

# ID-LoRA: Identity-Driven Audio-Video Personalization with In-Context LoRA

Aviad Dahan\*, Moran Yanuka\*, Noa Kraicer, Lior Wolf, and Raja Giryes

Tel Aviv University

\* Equal contribution

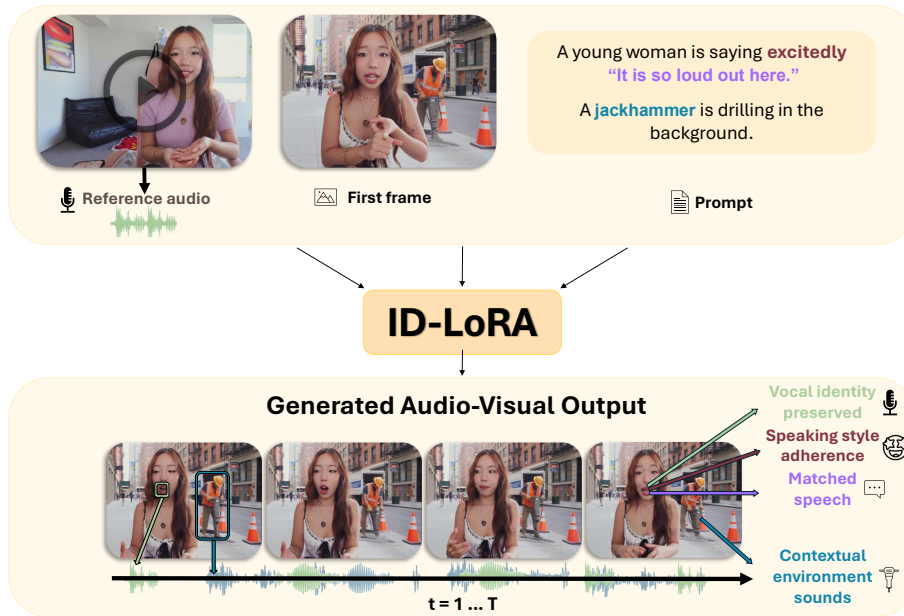
**Abstract.** Existing video personalization methods preserve visual likeness but treat video and audio separately. Without access to the visual scene, audio models cannot synchronize sounds with on-screen actions; and because classical voice-cloning models condition only on a reference recording, a text prompt cannot redirect speaking style or acoustic environment. Although prompt-conditioned audio models could offer such control, they lack access to the visual scene.

We propose **ID-LoRA** (Identity-Driven In-Context LoRA), which jointly generates a subject’s appearance and voice in a single model, letting a text prompt, a reference image, and a short audio clip govern both modalities together. ID-LoRA adapts the LTX-2 joint audio-video diffusion backbone via parameter-efficient In-Context LoRA and, to our knowledge, is the first method to personalize visual appearance and voice within a single generative pass. Two challenges arise from this formulation. Reference and generation tokens share the same positional-encoding space, making them hard to distinguish; we address this with *negative temporal positions*, which place reference tokens in a disjoint region of the RoPE space while preserving their internal temporal structure. Furthermore, speaker characteristics tend to be diluted during denoising; we introduce *identity guidance*, a classifier-free guidance variant that amplifies speaker-specific features by contrasting predictions with and without the reference signal. In human preference studies ID-LoRA is preferred over Kling 2.6 Pro, the leading commercial unified model with voice personalization capabilities, by 73% of annotators for voice similarity and 65% for speaking style. Automatic metrics confirm these gains: on cross-environment settings, speaker similarity improves by 24% over Kling, with the gap widening as reference and target conditions diverge. A preliminary user study further suggests that joint generation provides a useful inductive bias for physically grounded sound synthesis. ID-LoRA achieves these results with only  $\sim 3K$  training pairs on a single GPU. Code, models, and data are available at <https://id-lora.github.io>.

**Keywords:** Audio-Video Generation · Personalization · In-Context LoRA

## 1 Introduction

The field of generative media has witnessed a paradigm shift, moving from static image synthesis to the creation of high-fidelity, temporally coherent video. Indeed, foundational text-to-video (T2V) models, such as the open-source Wan series [33], demonstrate remarkable capabilities in generating complex scenes, and cross-modal personalization frameworks like Phantom [24] enable users to inject specific subject identities into these visual priors with high precision. However, human identity is inherently multimodal: a person is defined not only by their appearance but distinctly by the timbre, cadence, and prosody of their voice. Despite the visual fidelity of modern personalization methods, the resulting content remains predominantly silent, or relies on disjointed audio solutions that fracture the illusion of reality. Current approaches to audio-visual personalization typically rely on *cascaded pipelines*, where video generation is strictly conditioned on previously synthesized audio (*e.g.*, SadTalker [43], VASA-1 [41], Hallo [40]). These modular approaches share a fundamental limitation: the voice-



**Fig. 1: Unified Audio-Visual Personalization with ID-LoRA.** ID-LoRA takes a reference audio clip and a target first frame as input. Unlike cascaded pipelines that treat modalities separately, ID-LoRA jointly generates synchronized video and audio in a single pass. This unified approach allows a text prompt to simultaneously dictate novel content, such as a specific speaking style and environmental acoustics (*e.g.*, “a jackhammer is drilling in the background”), while ensuring the subject’s vocal identity and visual likeness are preserved across the generated sequence. Video examples with audio are available in the supplementary material.

cloning stage conditions only on an audio reference and a transcript, ignoring the text prompt that describes the target scene. As a result, if a prompt requests an angry shout in a windy outdoor setting but the reference audio was recorded in a quiet studio, a cascaded pipeline will simply propagate the studio-like acoustic characteristics and neutral speaking style, failing to follow the prompt’s intent. More broadly, cascaded generation prevents the prompt from jointly influencing both audio and video properties, limiting controllability over environment sounds and speaking style. While recent editing techniques like EditYourself [11] and Just-Dub-It [1] have moved toward unified generation, they are fundamentally constrained to *same-video* settings. They edit an existing video, maintaining the original speaker settings and acoustic environment, and cannot generalize to the *cross-video* settings required to synthesize a subject in a novel context.

