BEHAVE YOURSELF! PERSONALIZING TALKING HEADS BEYOND APPEARANCE

Anonymous authors
Paper under double-blind review

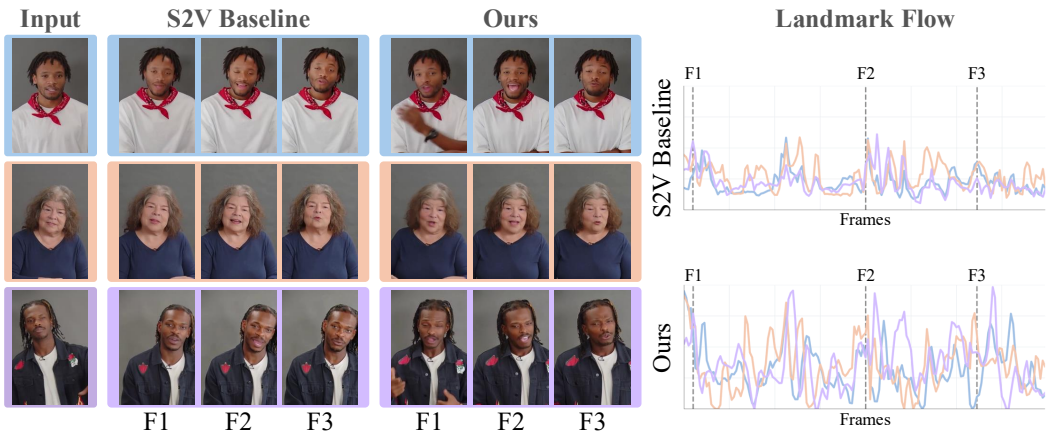


Figure 1: *Is appearance enough to personalize a talking video?* Given the same driving audio and an input frame (left), the S2V (Speech-To-Video) baseline produces similar facial dynamics across subjects, reflected in similar landmark-flow patterns (right). In contrast, our proposed spatio-temporal identity conditioning captures person-specific expressions, head movements, and gestures, yielding more distinct motion across subjects.

ABSTRACT

Recent video generation models enable appearance-based personalization by conditioning generation on reference images of a specific person. Yet a person's identity is not fully captured by appearance alone. It also includes characteristic behavior, such as distinctive head movements one makes while speaking or listening, which is not directly exposed through static reference images. In this work, we introduce a *spatio-temporal identity encoder* that aggregates a variable number of reference videos into a fixed number of identity tokens, capturing both visual details revealed across frames as well as the subject's characteristic behavior. These identity tokens condition a frozen audio-driven video diffusion transformer through dedicated cross-attention layers, enabling generation with new speech while preserving the subject's appearance and characteristic behavior. A three-stage training procedure first adapts the video encoder to talking-head videos through self-supervised training. It then establishes identity conditioning using the target clip as reference, before switching to different reference clips of the same subject. Through qualitative and quantitative experiments, we demonstrate that our method yields more authentic, person-specific videos compared to an appearance-only baseline.

1 INTRODUCTION

Recent video generation models (Wan et al., 2025; Kong et al., 2024; Blattmann et al., 2023; Polyak et al., 2024) have enabled the synthesis of remarkably realistic and temporally coherent videos. A particularly compelling capability is image-based personalization, where reference images provide visual cues for generating videos of a specific person (Hu, 2024; Tian et al., 2024; Xu et al., 2024a). These advances support content creation applications such as generating personalized talking-head videos and redubbing films.

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.

Yet a person's identity extends far beyond their visual appearance. It is also conveyed through characteristic behavior, from the way a smile develops to a habitual nod or head tilt during conversation. These temporal, person-specific traits form a fundamental part of how we recognize and relate to one another, and they vary substantially across individuals. Critically, such recurring patterns cannot be observed in a single static image. As a result, current appearance-based personalization methods can produce videos that visually resemble the subject but exhibit generic facial motion, as illustrated by the baseline's generated frames and landmark-flow traces in Figure 1. The generated person may look like the subject, but not behave like them.

One line of work animates a target identity by transferring motion from a driving video (Siarohin et al., 2019; Guo et al., 2024; Cheng et al., 2025). These methods typically operate through intermediate representations such as facial landmarks, 3D morphable model parameters, or keypoints, transferring motion frame by frame from the driving performance. However, when generating entirely new content, the goal is to synthesize motion that reflects the subject's characteristic behavior while remaining consistent with the model's conditioning, whether a text prompt or driving speech. Existing personalization techniques (Ruiz et al., 2023; Gal et al., 2022; Ye et al., 2023; Wang et al., 2024b) use reference images to reproduce a person's visual identity, but primarily focus on appearance rather than the subject's characteristic behavior. Motion customization methods learn reusable motion patterns through per-concept optimization (Zhao et al., 2024).

In this work, we personalize audio-driven talking-head generation by modeling the subject's characteristic behavior alongside their appearance. Different reference clips reveal complementary aspects of a subject's appearance and behavior. We use V-JEPA (Mur-Labadia et al., 2026) to jointly encode each reference clip's frames, capturing temporal context. Our Spatio-Temporal Identity Encoder aggregates a variable number of reference videos into a fixed number of identity tokens, forming a unified identity representation that guides generation. The tokens condition a frozen audio-driven video diffusion transformer through dedicated cross-attention layers, alongside an appearance reference image and driving audio, enabling new talking-head videos that preserve these subject-specific characteristics while remaining synchronized with new speech. Our three-stage training procedure begins by adapting our video encoder, V-JEPA 2.1 (Mur-Labadia et al., 2026), to talking-head videos through self-supervised learning. We then establish identity conditioning using the target clip itself as reference. Finally, we train with different reference clips of the same subject, encouraging the model to capture recurring aspects of their appearance and behavior. Our reference-video conditioning produces more individual expressions, head movements, and gestures, accompanied by more distinct landmark-flow patterns across subjects, as illustrated in Figure 1.

Through qualitative and quantitative experiments, we demonstrate that our method yields more authentic, person-specific videos compared to an appearance-only baseline. Human evaluations further show that our method improves how faithfully a generator already conditioned on a reference video reproduces the subject's characteristic behavior. Our results highlight the importance of modeling characteristic behavior alongside appearance in personalized talking-head generation.

