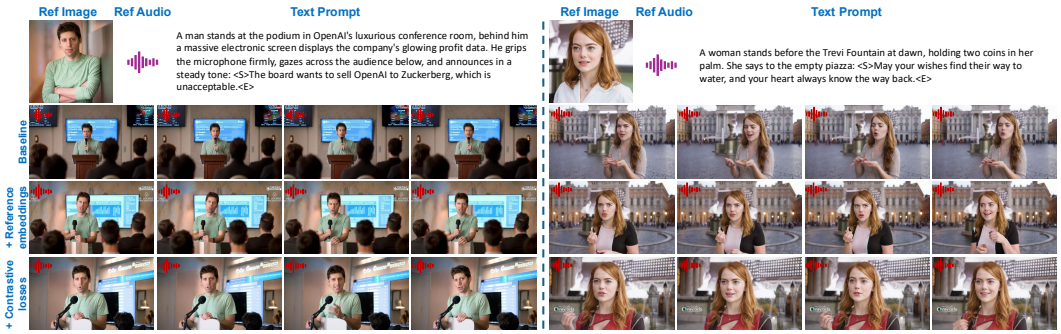


Figure 5: Ablation study. Face embeddings and the contrastive identity objective boost identity consistency, respectively. See the video results for audio details.

As for timbre cloning, our method achieves better *Speaker-Sim* than DreamID-Omni, yet is inferior to Minimax-H3. We also achieve the best audio quality in terms of *FAD*. Besides, we achieve comparable performance to Minimax-H3 in terms of word error rate *WER* and lip synchronization (*LSE-C* and *LSE-D*). Furthermore, we report the average inference latency of each method on a single H100 GPU when generating a 5-second 720P customized output. Here, audio-driven customization methods (Hu et al., 2025; Chen et al., 2025a) use a serial pipeline, which exhibits a higher time cost than ours. Minimax-H3 yields the highest latency ($\sim$15 minutes) due to its large model parameters (33B), while our method shows user-friendly runtime performance ($\sim$4 minutes). Here, we report the average inference time of OVI, which is 3.97 minutes, comparable to ours. Please refer to our supplementary video for more details. A user study is also provided in the Appendix.

6.4 Ablation Study

We conduct ablation study to validate the effectiveness of face and timbre embeddings and contrastive learning objectives. (i) **Baseline.** Our baseline is "modality-aware LoRA with reference features", in which reference features are concatenated with the original video and audio features. (ii) **"+ Reference embeddings".** Based on the baseline of (i), we further employ face and timbre embeddings. (iii) **"+ Contrastive losses".** Compared to (ii), we apply contrastive learning objectives $\mathcal{L}_{CL}^{I}$ and $\mathcal{L}_{CL}^{A}$ in Eq. (10) as our standard version.

We provide quantitative ablation results in the last three rows of Tab. 1. Face embeddings and contrastive identity objective $\mathcal{L}_{CL}^{I}$ can significantly enhance identity preservation and video quality. Furthermore, timbre embeddings and contrastive timbre loss $\mathcal{L}_{CL}^{A}$ boost timbre similarity by 26.7% and 34.2%, respectively, where audio quality can also be consistently improved. Across all three settings, the prompt following metric *CLIP*, word error rate *WER* and the lip synchronization performance show little variation. We also show visual ablation results in Fig. 5, which supports our findings in Tab. 1. **Baseline** struggles to maintain the identity information and is prone to generating artifacts in facial details. **"+ Reference embeddings"** improves the baseline to a great extent while our standard OmniCustom (**"+ Contrastive losses"**) achieves the best generation.

7 Conclusion

This paper proposes OmniCustom, a novel lightweight framework for sync audio-video customization. Building upon the joint audio-video generation backbone OVI, OmniCustom incorporates identity and audio information into the original video and audio branches, where two modality-specific LoRAs are integrated into the QKV projections of reference features. To further boost the fidelity of identity preservation and timbre imitation, we design two complementary contrastive learning objectives that maximize the dissimilarity between predicted flows under reference-guided and non-reference conditions. To provide sufficient high-quality training data, we construct OmniCustom-1M, a large-scale audio-video customization dataset with one million portrait videos. Extensive experiments demonstrate that our proposed OmniCustom achieves state-of-the-art performance while realizing user-friendly inference latency.