In this work, we address *unified audio-visual personalization* (Fig. 1): generating video clips of a specific identity where appearance and voice are synthesized jointly within a shared latent space. Unlike dubbing or editing approaches, our goal is to generalize to entirely new scenes. By operating within a unified latent space, we aim to place a specific subject into entirely *novel* contexts while preserving both their visual likeness and vocal identity. Because the modalities are coupled, the text prompt can simultaneously dictate the scene’s visual generation, environmental acoustics, and speaking style.

To achieve this, we propose **ID-LoRA** (Identity-Driven In-Context LoRA), a method that adapts a unified audio-video diffusion backbone (LTX-2 [15]) for audio-visual personalization. By leveraging a shared DiT backbone that processes audio and video latents, our model mimics the joint distribution of a speaker’s appearance and acoustic signature, enabling identity-preserving generation in novel contexts. Crucially, we generalize the In-Context LoRA approach to allow the model to attend to the reference identity sample (a single first-frame image and a short audio clip) and generate coherent outputs in novel settings. Because both modalities share the same text conditioning, the generated audio can follow prompt descriptions of the environment and speaking style rather than being locked to the reference clip’s acoustics, while simultaneously attending to the video modality.

Our contributions can be summarized as follows:

- We introduce **ID-LoRA**, the first In-Context LoRA framework for zero-shot joint audio-video personalization.
- We propose two architectural components that are key to effective identity transfer: *negative temporal positions* that place reference audio tokens in a negative region of the RoPE positional space, cleanly separating them from target tokens while keeping the target at the pretrained positions; and *identity guidance*, a classifier-free guidance generalization applied to the audio stream that amplifies speaker-specific features at inference time.
- We propose an evaluation protocol with curated splits that test different aspects of audio-visual personalization across multiple environments. We complement automatic metrics with two human evaluation paradigms: an

A/B preference study and a Mean Opinion Score (MOS) study evaluating environment sound interaction across diverse physical scenarios.

We demonstrate through extensive experiments and human evaluations that our unified approach outperforms state-of-the-art cascaded baselines and commercial models in speaker similarity, lip synchronization, and audio prompt adherence.

## 2 Related Work

### 2.1 Joint Audio-Video Generation

Recent work has explored shared diffusion or flow-matching backbones that jointly model video and audio latents. AV-DiT [37] achieves efficient synchronized synthesis using a modality-shared DiT backbone with lightweight modality-specific adapters. JarvisDiT [23] improves fine-grained audio-video synchronization via a hierarchical spatio-temporal synchronized prior, and Apollo [36] unifies multiple text- and image-conditioned generation tasks (T2AV/TI2AV/TI2V) under a single multi-task objective. Ovi [25] adopts a twin-backbone architecture with a dedicated 5B audio branch mirroring the visual backbone of Wan2.2. LTX-2 [15], which serves as our backbone, employs an asymmetric 48-layer DiT with bidirectional cross-modal attention, achieving state-of-the-art joint generation quality. While these models produce temporally coherent audio-visual content, they are designed for general text-to-audio-video synthesis and lack mechanisms for multimodal identity personalization.

### 2.2 Identity Personalization in Video and Audio

*Visual Identity Personalization.* Subject-driven image generation has advanced rapidly through optimization-based methods such as Textual Inversion [12] and DreamBooth [31], and through encoder-based approaches such as IP-Adapter [42] and InstantID [38] that inject identity features into diffusion models. Parameter-efficient fine-tuning via LoRA [16] has become a dominant paradigm, and In-Context LoRA (IC-LoRA) [17] extends it by concatenating reference and target latents along the sequence dimension, enabling zero-shot identity transfer through self-attention without per-subject optimization. Video personalization builds on these foundations: VACE [19] provides an all-in-one video creation framework, and Phantom [24] enables cross-modal subject insertion with high fidelity. Our work elevates the IC-LoRA paradigm from a visual-only technique to a comprehensive audio-visual framework. Because cross-modal identity transfer introduces severe positional entanglement during denoising, we formulate negative temporal positions and propose identity guidance, successfully handling the unique challenges of cross-modal personalization.

*Audio Identity and Scene Control.* Zero-shot voice cloning has progressed from speaker-embedding-conditioned TTS to neural codec language models [26, 34], flow-based approaches [7]), and masked generative codec models [39], that achieve



high-fidelity speaker transfer. However, these models generate clean speech with no control over background sounds or acoustic environment. Text-to-audio models such as AudioLDM 2 [22] produce rich environmental soundscapes from free-text descriptions but cannot preserve a specific speaker’s identity. Audiobox [32] bridges this gap by combining voice cloning with text-described background sounds, yet it operates purely in the audio domain without access to the visual scene. Our unified approach goes further: by generating audio and video jointly, a single text prompt controls visual content, environment sounds, and speaking style simultaneously, while the visual scene provides implicit conditioning for acoustic properties such as timing and reverberation. Throughout, the speaker’s vocal identity is preserved via the reference audio.

*Audio-Driven Talking Heads.* A related line of work generates talking-head video conditioned on a driving audio signal [40, 41, 43]. These methods animate a source portrait to match input speech, producing lip-synchronized results, but they require *pre-existing* audio rather than generating it, and typically operate on tightly cropped face regions without modeling the broader scene or environment sounds. Our setting is fundamentally different: we jointly *generate* both the audio and the video of a subject in a novel, prompt-specified context.

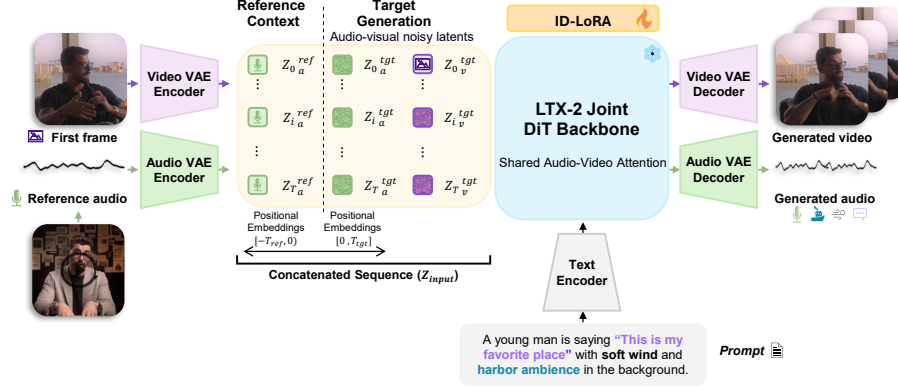
### 2.3 Concurrent Work

*Audio-Visual Editing and Dubbing* Just-Dub-It [1] adapts LTX-2 for transcript-driven video dubbing, EditYourself [11] enables transcript-based editing by inpainting the lip region, and Sync-LoRA [27] propagates visual appearance changes across a video but remains video-only with no audio branch. A common thread is the *same-video* constraint: all three methods modify a source video in place, inheriting its acoustic environment and speaker configuration, and none address *cross-video* personalization where the subject must be synthesized in an entirely new context with potentially different environment sounds and speaking style. ID-LoRA fills this gap by performing fully generative audio-visual personalization from only a reference audio clip, a first-frame image and a text prompt.