2 Related Work

Video and Audio-Driven Generation. Video diffusion and flow-based models, including Stable Video Diffusion (Blattmann et al., 2023), CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2024), Movie Gen (Polyak et al., 2024), and Wan (Wan et al., 2025), have improved video quality and temporal coherence. SadTalker (Zhang et al., 2023), AniPortrait (Wei et al., 2024a), EMO (Tian et al., 2024), Hallo (Xu et al., 2024a), VASA-1 (Xu et al., 2024b), and V-Express (Wang et al., 2024a) animate reference portraits from speech, focusing on realistic motion and lip synchronization. Wan-S2V (Gao et al., 2025) extends the Wan backbone to audio-driven character animation. We condition frozen Wan-S2V on identity features from reference videos.

Visual Identity Personalization. Textual Inversion (Gal et al., 2022) and DreamBooth (Ruiz et al., 2023) optimize subject-specific embeddings or model weights. IP-Adapter (Ye et al., 2023), PhotoMaker (Li et al., 2024), and InstantID (Wang et al., 2024b) encode reference images without subject-specific optimization. Omni-ID (Qian et al., 2025) uses learnable queries to aggregate a variable number of reference images into a fixed-size identity representation. Nested sAttention (Patashnik et al., 2025) explores semantic-aware attention values for concept personalization. VideoBooth (Jiang et al., 2024), ID-Animator (He et al., 2024), and ConsisID (Yuan et al., 2025)

extend reference-based appearance preservation to video. These methods emphasize appearance preservation and do not explicitly model recurring subject-specific facial and head motion.

Reference-Based Speaking-Style Personalization. Several methods infer speaking style from reference videos. StyleTalk (Ma et al., 2023) models reference expression trajectories through a 3D morphable model, while Say Anything with Any Style (Tan et al., 2024) uses learned style and pose codebooks. DiffPoseTalk (Sun et al., 2024) and Imitator (Thambiraja et al., 2023) generate personalized 3D facial animation. MimicTalk (Ye et al., 2024) combines in-context speaking-style conditioning with a NeRF renderer adapted to the target person. These approaches rely on structured motion representations or dedicated rendering pipelines, rather than leveraging the generative priors of video diffusion models pre-trained on large-scale video datasets. OmniTalker (Wang et al., 2025) addresses text-driven audio-video generation with one-shot multimodal style mimicking. TAVR (Guo et al., 2026) and our method leverage video diffusion models pre-trained on large-scale video data, benefiting from their visual and temporal priors. Directly extending TAVR's VAE-token conditioning to more reference videos increases the token sequence processed at each denoising step. We instead compress domain-adapted V-JEPA features from all references into a fixed number of identity tokens before generation. Additional references therefore increase only the one-time reference-processing cost, not the generator's per-step conditioning cost.

Self-Supervised Video Representation Learning. V-JEPA (Bardes et al., 2024) learns video representations by predicting masked target features in latent space rather than reconstructing pixels. V-JEPA 2 (Assran et al., 2025) scales this approach toward general video understanding, prediction, and planning, while V-JEPA 2.1 (Mur-Labadia et al., 2026) introduces dense feature prediction and deep self-supervision for spatially localized video representations. These models learn general-purpose visual features and do not explicitly disentangle appearance from subject-specific motion. We adapt V-JEPA 2.1 to talking-head videos and feed its dense patch features to our identity encoder.

3 METHOD

Given an appearance reference image, a driving audio signal, and one or more reference videos of the same subject, our goal is to generate a talking-avatar video. The generated video should preserve the appearance of the subject from the reference image, remain synchronized with the driving audio, and exhibit facial and head dynamics consistent with recurring motion patterns observed in the reference videos. The model is trained once across subjects and can personalize a previously unseen subject by encoding their reference videos, without updating any model parameters.

We build on Wan2.2-S2V (Gao et al., 2025; Wan et al., 2025), a pre-trained audio-driven video diffusion transformer that we refer to as Wan-S2V. While its image and audio conditioning provide appearance and speech-synchronized motion, the model has no explicit access to subject-specific motion observations. Although Wan-S2V also supports text conditioning, the dynamic identity of a person is a rich visual signal that cannot be expressed as a prompt. We therefore condition Wan-S2V directly on reference videos of the target subject. We extract spatio-temporal features from each reference video with a V-JEPA 2.1 video encoder (Mur-Labadia et al., 2026), and a trainable *Spatio-Temporal Identity Encoder* aggregates them into a fixed set of identity tokens Z. The identity tokens complement the appearance reference image by capturing recurring subject-specific dynamics and visual details revealed during motion, such as teeth and expression-dependent wrinkles. Newly added cross-attention layers inject these tokens into the frozen diffusion transformer (Figure 2).

We train our model in three stages. We first adapt a pre-trained video encoder to talking-head videos. We then train the identity encoder and the added cross-attention layers, using a target clip whose audio drives the generation, first with *same-clip conditioning*, where the target clip itself serves as the reference, and then with *cross-clip conditioning*, where the reference is a set of different clips of the same subject (Figure 3). This transition from same-clip to cross-clip references encourages the representation to retain subject-specific cues that recur across clips. At inference, the reference clips are encoded once, and the resulting identity tokens condition the standard iterative diffusion sampling process. No subject-specific fine-tuning or optimization is required. The rest of this section describes the identity encoder and its integration into Wan-S2V, the three-stage training procedure, and the guidance scheme during inference.

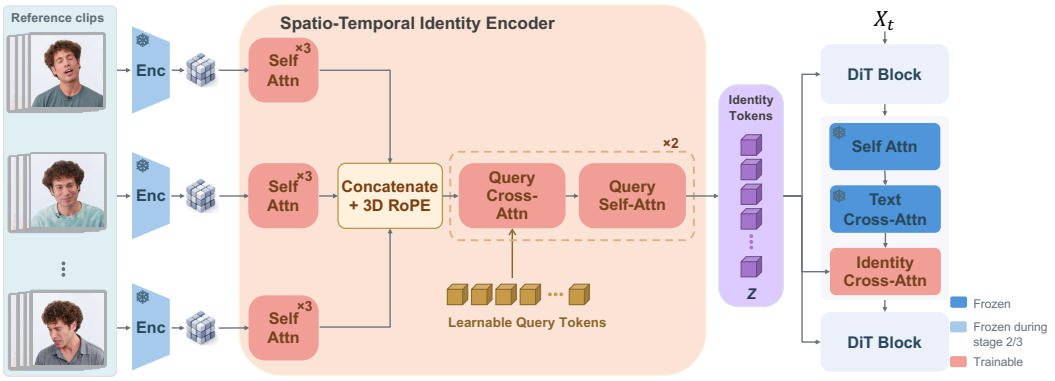


Figure 2: **Spatio-Temporal Identity Encoder and diffusion-transformer conditioning.** Each reference clip is processed by the frozen V-JEPA video encoder, labeled *Enc*, followed by three shared *Self Attn* blocks. The resulting clip features are concatenated and assigned 3D RoPE coordinates. Two aggregation blocks, each applying query cross-attention followed by query self-attention, produce a fixed set of identity tokens Z. The tokens condition every frozen DiT backbone block through a trainable identity cross-attention branch.