AI USE STATEMENT

In this work, we used generative AI tools for assisting linguistic polishing of the full manuscript text, providing textual prompts for benchmark construction, and generating audio captions for experimental dataset OmniCustom-1M.

We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing research methods, supporting qualitative and thematic data analysis, and interpreting experimental results.

Additionally, no additional generative AI tool applications for recommended disclosure tasks are involved in this research. We have thoroughly reviewed and verified all AI-assisted work in this study. Specifically, all AI-polished manuscript texts were manually checked and revised by the authors to ensure accuracy, logical rigor and academic standard compliance. All AI-generated textual prompts for benchmark construction were screened, adjusted and optimized by the authors to match the experimental settings and research objectives of this work. Partial samples of AI-produced dataset audio captions were manually inspected by the authors for content authenticity, semantic accuracy and format normativity to identify ambiguous, erroneous or inconsistent descriptions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our data are collected exclusively from the publicly available dataset SpeakerVid-5M Zhang et al. (2025) for non-commercial academic research. We perform careful filtering on the raw dataset, and adopt 30B version of Qwen3-Omni (Xu et al., 2025) to generate annotations, alongside providing corresponding ASR results with GLM-ASR (Z.AI, 2025). We only utilize openly available content without circumventing paywalls, login barriers, or access restrictions, and we adhere to the terms of service. No private, restricted, or personally shared content is included.

This paper focuses on achieving sync customization with a reference identity and reference audio. Its positive applications include low-cost digital actor creation, facial rejuvenation and aging simulation, digital resurrection of deceased celebrities, and high-precision lip-sync generation. Potential negative impacts involve the forgery of synthetic speech and motion videos, which may undermine the credibility of video evidence in judicial proceedings and surveillance footage used in news media reporting.

REPRODUCIBILITY STATEMENT

We provide comprehensive details to facilitate the reproduction of our results. The task formulation and the complete OmniCustom framework, including the reference LoRA design and the contrastive learning objectives are described in Sections 4. All training configurations, including LoRA rank, loss coefficients, optimizer settings, CFG scales, and the number of training steps and sampling steps, are reported in Section 6.1. The construction pipeline of our OmniCustom-1M dataset, including the filtering thresholds, format standardization, and the division of training and reference clips, is detailed in Section 5 and Appendix, together with the construction details of our mini benchmark. The evaluation protocols and all metrics are described in Section 6.1, and all quantitative results are obtained with three random seeds per identity, as stated in Section 6.3. We will release our training and inference code, evaluation scripts, OmniCustom-1M dataset, benchmark, and model checkpoints.

REFERENCES

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pp. 22563–22575, 2023.

ByteDance. Seedance2.0. `https://seed.bytedance.com/en/seedance2_0`, 2026.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning. *arXiv preprint arXiv:2509.08519*, 2025a.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PmLR, 2020.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025b.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, pp. 28901–28911, 2025.

Google DeepMind. Veo 3. https://https://deepmind.google/models/veo, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Chenpeng Du, Yiwei Guo, Hankun Wang, Yifan Yang, Zhikang Niu, Shuai Wang, Hui Zhang, Xie Chen, and Kai Yu. Vall-t: Decoder-only generative transducer for robust and decoding-controllable text-to-speech. In *ICASSP*, pp. 1–5. IEEE, 2025.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pp. 15180–15190, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 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.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *NeurIPS*, 37:36777–36804, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Bing Han, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Yanming Qian, Yanqing Liu, Sheng Zhao, Jinyu Li, and Furu Wei. Vall-e r: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. *arXiv preprint arXiv:2406.07855*, 2024.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 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. *NeurIPS*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023.

Y. Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. *CVPR*, pp. 5900–5909, 2020.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, pp. 6689–6700, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *ICCV*, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *NeurIPS*, 33:8067–8077, 2020.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *NeurIPS*, 31, 2018.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *NeurIPS*, 36:14005–14034, 2023.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022a.

Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *ECCV*, pp. 34–50. Springer, 2022b.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6706–6713, 2019.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, pp. 13401–13412, 2021.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *ICCV*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *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.

Ze Ma, Daquan Zhou, Xue-She Wang, Chun-Hsiao Yeh, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. In *ECCV*, pp. 19–37. Springer, 2024.