*Identity-Aware Audio-Visual Synthesis* DreamID-Omni [14] introduces a controllable framework for human-centric generation including multi-person scenarios, but requires  $\sim 1\text{M}$  training samples and substantial architectural modifications. MM-Sonate [29] trains on  $\sim 100\text{M}$  pairs across generation, editing, and animation tasks. In contrast, ID-LoRA demonstrates that parameter-efficient adaptation with only  $\sim 3\text{K}$  pairs can achieve strong audio-visual personalization, with particular emphasis on enabling the text prompt to govern environment sounds and speaking style.

## 3 Method

Our goal is to generate audio-video content that faithfully preserves both the visual appearance and vocal identity of a target individual, given only a short



**Fig. 2: ID-LoRA (overview).** Target first frame and reference audio are encoded into latents, concatenated with noisy targets, and fed to a shared LTX-2 DiT adapted with In-Context LoRA. Reference audio tokens receive *negative* temporal positions in the RoPE space, cleanly separating them from the target tokens. Joint text conditioning yields synchronized, identity-preserving video and audio whose acoustics follow the prompt and adapted to the new environment.

reference audio and a first frame. As illustrated in Fig. 2, we achieve this through **ID-LoRA** (Identity-Driven In-Context LoRA), which adapts a pre-trained joint audio-video diffusion backbone to perform identity transfer within a unified latent space.

### 3.1 Preliminaries: Joint Audio-Video Diffusion

We build upon LTX-2 [15], a joint audio-video generation model based on the Diffusion Transformer (DiT) architecture. LTX-2 employs an asymmetric dual-stream transformer with 48 layers that jointly processes video and audio latents through bidirectional cross-modal attention. The video stream (14B parameters) handles spatiotemporal dynamics using 3D RoPE positional encoding, while the audio stream (5B parameters) processes temporal audio features using 1D RoPE.

Both modalities are encoded into latent spaces: video pixels are compressed via a Video VAE ( $32 \times$  spatial,  $8 \times$  temporal compression), and audio waveforms are encoded through an Audio VAE that operates on mel spectrograms.

### 3.2 Identity-Driven In-Context LoRA

In-Context LoRA (IC-LoRA) [17] enables zero-shot adaptation by concatenating reference and target latents along the sequence dimension, allowing the model to learn identity correspondence through self-attention. We generalize this paradigm to the joint audio-video setting.

*Reference Conditioning.* Given a reference audio clip of the target speaker, we encode it into audio latents:

$$\mathbf{z}_a^{\text{ref}} = \mathcal{E}_a(A^{\text{ref}}) \quad (1)$$

where  $\mathcal{E}_a$  denotes the audio VAE encoder. The reference audio latents are concatenated with the target audio latents along the sequence dimension, while the video stream uses standard text-to-video generation with first-frame conditioning, which provides a strong visual anchor for face identity while allowing temporally coherent motion and synchronized audio:

$$\mathbf{z}^{\text{input}} = [\mathbf{z}_v^{\text{target}}, \mathbf{z}_a^{\text{ref}}, \mathbf{z}_a^{\text{target}}] \quad (2)$$

This audio-only in-context strategy allows the model to learn speaker identity transfer from the reference audio, while the video stream remains free to generate visuals guided by the text prompt and first frame while maintaining coherence with the audio.

*Negative Temporal Positions.* A key challenge in IC-LoRA is distinguishing reference tokens from target tokens during attention. In same-video settings such as dubbing or motion transfer, prior work assigns context tokens positional encodings identical to their target counterparts, enforcing strict spatiotemporal alignment between input and output [1, 27]. However, in our cross-video personalization setting, the reference audio originates from a different clip entirely and shares no temporal correspondence with the target. We therefore address this by assigning *negative* temporal positions to the reference audio tokens while keeping target positions positive. This creates a clear separation in the positional encoding space:

$$t^{\text{ref}} \in [-T_{\text{ref}}, 0), \quad t^{\text{target}} \in [0, T_{\text{target}}] \quad (3)$$

where  $T$  is the sequence length of the reference and target audio, respectively. This approach preserves the relative temporal structure within the reference while clearly demarcating reference from target.

### 3.3 Training Objective

We train ID-LoRA using the standard diffusion denoising objective. Given a reference audio clip and a target video-audio clip from the same speaker, the model learns to denoise the target latents while attending to the reference:

$$\mathcal{L} = \mathbb{E}_{t, \epsilon} \left[ \|\epsilon - \epsilon_\theta(\mathbf{z}_t^{\text{input}}, t, \mathbf{c})\|_2^2 \right] \quad (4)$$

where  $\epsilon_\theta$  is the denoising network with LoRA parameters, and  $\mathbf{c}$  includes text conditioning describing the target scene.

### 3.4 Identity-Guided Inference

At inference time, we introduce **Identity Guidance**, a classifier-free guidance variant that steers generation toward better identity preservation. We compute two forward passes: one with reference conditioning and one without, then extrapolate:

$$\hat{\epsilon} = \epsilon_{\theta}^{\text{uncond}} + s_{\text{id}} \cdot (\epsilon_{\theta}^{\text{ref}} - \epsilon_{\theta}^{\text{uncond}}) \quad (5)$$

where  $s_{\text{id}}$  is the identity guidance scale. Identity guidance follows the same principle as classifier-free guidance, but applied to the reference audio rather than the text prompt. Standard CFG extrapolates away from the unconditional prediction to amplify text influence; identity guidance similarly extrapolates between the unconditional and reference-conditioned predictions, amplifying identity-specific features, vocal timbre, speaking rhythm, pronunciation, while leaving scene content and environment sounds to be governed by the text prompt.