3.1 Model Architecture

Figure 2 provides an overview of the reference-video encoder and its integration into Wan-S2V.

Spatio-Temporal Identity Encoder. Each reference clip is processed by the video encoder, labeled *Enc* in Figure 2, to produce dense patch features. The encoder is adapted to talking-head videos and then kept frozen, as described in Section 3.2. The Spatio-Temporal Identity Encoder maps these features from a variable number of reference clips to a fixed set of identity tokens $Z \in \mathbb{R}^{Q\times d}$. The patch features of each clip first pass through shared self-attention blocks that preserve their spatial and temporal positions. We then concatenate the features of all clips and aggregate them with Q learnable query tokens through cross-attention, following Q-Former (Li et al., 2023). Each clip receives a distinct temporal offset in its positional encoding, so the queries attend jointly across clips while still distinguishing them. Architecture details are provided in Appendix A.

Conditioning the Diffusion Transformer. We condition the diffusion transformer on the identity tokens Z by adding an identity cross-attention branch to each Wan-S2V transformer block, in parallel to its frozen text cross-attention, following the decoupled attention design of IP-Adapter (Ye et al., 2023). The cross-attention update of block ℓ becomes

$$h_\ell^+ = h_\ell + \mathrm{CA}^\ell_{\mathrm{txt}}(\mathrm{LN}_\ell(h_\ell), c_{\mathrm{txt}}) + \mathrm{CA}^\ell_{\mathrm{id}}(\mathrm{LN}_\ell(h_\ell), Z)\,, \tag{1}$$

where h_ℓ are the hidden states at the input to the cross-attention sublayer, c_{txt} is the text context, $\mathrm{CA}^\ell_{\mathrm{txt}}$ is the original frozen branch, and $\mathrm{CA}^\ell_{\mathrm{id}}$ is the new branch with its own learnable projections. We initialize the output projection of each identity branch to zero, so the pre-trained generator is unchanged at the start of training.

3.2 Training

Domain Fine-Tuning of a Pretrained Video Encoder. Personalizing dynamics requires a representation that captures how a person moves in addition to how they look. We obtain this representation from V-JEPA 2.1 (Mur-Labadia et al., 2026), a self-supervised video encoder that is pre-trained for general-purpose video representations rather than specifically for talking-head personalization. We therefore adapt the encoder by continuing self-supervised training on face crops from our talking-head training set with the official V-JEPA 2.1 objective, which requires no subject labels. We analyze how this adaptation changes the representations of the encoder in Section 4.5. After fine-tuning, we freeze the encoder and do not propagate the flow-matching loss of the generator into it during the next stages. This self-supervised adaptation does not require forward or backward passes through Wan-S2V.

Training the Spatio-Temporal Identity Encoder. We now train the two modules that connect them, the Spatio-Temporal Identity Encoder and the identity cross-attention branches, while keeping both the video encoder and Wan-S2V frozen.

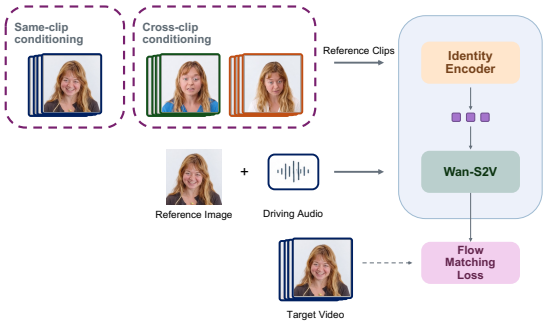


Figure 3: **Training the identity encoder via same-clip and cross-clip conditioning.** Same-clip conditioning (left) uses the target video itself as the identity reference, while cross-clip conditioning (right) uses different reference clips of the same subject. In both settings, the target video supplies the driving audio and flow-matching supervision.

Our goal is to extract subject-specific cues from reference clips that differ from the target and use them to guide the generated video. Training directly in this setting is difficult, because the reference and target motions are not temporally aligned and the loss indicates only weakly how the reference should shape the generated motion. The signal is weakest at initialization, where the zero-initialized output projections keep the identity branch inactive and the frozen generator already produces plausible motion from its image and audio conditions, so optimization can leave the identity tokens underused. We therefore split this training into the two identity-conditioning stages illustrated in Figure 3, and in both of them the target video supplies the driving audio and the flow-matching target. Under *same-clip conditioning*, the same five-second segment serves as both the identity reference and the training target, so their temporal alignment gives a direct signal for how the identity tokens should influence the generated motion. Under *cross-clip conditioning*, which starts from the same-clip weights, the identity encoder aggregates a variable number of different reference clips of the same subject, where the target is excluded from the reference set and no reference clip temporally overlaps with it. Both stages use the flow-matching objective of Wan-S2V (Lipman et al.; Esser et al., 2024), conditioned on the appearance image, text, driving audio, and identity tokens, and gradients update only the Spatio-Temporal Identity Encoder and the identity cross-attention branches. We also apply *conditioning dropout*, independently dropping the text, audio, and identity conditioning during training, while the appearance reference image remains available in every case. Text dropout replaces the fixed prompt with empty text and supports classifier-free text guidance, while audio and identity dropout zero their embeddings and expose the model to the zeroed configurations used at inference.