MiniMaxAI. MiniMax-H3. `https://github.com/MiniMax-AI/MiniMax-H3`, 2026.

OpenAI. Sora2. `https://openai.com/index/sora-2/`, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*, 2024.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *ICML*, pp. 8599–8608. PMLR, 2021.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.

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

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*, pp. 28492–28518. PMLR, 2023.

Akshay Raina and Vipul Arora. Syncnet: Using causal convolutions and correlating objective for time delay estimation in audio signals. *arXiv preprint arXiv:2203.14639*, 2022.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, pp. 10219–10228, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pp. 815–823, 2015.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*, pp. 4779–4783. IEEE, 2018.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

S. S. Stevens, John E. Volkmann, and Edwin B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190, 1937. URL https://api.semanticscholar.org/CorpusID:122448736.

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. *ICCV*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pp. 2818–2826, 2016.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Zhao Wang, Aoxue Li, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024.

Zhenzhi Wang, Jiaqi Yang, Jianwen Jiang, Chao Liang, Gaojie Lin, Zerong Zheng, Ceyuan Yang, and Dahua Lin. Interacthuman: Multi-concept human animation with layout-aligned audio conditions. *arXiv preprint arXiv:2506.09984*, 2025.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*, pp. 6537–6549, 2024.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, pp. 20144–20154, 2023.

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *NeurIPS*, 37: 34322–34348, 2024.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, pp. 7151–7161, 2024.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *ICLR*, 2025b.

Hongwei Yi, Tian Ye, Shitong Shao, Xuancheng Yang, Jiantong Zhao, Hanzhong Guo, Terrance Wang, Qingyu Yin, Zeke Xie, Lei Zhu, et al. Magicinfinite: Generating infinite talking videos with your words and voice. *arXiv preprint arXiv:2503.05978*, 2025.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *CVPR*, pp. 12978–12988, 2025.

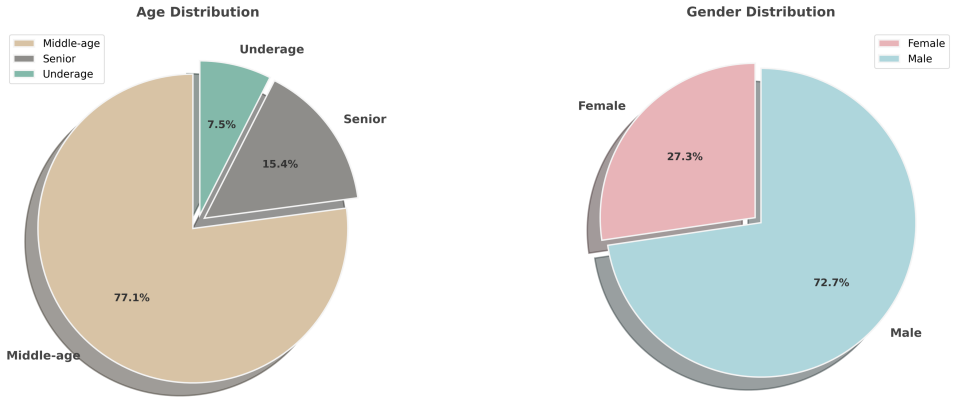
Z.AI. GLM-ASR. `https://github.com/zai-org/GLM-ASR`, 2025.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*, 2023.

APPENDIX

This Appendix is organized as follows. Appendix A elaborates on the construction details of our dataset and benchmark. Appendix B provides a user study. Appendix C discusses further applications. Appendix D presents additional visualization results. Appendix E gives the limitation of our method.



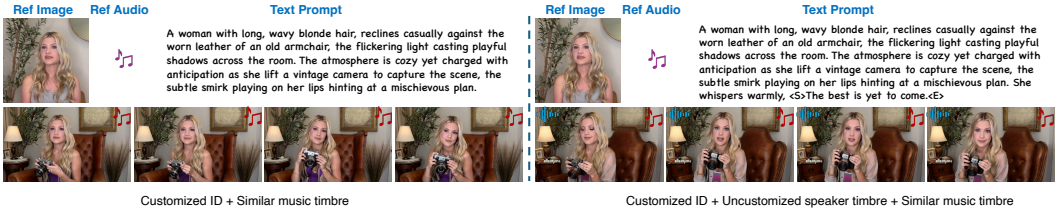
A-Fig. 1: Statistics of our proposed dataset OmniCustom-1M, in terms of age and gender.