## 4 Experimental Settings

### 4.1 Datasets

We train ID-LoRA on CelebV-HQ [44] and TalkVid [3], maintaining separate checkpoints for each dataset. Both datasets undergo a shared preprocessing pipeline consisting of segmentation, captioning, speaker clustering, and source separation of reference audio to avoid environment leakage. For automatic evaluation, we construct a test set of 120 video pairs from 63 held-out speakers across both datasets. Pairs follow objective, method-agnostic criteria (same- vs. cross-video pairing); the only manual step verifies caption correctness and speaker-identity matching, with no subjective selection, and the same fixed pairs are evaluated for every method. The CelebV-HQ portion is divided into an *easy* (same-video) split of 36 videos spanning 14 speakers and a *hard* (cross-video) split of 35 videos spanning 8 speakers. The TalkVid portion comprises 49 videos from 41 held-out speakers. To assess cross-dataset generalization, we also evaluate the CelebV-HQ checkpoint on the TalkVid test set (see Sec. 5.2). Full preprocessing and split details are provided in the supplementary material.

### 4.2 Baselines

We compare ID-LoRA against three cascaded pipelines that pair a zero-shot voice-cloning model (CosyVoice 3.0 [8], VoiceCraft [26], or ElevenLabs [9]) with WAN2.2 [33] as the video backbone. ElevenLabs supports built-in emotion and style control via automatic prompt enhancement, making it the strongest cascaded baseline for speaking style adherence. We also compare against Kling 2.6 Pro [20], a closed-source commercial model that generates talking-face video with voice cloning in a unified pipeline. For both ElevenLabs and Kling, prompts were adapted following each platform’s official guidelines. Per-baseline details are provided in the supplementary material.

### 4.3 Evaluation Metrics

We evaluate five axes against the GT target videos: **Speaker Similarity** via WavLM+ECAPA-TDNN [2, 6, 35] cosine similarity; **Face Similarity** via ArcFace [5] cosine similarity (5-frame average); **Lip Synchronization** via SyncNet [4, 28], LSE-D ( $\downarrow$ ) and LSE-C ( $\uparrow$ ); **Audio Prompt Adherence** via CLAP [10] similarity between generated audio and the combined environment and speaking style prompt; and **WER** via Whisper-large-v3 [30] transcription. Full details are provided in the supplementary material.

### 4.4 Implementation Details

We implement ID-LoRA on top of the LTX-2 model using LoRA with rank 128. Training is performed on a single NVIDIA H100 GPU for 6,000 steps, using AdamW optimizer with learning rate  $2 \times 10^{-4}$  and batch size of 4. First-frame conditioning is applied with probability 0.9 during training.

At inference, we use 30 denoising steps with the following guidance scales: video CFG = 3.0, audio CFG = 7.0, identity guidance = 4.0, and AV-bimodal CFG = 3.0. STG (Spatio-Temporal Guidance) is applied with scale 1.0 on block 29. Generated videos are  $1024 \times 1024$  resolution with 121 frames at 25fps for both CelebV-HQ and TalkVid. AV-bimodal and STG guidance have been introduced in [15] and [18] respectively. Following the standard LTX-2 inference recipe [15], we render videos with its *two-stage* pipeline to all LTX-based methods.

## 5 Results

### 5.1 Main Results on CelebV-HQ

Table 1 presents our main comparison between ID-LoRA, three cascaded baselines (CosyVoice 3.0 + WAN2.2, VoiceCraft + WAN2.2, and ElevenLabs + WAN2.2), and the state-of-the-art closed-source Kling 2.6 Pro model across all evaluation splits: CelebV-HQ easy (same-video), CelebV-HQ hard (cross-video), and TalkVid. All cascaded baselines use WAN2.2 [33] as the video backbone. We additionally isolate architecture effects by comparing against LTX-based variants in Table 2.

On the **easy** (same-video) subset, all methods benefit from high source-target similarity, representing a voice replication scenario. On the **hard** (cross-video) subset, which tests generalization to novel acoustic settings, our advantages become more pronounced: the speaker similarity gap over the best cascaded baseline widens from +0.063 (easy) to +0.086 (hard). This suggests that our unified approach generalizes more robustly to new settings, whereas cascaded pipelines degrade more sharply when reference and target conditions diverge.

ID-LoRA outperforms all cascaded baselines on **speaker similarity** and **lip synchronization**, demonstrating that a unified generation scheme better preserves the target voice and produces more coherent audio-visual alignment than

**Table 1:** Comparison on CelebV-HQ easy (same-video), hard (cross-video), and TalkVid splits.  $\uparrow$  higher is better,  $\downarrow$  lower is better. Best in **bold**, second best underlined.  
 $\dagger$  Commercial model.