3.3 Dual Classifier-Free Guidance

At inference, we use a chained form of classifier-free guidance (CFG) (Ho & Salimans, 2022) so that the original Wan-S2V conditions and the added identity condition can be controlled separately. Following the chained formulation of InstructPix2Pix (Brooks et al., 2023), we evaluate three velocity predictions at each denoising step. v_{uncond} uses zeroed audio and identity tokens, v_{id0} uses driving audio with zeroed identity tokens, and v_{full} additionally uses the identity tokens Z. All three share the appearance image, and we combine them as

$$\hat{v} = v_{\text{uncond}} + \omega_{\text{S2V}} \left(v_{\text{id0}} - v_{\text{uncond}}\right) + \omega_{\text{id}}(p) \left(v_{\text{full}} - v_{\text{id0}}\right), \tag{2}$$

where ω_{S2V} controls the text and audio guidance and $\omega_{\text{id}}(p)$ controls the additional identity contribution at normalized sampling progress $p \in [0, 1]$. Conditioning dropout exposes the model to the v_{id0} configuration during training, so this branch is a learned zero-identity prediction of our model rather than of the unmodified Wan-S2V. Strong identity guidance transfers more of the subject-specific cues in the reference clips, but keeping it strong through the final denoising steps can introduce visual artifacts. We therefore decrease ω_{id} over sampling, applying stronger identity guidance early and reducing it during later refinement, following the piecewise-constant schedule given in Appendix C.

4 Experiments

4.1 Implementation Details

We train our model on Seamless Interaction (Agrawal et al., 2025), using 3,540 subjects and 3,814,624 five-second clips after quality filtering. Reference backgrounds are masked with MediaPipe (Lugaresi et al., 2019). Our model combines the 14B-parameter Wan-S2V generator with a domain-adapted V-JEPA 2.1 encoder and 256 identity tokens. We train the identity encoder and conditioning layers using same-clip conditioning with one reference clip, followed by cross-clip conditioning with up to 16 reference clips, keeping the adapted V-JEPA and generator frozen. At inference, we encode each reference set once and generate 81-frame videos at 480×832 resolution

and 16 FPS using dual classifier-free guidance. Full dataset and preprocessing details are provided in Appendix B, and model, training, and inference settings in Appendix C.

4.2 COMPARISONS

Test set. We evaluate on 119 unseen subjects and 581 target clips from the identity-filtered Seamless Interaction test partition. Each target clip is a real test video whose audio drives generation and whose visual content serves as ground truth for evaluation. The reference videos do not temporally overlap with their corresponding target clips. We generate one video per target clip using a single fixed random seed, yielding 581 videos per evaluated configuration.

Table 1: Reference-video inputs of the Wan-S2V configurations. The preceding clip is the ground-truth clip immediately before the target, while identity clips are always separate from it.

Configuration	FramePack	Identity clips
Image Reference	–	–
Direct Video Reference	preceding clip	–
Direct + **Ours**	preceding clip	8
Non-Direct Video Reference	separate clip	–
Ours (1 ref)	–	1
Ours	–	8
Non-Direct + **Ours**	separate clip	8

Compared configurations. We compare in two settings. First, we evaluate configurations of Wan-S2V that share the same frozen generator but differ in their reference-video conditioning. Table 1 defines them. FramePack (Zhang et al., 2025) supports autoregressive video continuation. *Direct* uses the ground-truth clip immediately preceding the target, while *Non-Direct* uses a separate clip of the same subject. Direct is a non-deployable oracle because its input is unavailable when generating novel content. Second, we compare against external methods, which use their own generators, so differences there reflect whole systems rather than the conditioning alone. *MimicTalk* (Ye et al., 2024) adapts a subject-specific renderer to each person, which requires per-subject optimization rather than a single feed-forward pass. We compare with *OmniTalker* (Wang et al., 2025), which uses a short audio-video reference to model speaking style. We additionally evaluate the released pre-trained *TAVR* (Guo et al., 2026) model, which uses reference-video tokens and reference audio to condition a Wan backbone for audio-driven generation. Both take a single reference clip, which the Ours (1 ref) row matches.

Metrics. Our evaluation builds on the protocol of OmniTalker (Wang et al., 2025). We measure appearance and identity using cosine similarity (CSIM) between ArcFace (Deng et al., 2022) embeddings. *CSIM to ref* compares generated frames with the appearance reference image used to condition generation, while *CSIM to GT* compares them with frames from the corresponding ground-truth target clip. Expression-FID (E-FID), Pose-FID (P-FID), and Mesh-FID are Fréchet distances between generated and target distributions of per-frame expression coefficients, head-pose parameters, and 3D facial landmarks, respectively. We report Fréchet Video Distance (FVD) using two real-video reference distributions. $\mathrm{FVD}_{\mathrm{target}}$ compares generated clips with each subject's corresponding ground-truth target clips, reflecting similarity in both appearance and motion. $\mathrm{FVD}_{\mathrm{test\ set}}$ compares generated clips with the entire real test set to assess dataset-level video realism rather than fidelity to a particular subject. Mean Sync-C measures audiovisual synchronization. Higher values are better for CSIM and Sync-C, while lower values are better for E-FID, P-FID, Mesh-FID, and both FVD variants. Feature extraction, frame pairing, preprocessing, and metric-specific aggregation details are provided in Appendix C.2.

Qualitative comparison. We provide video comparisons in an HTML viewer included with the supplementary material. These videos allow readers to compare facial expressions, head movements, lip synchronization, and temporal consistency throughout the generated sequences. In our comparisons, MimicTalk, which relies on a subject-specific NeRF renderer rather than a pre-trained video diffusion model, produces lower visual quality than other methods. TAVR generates visually convincing videos, but our method more faithfully captures the subject's characteristic behavior, including expressions and head movements observed in the reference videos. OmniTalker's redubbing outputs closely follow the reference clip's head movements and temporal progression, while adapting mouth movements and facial expressions to the driving audio. Although suitable for preserving a source performance during redubbing, this keeps the generated behavior closely tied to that particular clip. Our method instead aggregates subject-specific cues across unaligned reference videos to condition generation without prescribing a framewise motion trajectory.

Quantitative comparison. Table 2 reports the quantitative comparison over the complete evaluation set. We evaluate MimicTalk, OmniTalker, and TAVR as external reference-video baselines alongside controlled comparisons within Wan-S2V. Compared with Image Reference, Ours improves appearance and identity similarity (CSIM) and more closely matches the subject's expres-

Table 2: Quantitative comparison on unseen Seamless Interaction subjects. Table 1 defines the configurations, of which the two Direct rows are non-deployable oracles. CSIM measures appearance similarity, to the conditioning image and to ground-truth frames. E-FID, P-FID, and Mesh-FID compare expression, head-pose, and facial-geometry distributions, while the two FVD variants compare video distributions against each subject and against the whole real test set.