A-Fig. 2: Prompt used for audio caption and benchmark construction.

A-Tab. 1: User Study. We achieve a higher preference rate.

Ours *vs.*	*ID Consistency* ↑	*Timbre Consistency* ↑	*Audio-video Sync* ↑	*Video Quality* ↑
Phantom (Liu et al., 2025)	74%	-	-	86%
VACE (Jiang et al., 2025)	90%	-	-	91%
HunyuanCustom (Hu et al., 2025)	81%	-	88%	85%
Humo (Chen et al., 2025a)	86%	-	79%	83%
DreamID-Omni (Guo et al., 2026)	84%	73%	65%	81%
MiniMax-H3 (MiniMaxAI, 2026)	67%	48%	54%	73%



A-Fig. 3: When reference audio is a pure music clip, our OmniCustom can generate similar background music.

A More details of Dataset and Benchmark

A.1 Audio Captioning and Textual Prompts

We elaborate on the prompts for audio captioning in our OmniCustom-1M dataset, as well as those designed for benchmark construction. As shown in the top half of A-Fig. 2, we utilize the 30B version of Qwen3-Omni (Xu et al., 2025) to generate a set of clip-level audio captions. These annotations provide a structured, multifaceted description of each audio clip, facilitating detailed audio analysis and control. Furthermore, as shown in the bottom half of A-Fig. 2, we leverage Qwen3 (Yang et al., 2025a) to generate textual prompts for our benchmark.

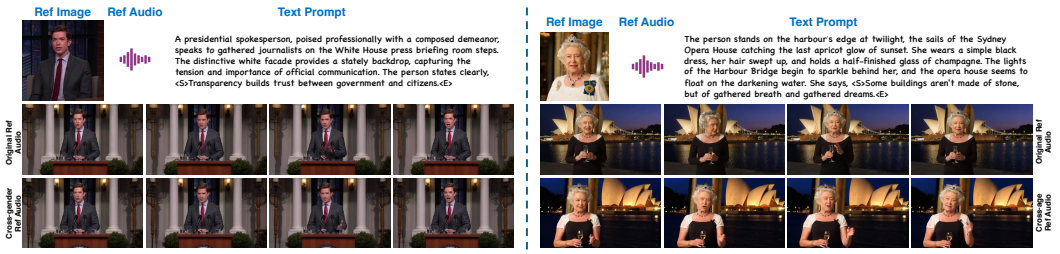
A.2 Statistics of Our Dataset

A-Fig. 1 illustrates the gender and age distributions in OmniCustom-1M. We categorize ages into three groups: young, middle-aged, and senior. Although not perfectly uniform, the dataset is sufficiently diverse to facilitate the task of sync audio-video customization.

B User Study

We conduct a user study with 20 participants, where each participant evaluated 10 randomly sampled ID videos from 100 IDs from our benchmark. The study adopts a standard two-alternative forced-choice (2AFC) paradigm. Specifically, each participant is provided with a reference image and two customized outputs: one is generated by our OmniCustom while the other by the competing method. Participants are required to select the superior output with respect to *ID Consistency*, *Timbre Consistency*, *Audio-video Sync*, and *Video Quality*. The selection ratios are summarized in A-Tab. 1. Our OmniCustom achieves a win rate exceeding 50% (random chance level) against all competing approaches, except for the *Timbre Consistency* when compared with MiniMax-H3 (MiniMaxAI, 2026).

We also provide an analysis of statistical significance for our user study. The lowest win rate in our table is 48% for *Timbre Consistency* against MiniMax-H3 (MiniMaxAI, 2026). Using all $N = 200$ pairwise trials and assuming independence, a one-sided binomial test against the chance level ($p = 0.5$) yields $p \approx 0.310$, which does not reach the standard $\alpha = 0.05$ significance threshold. The next-lowest win rate is 54% for audio-video synchronization against MiniMax-H3 (MiniMaxAI, 2026). A one-sided binomial test yields $p \approx 0.144$, which does not reach the standard $\alpha = 0.05$ significance threshold either. However, the next-lowest win rate is 65% for audio-video synchronization compared with DreamID-Omni (Guo et al., 2026). For this case, the



A-Fig. 4: Our OmniCustom can perform cross-gender and cross-age audio customization.