|         | Method                                                  | Spk Sim $\uparrow$ | Face Sim $\uparrow$ | LSE-D $\downarrow$ | LSE-C $\uparrow$ | CLAP $\uparrow$ | WER $\downarrow$ |
|---------|---------------------------------------------------------|--------------------|---------------------|--------------------|------------------|-----------------|------------------|
| C-Easy  | CosyVoice 3.0 + WAN2.2                                  | <u>0.510</u>       | <b>0.886</b>        | 11.27              | 1.53             | 0.301           | 0.273            |
|         | VoiceCraft + WAN2.2                                     | 0.412              | <u>0.879</u>        | 10.99              | 1.48             | <u>0.335</u>    | 0.241            |
|         | ElevenLabs + WAN2.2                                     | 0.413              | <u>0.879</u>        | 11.98              | 1.51             | 0.266           | 0.180            |
|         | Kling 2.6 Pro $^\dagger$                                | 0.487              | 0.847               | <u>10.24</u>       | <u>3.01</u>      | 0.327           | <b>0.046</b>     |
|         | <b>ID-LoRA (Ours)</b>                                   | <b>0.573</b>       | 0.870               | <b>8.20</b>        | <b>4.43</b>      | <b>0.372</b>    | <u>0.106</u>     |
| C-Hard  | CosyVoice 3.0 + WAN2.2                                  | <u>0.391</u>       | 0.890               | 11.40              | 1.50             | 0.249           | 0.362            |
|         | VoiceCraft + WAN2.2                                     | 0.344              | <u>0.892</u>        | 10.60              | 1.33             | 0.258           | 0.427            |
|         | ElevenLabs + WAN2.2                                     | 0.357              | <b>0.894</b>        | 11.86              | 1.72             | 0.238           | 0.154            |
|         | Kling 2.6 Pro $^\dagger$                                | 0.385              | 0.854               | <u>9.49</u>        | <u>3.47</u>      | <u>0.316</u>    | <u>0.121</u>     |
|         | <b>ID-LoRA (Ours)</b>                                   | <b>0.477</b>       | 0.874               | <b>8.49</b>        | <b>3.90</b>      | <b>0.363</b>    | <b>0.113</b>     |
| TalkVid | CosyVoice 3.0 + WAN2.2                                  | 0.579              | 0.770               | 12.34              | 1.20             | 0.315           | 0.223            |
|         | VoiceCraft + WAN2.2                                     | 0.457              | <b>0.773</b>        | 12.15              | 1.22             | <u>0.407</u>    | 0.171            |
|         | ElevenLabs + WAN2.2                                     | 0.491              | 0.772               | 12.42              | 1.31             | 0.319           | 0.064            |
|         | Kling 2.6 Pro $^\dagger$                                | 0.506              | 0.754               | 11.59              | 2.40             | 0.326           | <b>0.040</b>     |
|         | <b>ID-LoRA (Ours)</b>                                   | <b>0.599</b>       | <u>0.772</u>        | <u>10.62</u>       | <u>3.09</u>      | 0.385           | <u>0.054</u>     |
|         | <b>ID-LoRA (Celeb <math>\rightarrow</math> TalkVid)</b> | <u>0.595</u>       | 0.767               | <b>10.32</b>       | <b>3.12</b>      | <b>0.412</b>    | 0.065            |

cascading separately trained models. Kling 2.6 Pro, a unified commercial solution trained on significantly larger data, does not surpass our open approach on these axes. Both unified approaches (ours and Kling) also achieve substantially higher audio prompt adherence (CLAP) than the cascaded baselines, with the gap widening on the hard split (+0.105 vs. best cascaded) where environments diverge and cascaded pipelines blindly propagate reference acoustics. Statistical significance tests confirm our speaker similarity gains over all baselines at  $p < 0.001$ , and our CLAP advantage over Kling reaches  $p < 0.01$ .

## 5.2 Cross-Dataset Generalization

We additionally report TalkVid results in Table 1. ID-LoRA maintains its advantages in speaker similarity and lip synchronization on this dataset. Notably, the CelebV-HQ checkpoint applied to TalkVid without fine-tuning achieves speaker similarity of 0.595, only 0.004 below the in-domain checkpoint, still surpassing all other baselines, while actually improving CLAP (0.412 vs. 0.385) and LSE-D (10.32 vs. 10.62). This likely reflects CelebV-HQ’s greater acoustic diversity, indicating that ID-LoRA learns transferable identity representations rather than dataset-specific shortcuts. We also evaluate  $\sim 9.7$ s clips on a TalkVid subset; we report the results in the supplementary material.

**Table 2:** Ours vs LTX variants across three datasets: CelebV-HQ easy (same-video), CelebV-HQ hard (cross-video), and TalkVid. Includes TTS-conditioned LTX and LTX-Zeroshot variants.  $\uparrow$  higher is better,  $\downarrow$  lower is better. Best in **bold**, second best underlined.

|         | Method                                          | Spk Sim $\uparrow$ | Face Sim $\uparrow$ | LSE-D $\downarrow$ | LSE-C $\uparrow$ | CLAP $\uparrow$ | WER $\downarrow$ |
|---------|-------------------------------------------------|--------------------|---------------------|--------------------|------------------|-----------------|------------------|
| C-Easy  | CosyVoice 3.0 + LTX                             | 0.522              | 0.875               | 9.42               | 2.81             | 0.307           | 0.287            |
|         | VoiceCraft + LTX                                | 0.412              | <b>0.885</b>        | 9.19               | 2.32             | <u>0.335</u>    | 0.241            |
|         | ElevenLabs + LTX                                | 0.420              | 0.873               | 9.44               | <u>2.99</u>      | 0.262           | <u>0.185</u>     |
|         | LTX-Zeroshot                                    | 0.462              | <u>0.881</u>        | <b>8.05</b>        | 2.41             | 0.308           | 4.634            |
|         | <b>ID-LoRA</b>                                  | <b>0.573</b>       | 0.870               | <u>8.20</u>        | <b>4.43</b>      | <b>0.372</b>    | <b>0.106</b>     |
| C-Hard  | CosyVoice 3.0 + LTX                             | 0.409              | 0.888               | 9.39               | 2.13             | 0.246           | 0.362            |
|         | VoiceCraft + LTX                                | 0.344              | 0.883               | 8.72               | 1.52             | <u>0.258</u>    | 0.427            |
|         | ElevenLabs + LTX                                | 0.364              | 0.886               | 10.31              | <u>2.21</u>      | 0.235           | <u>0.152</u>     |
|         | LTX-Zeroshot                                    | 0.374              | <b>0.896</b>        | <b>7.89</b>        | 2.07             | 0.242           | 2.393            |
|         | <b>ID-LoRA</b>                                  | <b>0.477</b>       | 0.874               | <u>8.49</u>        | <b>3.90</b>      | <b>0.363</b>    | <b>0.113</b>     |
| TalkVid | CosyVoice 3.0 + LTX                             | 0.578              | 0.768               | 11.33              | 2.33             | 0.324           | 0.223            |
|         | VoiceCraft + LTX                                | 0.457              | 0.768               | 11.37              | 1.81             | <u>0.405</u>    | 0.165            |
|         | ElevenLabs + LTX                                | 0.493              | 0.768               | 11.13              | 2.71             | 0.322           | <u>0.064</u>     |
|         | LTX-Zeroshot                                    | 0.530              | <b>0.850</b>        | <b>9.81</b>        | 2.45             | 0.381           | 0.471            |
|         | <b>ID-LoRA</b>                                  | <b>0.599</b>       | <u>0.772</u>        | 10.62              | <u>3.09</u>      | 0.385           | <b>0.054</b>     |
|         | <b>ID-LoRA</b><br>(Celeb $\rightarrow$ TalkVid) | <u>0.595</u>       | 0.767               | <u>10.32</u>       | <b>3.12</b>      | <b>0.412</b>    | 0.065            |

### 5.3 Controlling for Backbone Effects

To isolate pipeline design from model family effects, Table 2 compares ID-LoRA with LTX-based cascaded variants (CosyVoice-LTX, VoiceCraft-LTX, ElevenLabs-LTX) and LTX-Zeroshot.