	Appearance and Identity		Subject Dynamics				Video Quality	
	CSIM to ref ↑	CSIM to GT ↑	E-FID ↓	P-FID ↓	Mesh-FID ↓	$\mathrm{FVD}_{\mathrm{target}}$ ↓	$\mathrm{FVD}_{\mathrm{test\,set}}$ ↓	Sync-C ↑
Image Reference	0.582	0.589	0.98	9.99	6.41	218.121	316.65	6.421
Direct Video Reference	0.653	0.671	0.86	8.45	4.70	127.016	238.31	6.339
Direct + **Ours**	0.658	0.680	0.63	7.15	3.72	104.941	228.42	6.262
Non-Direct Video Reference	**0.694**	**0.665**	0.82	10.75	5.96	147.524	278.67	6.341
MimicTalk	0.525	0.531	0.89	9.32	6.48	283.872	414.849	6.005
OmniTalker	0.614	0.609	0.81	9.41	5.21	137.842	**259.53**	6.305
TAVR	0.563	0.542	0.92	10.25	6.13	248.864	283.72	**6.439**
Ours (1 ref)	0.613	0.618	0.85	8.10	5.08	184.984	297.83	6.403
Ours	0.631	0.638	**0.72**	**7.02**	**4.52**	165.277	290.16	6.419
Non-Direct + **Ours**	0.693	0.664	0.73	7.78	4.67	**126.727**	271.48	6.276

sion, head-pose, and facial-geometry distributions (E-FID, P-FID, and Mesh-FID). It also lowers both FVD scores while obtaining a Sync-C score numerically close to that of Image Reference.

Direct and Non-Direct Video Reference achieve higher CSIM than Ours, consistent with the pixel-level appearance cues supplied by their FramePack reference frames. Their E-FID, P-FID, and Mesh-FID scores, however, remain higher than those of our method. Non-Direct Video Reference even has a higher P-FID than Image Reference, consistent with carrying over reference head poses that may differ from the target distribution. Adding our identity conditioning to either FramePack configuration lowers E-FID, P-FID, Mesh-FID, and both FVD scores relative to the corresponding baseline. CSIM remains nearly unchanged in the Non-Direct configuration, with small numerical gains in the Direct configuration. Both combinations obtain lower Sync-C than their corresponding FramePack baselines, indicating a trade-off rather than an improvement across all evaluation metrics.

MimicTalk obtains the worst $\mathrm{FVD}_{\mathrm{test\,set}}$ among the evaluated methods. TAVR achieves a lower $\mathrm{FVD}_{\mathrm{test\,set}}$ than Image Reference, but also lower CSIM scores. Compared with MimicTalk, OmniTalker, and TAVR, Ours achieves higher appearance and identity similarity and more closely matches the target expression, head-pose, and facial-geometry distributions. OmniTalker's higher P-FID is similarly consistent with its redubbing outputs retaining source head movements that may differ from the target pose distribution.

User study. In a pairwise user study, participants preferred Non-Direct + Ours over Non-Direct Video Reference, MimicTalk, OmniTalker, and TAVR for appearance similarity, behavioral similarity, and overall preference. The protocol and results are provided in Appendix E.

4.3 Ablation Studies

The ablations in Tables 3 and 4 follow the evaluation protocol of the main comparisons, using the same 119 subjects and 581 target clips, with all inputs and generation settings held fixed except for the factor under study. They report CSIM to ref, E-FID, and P-FID, using the metric definitions in Table 2 and the aggregation rules in Appendix C.2. The guidance ablation additionally reports $\mathrm{FVD}_{\mathrm{test\,set}}$ and Sync-C to assess video realism and audiovisual synchronization.

Table 3: Reference-video encoder and training-procedure ablations. Encoder variants use same-clip conditioning followed by cross-clip conditioning, while training variants use domain-adapted V-JEPA 2.1.

	Appearance	Subject Dynamics	
Configuration	CSIM to ref ↑	E-FID ↓	P-FID ↓
Reference-video encoder			
DINOv3	**0.639**	0.84	8.43
V-JEPA 2.1	0.613	0.79	7.87
FT V-JEPA 2.1	0.631	**0.72**	**7.02**
Identity encoder training procedure			
Same-clip only	0.619	0.87	9.23
Cross-clip only	0.596	0.95	9.55
Same-clip then cross-clip	**0.631**	**0.72**	**7.02**

Video encoder. The first part of Table 3 compares DINOv3 (Siméoni et al., 2025), an image encoder of comparable scale, with the off-the-shelf and domain fine-tuned V-JEPA 2.1 encoders. All three use same-clip conditioning followed by cross-clip conditioning and the same identity encoder architecture, and each receives 64 uniformly sampled frames per reference clip. For DINOv3, we encode the frames independently and concatenate their patch features before applying the same Spatio-Temporal Identity Encoder, with the V-JEPA 2.1 input protocol given in Appendix C and the DINOv3 checkpoint in Appendix C.1. DINOv3 achieves the highest appearance similarity among the three encoders, while both V-JEPA variants yield better



Figure 4: Effect of removing a smiling (left group) or laughing (right group) reference video. Each group shows two time steps of a different subject, comparing ground truth (GT), generation with all references, and generation with the specified reference removed.

expression and head-pose results. Since both encoders receive the same frames and feed the same identity encoder, the gain in subject dynamics comes from modeling the frames jointly over time rather than from stronger per-frame features. Domain fine-tuning improves all three metrics over unadapted V-JEPA, yielding the lowest expression and head-pose distances among these encoders.

Training procedure. With the domain fine-tuned encoder fixed, the second part of the table compares same-clip conditioning only, cross-clip conditioning only, and the two in sequence, using the settings in Appendix C.1. Same-clip conditioning alone outperforms cross-clip conditioning alone on all three metrics, consistent with the benefit of an aligned learning signal for establishing reference conditioning. Running the two in sequence improves appearance similarity and lowers both facial-feature distances further.

Dual classifier-free guidance. Table 4 compares standard classifier-free guidance, constant dual guidance at several identity scales, and scheduled dual guidance for the same trained model. The guidance settings appear in Appendix C.1. Standard classifier-free guidance applies one weight to the joint conditioning signal, while the two dual-guidance variants assign an independent weight to the identity contribution. Standard CFG improves appearance similarity and lowers expression and head-pose distances relative to Image Reference.