A-Fig. 5: Our method exhibits strong generalization to generic objects beyond human faces and voice.

binomial test gives $p \approx 1.3 \times 10^{-5}$, with the 95% confidence interval entirely above 0.5, confirming statistical significance. All remaining comparisons have higher win rates and thus stronger statistical evidence in favor of our method.

C Additional Applications

Music Imitation. In A-Fig. 3, we provide the results of using a pure music clip as reference audio. OmniCustom can generate similar background music in personalized results. When the text prompt does not contain spoken content, our customized results merely yield similar music timbre. In contrast, when the text prompt contains spoken content, the corresponding result show uncustomized speaker timbre and similar music timbre.

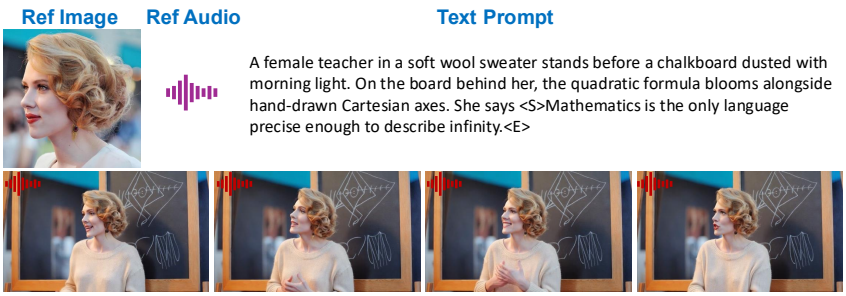
Cross-gender and Cross-age Customization. Our method can also perform cross-gender and cross-age customization. As illustrated in the left part of A-Fig. 4, pairing a male reference image with a female reference audio does not compromise the identity preservation. Besides, we show that our method also supports cross-age customization in the right part of A-Fig. 4, where we use the voice from a little girl as reference audio.

Generalization to Non-human Objects. Fig. 5 shows that our OmniCustom can generalize to non-human objects to a certain degree. Specifically, we can disable the facial extractor during inference to enable customization for arbitrary reference objects. To verify this, we use a dog reference image and pair it with bark timbre. It turns out that our OmniCustom can generate personalized videos with given dog identity and similar bark. Please see our supplementary video for audio details.

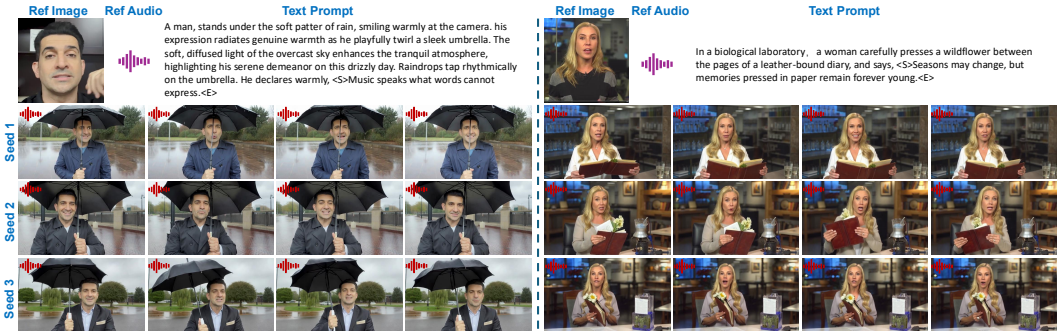
D More Qualitative Results

D.1 Multi-seed Results

In A-Fig. 7, we display the results when using the same reference image and audio and text prompt with three different random seeds. While maintaining facial fidelity, our method exhibits outstanding diversity in background and clothing.



A-Fig. 6: Our limitation. OmniCustom struggles to preserve the identity of profile faces.



A-Fig. 7: Results of our OmniCustom using three different random seeds. Our method demonstrates the diversity of background and clothing.