Across CelebV-HQ easy/hard, ID-LoRA remains the strongest on speaker similarity and synchronization-oriented metrics (LSE-C and competitive LSE-D), while several LTX variants achieve higher face similarity, an effect we analyze below. On TalkVid, ID-LoRA still leads in speaker similarity and low WER, whereas LTX-Zeroshot can obtain very low LSE-D together with degraded transcription quality (high WER), indicating that low SyncNet distance alone can overestimate quality when articulation is poor. This motivates evaluating synchronization jointly with intelligibility and identity rather than in isolation.

*Face similarity and lip synchronization.* All baselines achieve comparable or higher face similarity than ID-LoRA, while Kling 2.6 Pro scores lower (0.847–0.854). This reflects a systematic bias in ArcFace: methods with less lip motion keep face embeddings closer to the reference, inflating the metric. A similar confound affects lip synchronization: LTX-Zeroshot achieves the lowest LSE-D across all splits yet also the lowest LSE-C confidence and largely unintelligible speech (WER > 2.3 on CelebV-HQ), indicating that its low lip-sync distance stems from

**Table 3:** Full  $2 \times 2$  ablation of our two core components: negative temporal positions (Neg. PEs) and identity guidance (IdG), on the CelebV-HQ hard split (cross-video). The two components combine *super-additively* on speaker similarity, while Neg. PEs is the dominant factor for intelligibility (WER). Best in **bold**, second best underlined.

| Method                       | Spk Sim $\uparrow$ | Face Sim $\uparrow$ | LSE-D $\downarrow$ | CLAP $\uparrow$ | WER $\downarrow$ |
|------------------------------|--------------------|---------------------|--------------------|-----------------|------------------|
| Neither                      | 0.430              | 0.873               | 8.84               | <b>0.373</b>    | 0.188            |
| Neg. PEs only                | 0.438              | <b>0.875</b>        | <u>8.71</u>        | <u>0.364</u>    | <b>0.113</b>     |
| IdG only                     | <u>0.441</u>       | 0.872               | 9.02               | 0.359           | 0.252            |
| <b>OURS (Neg. PEs + IdG)</b> | <b>0.477</b>       | <u>0.874</u>        | <b>8.50</b>        | 0.363           | <b>0.113</b>     |

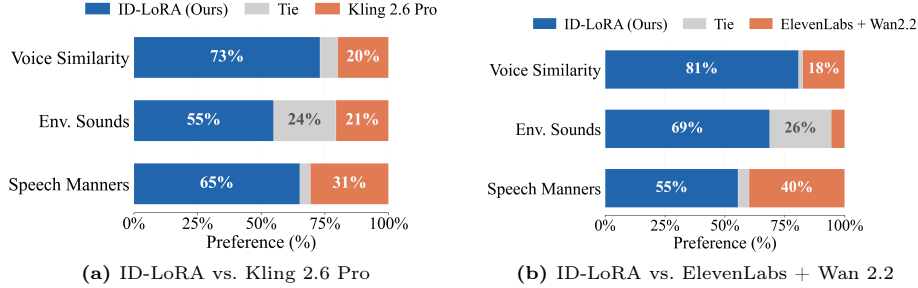
near-static frames rather than accurate articulation. These observations motivate evaluating synchronization jointly with intelligibility and identity rather than in isolation, and suggest that face similarity alone is an incomplete measure for talking-head generation, as it penalizes realistic speech articulation. A quantitative lip motion analysis is provided in the supplementary material.

#### 5.4 Ablation Studies

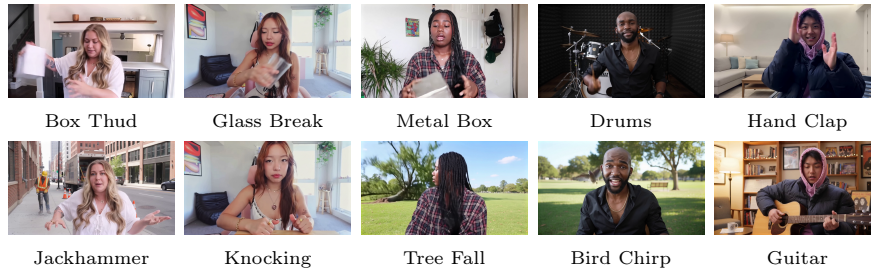
*Negative Temporal Positions and Identity Guidance.* Table 3 reports the full  $2 \times 2$  ablation of our two core components on the hard split. Each component is of limited value in isolation: relative to the bare baseline (neither component, 0.430 speaker similarity), adding negative temporal positions or identity guidance alone improves speaker similarity by only +0.008 and +0.011, respectively, whereas enabling both together yields a markedly larger +0.047 gain. This super-additive interaction indicates that the two components reinforce one another rather than contributing independently. Intelligibility, by contrast, is governed primarily by the positional encoding: every configuration with negative temporal positions attains  $\text{WER} \leq 0.113$ , while standard encodings raise WER to 0.188–0.252, confirming that the model struggles to separate reference from target speech without a clear positional boundary. Face similarity and audio prompt adherence (CLAP) are largely insensitive to either component. A sensitivity analysis of the identity guidance scale is provided in the supplementary material.

*Backbone Generalization.* Our three core ingredients are backbone-agnostic: negative temporal positions reuse the pretrained RoPE and apply to any RoPE-based transformer, identity guidance is inference-time, and audio-visual in-context conditioning needs only shared cross-modal attention. Applying the identical recipe without re-tuning to LTX-2.3 [21], a larger LTX variant (22B) with a different audio model, transfers directly: on the CelebV-HQ hard split it yields intelligible, synchronized speech (WER 0.165, LSE-D 8.97) and even improves speaker similarity (0.481 vs. 0.477). Our recipe is thus not tied to one architecture and benefits from stronger backbones.





**Fig. 3:** Human evaluation results: A/B preference rates (%) on the hard (cross-video) split. Annotators on AMT evaluated 35 pairs across 8 speakers along three axes: voice similarity, environment sounds, and speech manners.