Table 4: Identity-guidance ablation. Parentheses denote the constant identity-guidance scale. The scheduled variant decreases this scale during sampling.

	Appearance	Dynamics		Video Quality	
Guidance	CSIM to ref ↑	E-FID ↓	P-FID ↓	FVD ↓	Sync-C ↑
Standard CFG	0.597	0.93	9.05	**286.76**	6.414
Constant dual (3)	0.609	0.89	9.01	302.57	6.402
Constant dual (7.5)	0.615	0.83	8.32	478.24	6.365
Constant dual (9)	0.547	0.78	7.46	548.89	6.287
Scheduled dual CFG	**0.631**	**0.72**	**7.02**	290.16	**6.419**

Increasing the constant identity-guidance scale lowers expression and head-pose distribution distances, but increases $\mathrm{FVD}_{\text{test set}}$ and reduces Sync-C. At the highest tested scale, appearance similarity also deteriorates. The scheduled variant, Ours in Table 2, achieves the highest CSIM and lowest E-FID and P-FID among the tested settings, while keeping $\mathrm{FVD}_{\text{test set}}$ and Sync-C numerically close to standard CFG. These results support decreasing identity guidance during sampling to balance appearance and behavioral matching with video quality and audiovisual synchronization.

4.4 Reference Set Analysis

Content of reference videos. We next examine how the content of the reference videos influences the generated behavior. Figure 4 compares generation using all references with generation after removing a smiling reference video for one subject and a laughing reference video for another, alongside the corresponding ground-truth frames. Smiling and laughter remain visible after removal, but the outputs differ in mouth shape, eye closure, and head pose. These examples illustrate how the reference set influences the way expressions are rendered and motivate selecting references that cover a subject's range of expressions and poses.

Number of reference videos. Table 5 compares the zero-reference Image Reference baseline with our model conditioned on 1, 2, 4, 8, and 16 videos of the same subject. The comparison uses the main evaluation protocol with the same 119 subjects and 581 target clips, holding the trained model and all other inputs and generation settings fixed across nonzero reference counts. We report CSIM to ref, E-FID, and P-FID using the metric definitions in Table 2 and the aggregation rules in Appendix C.2.

The largest improvements in all three metrics occur when moving from the zero-reference baseline to a single reference video. Appearance similarity largely saturates beyond four references, while expression and head-pose distances continue to decrease up to eight in this evaluation. Increasing the count from eight to sixteen yields only marginal improvements in appearance similarity and expression distance, while head-pose distance increases slightly.

4.5 Representation Analysis

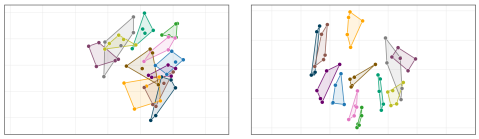
We examine how domain fine-tuning changes the geometry of V-JEPA 2.1 backbone features before the learned identity encoder and generator. Appendix D details the evaluation protocol. Following Wang & Isola (2020), uniformity U measures overall representation spread (lower means broader), while same-person alignment A is the mean squared distance between same-person clip representations (lower means closer). Table 6 shows that fine-tuning lowers U but increases A, indicating greater dispersion both overall and within identities rather than tighter identity grouping. Figure 5 provides a complementary PCA visualization of selected subjects. Although same-person distances increase on average in the full feature space, clips of the same person still tend to group together relative to clips of other people in the fine-tuned projection. The more differentiated clip representations may help the identity encoder combine complementary observations of the same subject. Consistent with the value of multiple references, Table 5 shows continued improvements in expression matching even after appearance similarity largely saturates.

Table 5: Effect of the number of reference clips from the same subject. Zero references corresponds to the Image Reference baseline.

	Appearance	Subject Dynamics	
References	CSIM to ref ↑	E-FID ↓	P-FID ↓
0	0.582	0.98	9.99
1	0.613	0.85	8.10
2	0.624	0.78	7.76
4	0.628	0.75	7.27
8	0.631	0.72	7.02
16	0.632	0.71	7.09

Table 6: V-JEPA 2.1 representation geometry. Lower U indicates broader spread, while lower A indicates closer same-person representations. Evaluation details are in Appendix D.

Encoder	U	A
Original	-0.640	0.107
Fine-tuned	-0.822	0.148



Pretrained Finetuned

Figure 5: Separately fitted PCA projections of V-JEPA 2.1 clip representations before (left) and after (right) domain fine-tuning. Colors identify subjects, points denote clips, and shaded polygons outline subject groups.

5 Conclusions, Limitations and Future Work

We introduced a method for personalizing audio-driven talking-head generation beyond visual appearance, capturing characteristic behavior from reference videos. Our Spatio-Temporal Identity Encoder aggregates complementary observations from a variable number of clips into compact identity tokens that condition a pre-trained video diffusion model. This enables generation with new speech while drawing on recurring subject-specific cues observed across different reference videos.

Our experiments show that reference-video conditioning improves appearance similarity and more closely captures the subject's expression and head-pose distributions compared with image-only conditioning and competing reference-based methods. Our analysis further shows that multiple reference videos provide complementary information, and that the behaviors represented in the reference set influence the resulting performance. Together, these results support a broader view of personalization that captures not only how a person looks, but also how they behave.

Our current model is trained on a relatively constrained domain of dyadic conversations, leaving substantial room to extend behavioral personalization to more diverse recording conditions and behaviors. Such an extension will require datasets with sufficient footage of each individual across different environments, viewpoints, and interactions. More broadly, we believe that modeling the dynamics of an individual is an interesting and still underexplored dimension of generative personalization. The long-standing success of gait recognition provides a striking example: characteristic patterns of motion can carry a strong signature of identity, and can remain informative even when facial appearance is difficult to resolve Nixon et al. (2006). Our work explores this principle at the finer scale of facial expressions and head motion during conversation. Extending it to richer behavioral cues and understanding which aspects of human dynamics are persistent and identity-specific opens an interesting direction for future research.

AI USE STATEMENT

AI assistants supported implementation development, evaluation scripts, literature searches, manuscript drafting and editing, and figure preparation. The authors are responsible for verifying the implementation, references, experimental results, and final text, and retain full responsibility for the work. The generated talking-head videos are the experimental outputs evaluated in this study.