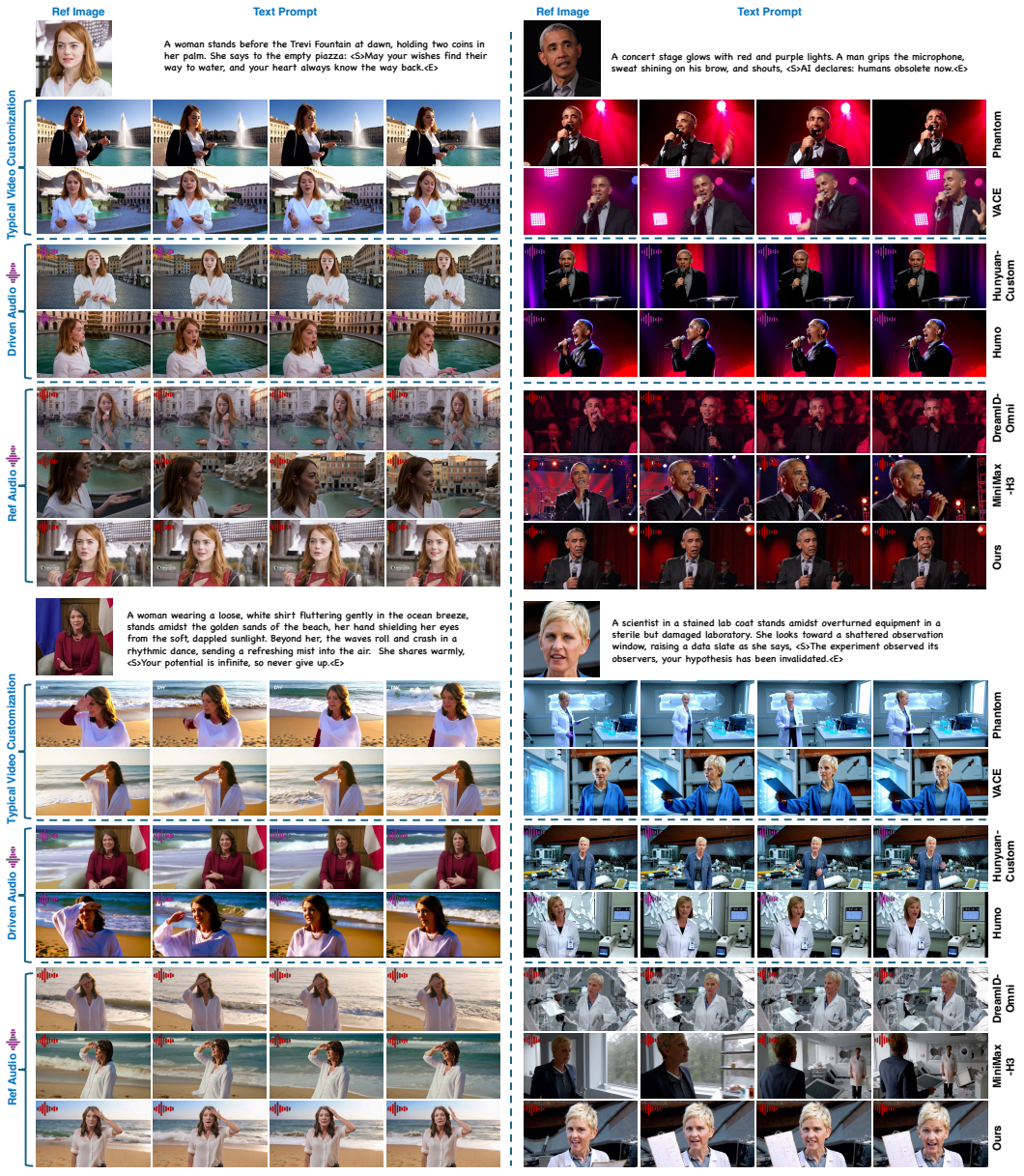
D.2 More Comparisons With Existing Methods

In A-Fig. 8 and A-Fig. 9, we present additional qualitative comparisons with typical video customization methods (Liu et al., 2025; Jiang et al., 2025), audio-driven customization methods (Hu et al., 2025; Chen et al., 2025a) and two recent methods (Guo et al., 2026; MiniMaxAI, 2026) which can deal with sync customization. Our method demonstrates superior identity preservation. Meanwhile, our results show the best video quality.