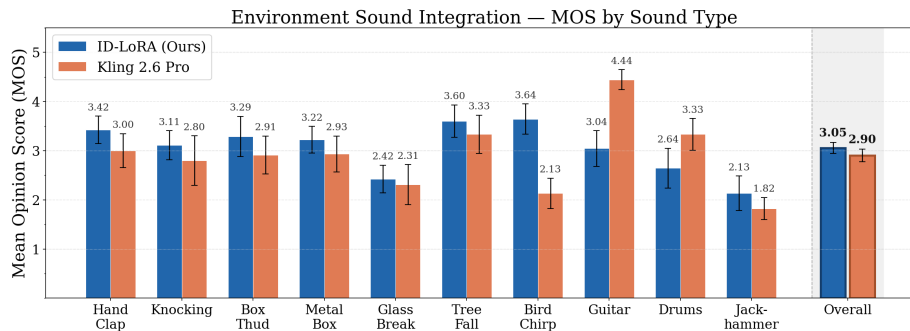


**Fig. 4:** Examples from the MOS evaluation set. Each column shows two generated videos from the same speaker under different environmental sound conditions.

## 5.5 Human Evaluation

To complement automatic metrics, we conduct two human evaluations on Amazon Mechanical Turk (AMT), with 9 annotators per item. Participation is restricted to workers with Masters qualification, a HIT approval rate above 97%, and location in English-speaking countries.

*A/B Preference Study.* We perform two A/B preference tests comparing ID-LoRA against Kling 2.6 Pro and ElevenLabs + WAN2.2, using the CelebV-HQ hard split. For each pair, annotators are shown videos from both methods in randomized A/B order and select a preference (A, B, or Equal) along three axes: voice similarity (with a reference video shown), environment sounds (with the target description shown), and speech manners (with the target style description shown). To ensure annotation quality, we employ A/B randomization independent per (pair, question), a cyclic Latin square so that no annotator evaluates the same pair under two different question types, watch-gating that disables rating until all videos are watched to completion, guided practice rounds, and two hidden control pairs with known answers to filter inattentive workers (see supplementary for full protocol).



**Fig. 5:** Environment sound interaction MOS study: per-scenario mean opinion scores (1–5) with 95% confidence intervals for ID-LoRA and Kling 2.6 Pro across ten physical interaction scenarios. ID-LoRA scores higher on 8 of 10 scenarios and in overall score.

Figure 3 shows the preference rates, with  $\sim 285$ – $290$  annotations per question type. All p-values are from two-sided binomial tests on win counts excluding equal responses. Against ElevenLabs + Wan 2.2, ID-LoRA is overwhelmingly preferred for voice similarity (80.7% vs. 17.5%; win rate 82.1%,  $p < 0.001$ ) and environment sounds (68.7% vs. 5.6%; win rate 92.4%,  $p < 0.001$ ), with a moderate advantage for speech manners (55.5% vs. 39.9%; win rate 58.1%,  $p < 0.01$ ). Against Kling 2.6 Pro, ID-LoRA is significantly preferred on all three axes ( $p < 0.001$ ): voice similarity (73.1% vs. 20.0%; win rate 78.5%), environment sounds (54.8% vs. 20.7%; win rate 72.6%), and speech manners (65.2% vs. 30.7%; win rate 68.0%). Across both comparisons, the strongest advantages appear in voice similarity and environment sounds, validating that unified generation jointly preserves speaker identity and enables text-driven control over the audio scene.

*Environment Sound Interaction MOS Study.* The A/B study evaluates ambient sounds described in the prompt (*e.g.*, wind, traffic). A more demanding test is whether the model can generate sounds corresponding to *physical interactions* depicted in the scene—*e.g.*, a thud when a box drops or guitar music when strumming. Cascaded baselines cannot perform this task as they generate speech only; we therefore compare ID-LoRA against Kling 2.6 Pro only. We design ten interaction scenarios, *e.g.* dropping a box, clapping, drumming. For each scenario, we select five speakers from TalkVid and use Nano-banana [13] to insert a scene-appropriate object and background into the first frame (*e.g.*, a guitar in the speaker’s hands). Both methods generate videos from these edited frames with prompts describing the interaction, yielding new 50 paired samples. Annotators rate each video independently on a 1–5 MOS scale (1 = Bad, 5 = Excellent) assessing how well the generated audio matches the depicted interaction. Representative examples are shown in Fig. 4; Scenario details and per-scenario analysis are in the supplementary.

Results are shown in Figure 5. ID-LoRA achieves a higher overall MOS (3.05 vs. 2.90), wins on 8 of 10 scenarios, and exhibits lower variance with fewer

“Bad” ratings (15.3% vs. 23.3%). The fact that ID-LoRA, trained on only  $\sim 3K$  pairs, matches a large-scale commercial system confirms that unified generation provides a strong inductive bias for physically grounded audio-visual correspondence.

## 6 Conclusions

We presented **ID-LoRA**, a unified audio-visual personalization method that jointly synthesizes a subject’s appearance and vocal identity. Unlike cascaded pipelines, our approach benefits from cross-modal attention and allows text conditioning to influence both modalities simultaneously by leveraging the LTX-2 backbone for zero-shot personalization. Key architectural contributions include negative temporal positions for reference-target separation and identity guidance for enhanced speaker preservation.

Experiments demonstrate that ID-LoRA outperforms cascaded baselines and competes with state-of-the-art closed-source models in speaker similarity and lip synchronization. Crucially, human evaluations show our method significantly surpasses both cascaded baselines and Kling 2.6 Pro in matching speech manners, generating appropriate environment sounds, and integrating physically grounded audio. Overall, ID-LoRA yields more coherent audio-visual outputs with stronger prompt adherence and identity preservation than dedicated TTS pipelines. We focus on single-speaker personalization to study identity transfer in isolation; future work will extend to multi-speaker, cross-lingual, and general Audio-Visual Reference-to-Video generation.

## Acknowledgements

We want to thank Naomi Ken Korem and the LTX team for their support surrounding the LTX model family. The work was supported by a grant from Google and the Tel Aviv University Center for AI and Data Science (TAD).