ETHICS STATEMENT

Personalizing talking-head generation can support creative media production, but reproducing a person's appearance and characteristic behavior also creates risks of impersonation, nonconsensual depictions, and deceptive attribution of speech. Our experiments use the Seamless Interaction dataset, with data partitions and preprocessing described in Appendix B. The limited training domain and dependence on available reference footage are discussed in Section 5.

REPRODUCIBILITY STATEMENT

Section 3 describes the model architecture, three-stage training procedure, and inference guidance, while Section 4 presents the evaluation protocol and compared configurations. Appendix A provides architectural details, Appendix B documents data partitions and preprocessing, and Appendix C describes training and inference settings. Ablation configurations and metric computation are described in Appendices C.1 and C.2, respectively. Appendix D specifies feature extraction and statistical measures for the encoder representation analysis.

REFERENCES

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3087709. URL http://dx.doi.org/10.1109/TPAMI.2021.3087709.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow

transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

Zujin Guo, Zhenhui Ye, Yi Ren, Yuanming Li, Ce Chen, Zhibin Hong, and Chen Change Loy. Generate your talking avatar from video reference. *arXiv preprint arXiv:2604.27918*, 2026. URL https://arxiv.org/abs/2604.27918.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8153–8163, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6689–6700, 2024.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. StyleTalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1896–1904, 2023.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Mark S. Nixon, Tieniu Tan, and Rama Chellappa. *Human Identification Based on Gait*, volume 4 of *International Series on Biometrics*. Springer, 2006. doi: 10.1007/978-0-387-29488-9.

Or Patashnik, Rinon Gal, Daniil Ostashev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Nested attention: Semantic-aware attention values for concept personalization, 2025. URL https://arxiv.org/abs/2501.01407.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Guocheng Qian, Kuan-Chieh Wang, Or Patashnik, Negin Heravi, Daniil Ostashev, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Omni-id: Holistic identity representation designed for generative tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8786–8795, 2025.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: 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.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-Jin Liu. DiffPoseTalk: Speech-driven stylistic 3D facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics*, 43(4):1–9, 2024.

Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. Say anything with any style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5088–5096, 2024.

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3D facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20564–20574, 2023.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

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.

Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation i n seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Zhongjian Wang, Peng Zhang, Jinwei Qi, Guangyuan Wang, Chaonan Ji, Sheng Xu, Bang Zhang, and Liefeng Bo. OmniTalker: One-shot real-time text-driven talking audio-video generation with multimodal style mimicking. *Advances in Neural Information Processing Systems*, 38:21834–21862, 2025.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6537–6549, 2024b.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024a.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024b.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024c.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Zhenhui Ye, Tianyun Zhong, Yi Ren, Ziyue Jiang, Jiawei Huang, Rongjie Huang, Jinglin Liu, Jinzheng He, Chen Zhang, Zehan Wang, Xize Chen, Xiang Yin, and Zhou Zhao. MimicTalk: Mimicking a personalized and expressive 3D talking face in minutes. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12978–12988, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 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.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024.

A Architecture Details

This section provides the details of the Spatio-Temporal Identity Encoder that are omitted from Section 3. The frozen V-JEPA 2.1 encoder produces a spatio-temporal grid of patch features for each reference clip, which we flatten into $P = T' \cdot H' \cdot W'$ tokens of width d, where T', H', and W' are the temporal and spatial extents of the feature grid.

The patch features of each clip are first processed independently by three self-attention blocks with weights shared across all reference clips. We apply three-dimensional rotary position encodings (3D RoPE) in these blocks so that attention is aware of the spatial and temporal coordinates of the patches.

We then concatenate the refined features of all clips and aggregate them with Q learnable query tokens through two aggregation blocks. Each block applies cross-attention from the queries to the concatenated clip features, with 3D RoPE applied to the keys, followed by self-attention among the queries. To assign non-overlapping positions to different clips, we shift the temporal coordinate τ of clip i to $\tau + i\,T'$ before applying the positional encoding to the keys. This lets the queries attend jointly across all clips while still distinguishing them. We randomly permute the order of the reference clips during training so that no clip index acquires a consistent role. A final layer normalization over the query tokens produces the identity tokens $Z \in \mathbb{R}^{Q \times d}$, whose number is independent of the number of reference clips. At inference, we compute Z once from the reference set of each subject and reuse it throughout all denoising steps.

B Dataset and Preprocessing

We use both the *Naturalistic* and *Improvised* portions of the Seamless Interaction dataset (Agrawal et al., 2025), which contains in-person dyadic conversations. We follow its official data partitions and use only subjects in the training partition for V-JEPA domain adaptation and both downstream training stages. We form the evaluation split from the official test partition and exclude any identity that appears in the training manifests.

Source recordings are divided into nominal five-second clips at their native 29.97 or 30 frames per second, and incomplete final segments are discarded. For each clip, we independently generate a vertical $9{:}16$ target crop and a face-centered $1{:}1$ reference crop from the same skeleton track. A fixed crop over the complete clip extends from the subject's shoulders to above the head and is expanded to the required aspect ratio. Before V-JEPA encoding, we use MediaPipe (Lugaresi et al., 2019) to mask the background of each reference crop so that the reference pathway observes the foreground subject rather than the scene background. The associated participant-specific audio is stored as a synchronized mono waveform at 16 kHz. We require a peak-window audio RMS of at least 0.001, facial-keypoint confidence of at least 0.9, a minimum face-area fraction of 0.05, and at most $1.5\times$ spatial upsampling. We also apply a blur threshold of 75 with a luminance gate of 90 and remove clips with failed tracking or quality measurements.

We construct two training manifests for identity conditioning from the training partition. For same-clip conditioning, each target uses a temporally aligned reference from the same source video. For cross-clip conditioning, each target is paired with 1, 2, 4, 8, or 16 reference clips of the same subject that do not temporally overlap with the target clip. The reference clips are drawn from other segments of the same longer recording. They are selected without replacement, and the resulting training manifests remain fixed across epochs.