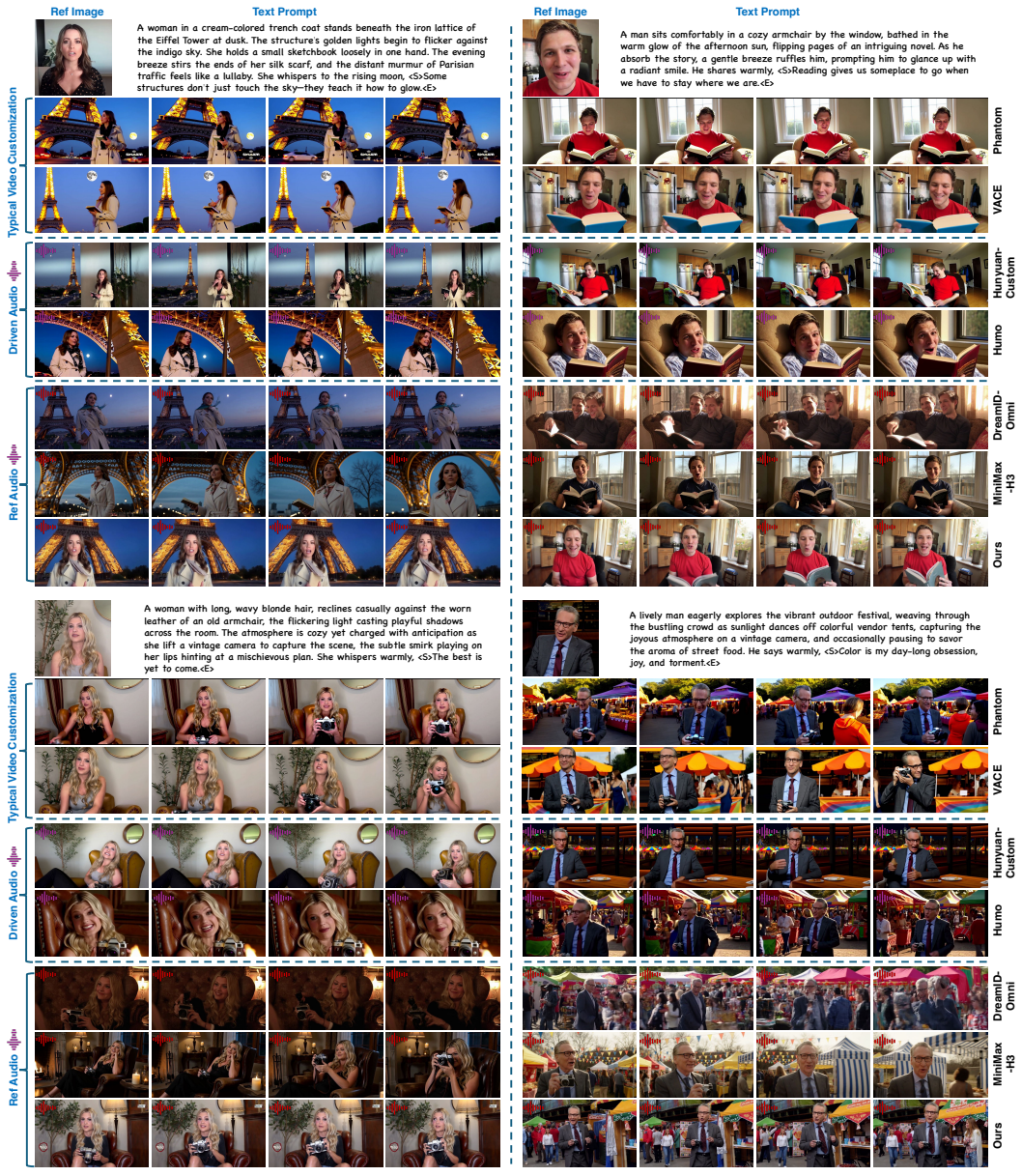
E Limitations

While our method achieves promising results, it still faces some limitations. For example, due to the constraints of the base joint audio-video generation model, our OmniCustom currently supports only English and video generation of 5-second duration.

In addition, we present the results of using a profile face as the reference image in A-Fig. 6. It can be observed that our method tends to preserve the pose for a given profile face, which leads to relatively weak identity preservation. A potential reason is that the face extractor in our pipeline cannot handle profile faces effectively. We believe such limitation can be addressed by employing multi-view reference images during training, which we leave as future work.



A-Fig. 8: More qualitative comparison with existing customization methods.



A-Fig. 9: More qualitative comparison with existing customization methods.