## References

1. Chen, A., Korem, N.K., Zeevi, G., Halperin, T., Yosef, M.B., Jelercic, U., Bibi, O., Patashnik, O., Cohen-Or, D.: Just-dub-it: Video dubbing via joint audio-visual diffusion (2026), <https://arxiv.org/abs/2601.22143>
2. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., 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)
3. Chen, S., Huang, H., Liu, Y., Ye, Z., Chen, P., Zhu, C., Guan, M., Wang, R., Chen, J., Li, G., Lim, S.N., Yang, H., Wang, B.: Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis (2025)
4. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: *Asian conference on computer vision*. pp. 251–263. Springer (2016)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4690–4699 (2019)
6. Desplanques, B., Thienpondt, J., Demuynck, K.: Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143* (2020)
7. Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., et al.: Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407* (2024)
8. Du, Z., Gao, C., Wang, Y., Yu, F., Zhao, T., Wang, H., Lv, X., Wang, H., Shi, X., An, K., et al.: Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589* (2025)
9. ElevenLabs: Elevenlabs. <https://elevenlabs.io/> (2026), accessed: 2026-03-02
10. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap learning audio concepts from natural language supervision. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2023)
11. Flynn, J., Paier, W., Dinev, D., Nguyen, S.N., Poghosyan, H., Toribio, M., Banerjee, S., Gafni, G.: Edityourself: Audio-driven generation and manipulation of talking head videos with diffusion transformers. *arXiv preprint arXiv:2601.22127* (2026)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022). <https://doi.org/10.48550/ARXIV.2208.01618>, <https://arxiv.org/abs/2208.01618>
13. Google DeepMind: Nano Banana Pro (gemini 3 pro image). <https://blog.google/innovation-and-ai/products/nano-banana-pro/> (2026)
14. Guo, X., Ye, F., Sun, Q., Chen, L., Li, B., Zhang, P., Liu, J., Zhao, S., He, Q., Hou, X.: Dreamid-omni: Unified framework for controllable human-centric audio-video generation (2026), <https://arxiv.org/abs/2602.12160>
15. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., Richardson, E., Shiran, G., Chachy, I., Chetboun, J., Finkelson, M., Kupchick, M., Zabari, N., Guetta, N., Kotler, N., Bibi, O., Gordon, O., Panet, P., Benita, R., Armon, S., Kulikov, V., Inger, Y., Shiftan, Y., Melumian, Z., Farbman, Z.: Ltx-2: Efficient joint audio-visual foundation model (2026), <https://arxiv.org/abs/2601.03233>

16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), <https://openreview.net/forum?id=nZevKeeFYf9>
17. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)
18. Hyung, J., Kim, K., Hong, S., Kim, M.J., Choo, J.: Spatiotemporal skip guidance for enhanced video diffusion sampling. arXiv preprint arXiv:2411.18664 (2024)
19. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing (2025), <https://arxiv.org/abs/2503.07598>
20. Kuaishou Technology: Kling 2.6 pro: Audio-visual generation model. <https://app.klingai.com/global/> (2026), accessed: 2026
21. Lightricks: LTX-2.3. <https://ltx.io/model/ltx-2-3> (2026), accessed: 2026-06-21
22. Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., Plumbley, M.D.: Audioldm 2: Learning holistic audio generation with self-supervised pretraining. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 2871–2883 (2024). <https://doi.org/10.1109/TASLP.2024.3399607>
23. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Jiang, R., Luo, J., Fei, H., et al.: Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025)
24. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14951–14961 (2025)
25. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation (2025), <https://arxiv.org/abs/2510.01284>
26. Peng, P., Huang, P.Y., Li, S.W., Mohamed, A., Harwath, D.: VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12442–12462. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). <https://doi.org/10.18653/v1/2024.acl-long.673>, <https://aclanthology.org/2024.acl-long.673/>
27. Polaczek, S., Patashnik, O., Mahdavi-Amiri, A., Cohen-Or, D.: In-context sync-lora for portrait video editing (2025), <https://arxiv.org/abs/2512.03013>
28. Prajwal, K., Mukhopadhyay, R., Nambodiri, V.P., Jawahar, C.: 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)
29. Qiang, C., Wang, J., Wang, X., Yin, K., Guo, Y., Zeng, X., Li, N., Li, Z., Liang, Y., Zhang, Z., et al.: Mm-sonate: Multimodal controllable audio-video generation with zero-shot voice cloning. arXiv preprint arXiv:2601.01568 (2026)
30. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022). <https://doi.org/10.48550/ARXIV.2212.04356>, <https://arxiv.org/abs/2212.04356>
31. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)

32. Vyas, A., Shi, B., Le, M., Tjandra, A., Wu, Y.C., Guo, B., Zhang, J., Zhang, X., Adkins, R., Ngan, W., et al.: Audiobox: Unified audio generation with natural language prompts. arXiv preprint arXiv:2312.15821 (2023)
33. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), <https://arxiv.org/abs/2503.20314>
34. Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F.: Neural codec language models are zero-shot text to speech synthesizers (2023), <https://arxiv.org/abs/2301.02111>
35. Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., Zeng, M., Huang, X.: Unispeech: Unified speech representation learning with labeled and unlabeled data. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 10937–10947. PMLR (2021), <http://proceedings.mlr.press/v139/wang21y.html>
36. Wang, J., Qiang, C., Guo, Y., Wang, Y., Zeng, X., Deng, F.: Apollo: Unified multi-task audio-video joint generation (2026), <https://arxiv.org/abs/2601.04151>
37. Wang, K., Deng, S., Shi, J., Hatzinakos, D., Tian, Y.: Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation (2024), <https://arxiv.org/abs/2406.07686>
38. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
39. Wang, Y., Zhan, H., Liu, L., Zeng, R., Guo, H., Zheng, J., Zhang, Q., Zhang, X., Zhang, S., Wu, Z.: Maskgct: Zero-shot text-to-speech with masked generative codec transformer. arXiv preprint arXiv:2409.00750 (2024)
40. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024)
41. Xu, S., Chen, G., Guo, Y.X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B.: Vasa-1: Lifelike audio-driven talking faces generated in real time. Advances in Neural Information Processing Systems **37**, 660–684 (2024)
42. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
43. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8652–8661 (2023)
44. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: Celebv-hq: A large-scale video facial attributes dataset. In: European conference on computer vision. pp. 650–667. Springer (2022)