After preprocessing, the training partition contains $3{,}540$ subjects and $3{,}814{,}624$ unique five-second clips, corresponding to approximately $5{,}298.09$ nominal hours of video. The same-clip conditioning and cross-clip conditioning manifests contain $3{,}814{,}624$ and $584{,}049$ training samples, respectively. The same-clip conditioning manifest contains $2{,}408{,}707$ *Naturalistic* and $1{,}405{,}917$ *Improvised* samples, while the cross-clip conditioning manifest contains $370{,}335$ and $213{,}714$, respectively. After identity-level filtering, the evaluation set contains 119 unseen subjects and 581 target clips.

C Implementation Details

Reference encoder. We build on the 14B-parameter Wan-S2V generator and the V-JEPA 2.1 ViT-g/16 reference encoder. The generator operates on 9:16 videos, while V-JEPA receives corresponding face-centered 1:1 crops to match its pre-training aspect ratio. V-JEPA uniformly samples 64 frames from each five-second reference crop and processes them at 384×384 resolution. We adapt the reference encoder for 3 epochs using the V-JEPA 2.1 objective, a learning rate of 1×10^{-4}, and masked face crops from the training set. The adapted encoder produces a spatio-temporal patch grid with feature width d=1408.

Identity-conditioning model. The Spatio-Temporal Identity Encoder uses Q=256 query tokens, three shared self-attention blocks for each clip, and two query aggregation blocks. We inject the resulting identity tokens into every Wan-S2V transformer block with λ=1.

Optimization. We train the identity encoder and injection layers using same-clip conditioning followed by cross-clip conditioning illustrated in Figure 3. Both stages use AdamW with a learning rate of 5×10^{-5} and an effective batch size of 1024, achieved through gradient accumulation. Same-clip conditioning runs for 10,000 optimization steps. Cross-clip conditioning initializes from the same-clip conditioning checkpoint and runs for another 10,000 steps with 1, 2, 4, 8, or 16 reference clips. During both stages, we apply identity dropout with probability 0.05, audio dropout with probability 0.1, and text dropout with probability 0.5. We train on NVIDIA A100 GPUs using BF16 numerical precision. We precompute and cache the frozen V-JEPA features and VAE latents for all training clips.

Inference. We generate 81-frame videos at 480×832 resolution and 16 frames per second using 40 denoising steps with the default Wan-S2V sampler. All positively conditioned predictions use the fixed prompt "`a video of a person`", while the negative branch uses an empty prompt. We set ω_{S2V}=5. We use a piecewise-constant identity-guidance schedule. The weight is 9 for the first 25% of sampling, 5 from 25% to 50%, 3 from 50% to 75%, 1.5 from 75% to 90%, and 0.75 for the final 10%. Encoding one subject's reference set of eight videos takes a median of 7.90 seconds over five runs and is performed once before generation. This includes preprocessing, host-to-GPU transfer, V-JEPA encoding, and joint query pooling.

OmniTalker baseline. We generated the OmniTalker (Wang et al., 2025) comparison videos using its public web interface.[1]

C.1 Ablation Implementation Details

DINOv3 encoder. The DINOv3 ablation uses the checkpoint `facebook/dinov3-vith16plus-pretrain-lvd1689m`.[2] We uniformly sample 64 frames from each reference clip, encode them independently, and concatenate their patch features before applying the Spatio-Temporal Identity Encoder.

Training variants. The variants using same-clip conditioning only or cross-clip conditioning only each use 10,000 optimization steps. The sequential variant uses 10,000 steps of same-clip conditioning followed by 10,000 steps of cross-clip conditioning, for 20,000 steps in total. All variants use the AdamW optimizer, learning rate, and effective batch size specified above.

Guidance settings. Standard CFG uses a shared guidance scale of 5.0, matching the Wan-S2V default. For dual guidance, we keep ω_{S2V}=5.0 and compare constant identity scales of 3, 7.5, and 9 with the decreasing schedule specified above.

[1] `https://omnitalker.com/`

[2] `https://huggingface.co/facebook/dinov3-vith16plus-pretrain-lvd1689m`

C.2 Metric Implementation Details

Appearance and identity. We extract ArcFace (Deng et al., 2022) embeddings using InsightFace's `buffalo_l` model pack, with the `w600k_r50.onnx` recognizer and `det_10g.onnx` detector. We process native-resolution frames with a 640×640 detector input, select the largest detected face, and apply five-landmark similarity alignment to the standard 112×112 ArcFace template. We ℓ_2-normalize the 512-dimensional embeddings and compute cosine similarity. Each video is independently sampled to at most 32 evenly spaced frames. CSIM to ref compares the embedding of each generated frame with that of the appearance reference image used to condition generation. CSIM to GT compares generated frames with frames from the corresponding ground-truth target clip, paired using their sampled-list indices rather than timestamps. Only comparisons with successful face detection in both images contribute to the mean. We average frame-level similarities within each clip and then average the 581 clip-level scores with equal weight.

Expression, head pose, and facial geometry. All three metrics use Google's float16 MediaPipe Face Landmarker checkpoint (SHA256 prefix `64184e229b26`) in IMAGE mode with one face and minimum detection confidence 0.5. We process native RGB frames without external cropping or alignment. E-FID uses the raw 52-dimensional blendshape scores without normalization. P-FID uses a 6-dimensional vector of rotation and translation estimated with OpenCV's `solvePnP` from six face landmarks, with featurewise standardization fitted jointly to generated and GT descriptors. Mesh-FID uses all 478 normalized three-dimensional landmarks, including iris points, concatenated into a 1434-dimensional vector. Generated and GT mesh descriptors are jointly centered and projected using PCA fitted to their pooled descriptors, retaining at most 32 components without ℓ_2 normalization. E-FID, P-FID, and Mesh-FID compare generated descriptors with descriptors from the ground-truth target clips. We pool per-frame descriptors across the clips of each subject and compute each distance between that subject's generated and ground-truth descriptor distributions. We then average the 119 subject-level distances with equal weight. We sample at most 32 uniformly spaced frames across each full clip.

Video quality and audiovisual synchronization. Both FVD protocols use 400-dimensional pre-softmax RGB I3D features from the Kinetics-400 checkpoint `flateon/FVD-I3D-torchscript`, revision `1c2d617`. We sample exactly 77 frames at 16 frames per second over the first 4.8125 seconds, normalize RGB values to $[-1, 1]$, and bilinearly resize the entire frame to 224×224 without preserving its aspect ratio. Short or undecodable clips raise an error and are not padded. We compute $\mathrm{FVD}_{\mathrm{target}}$ separately for each subject between the video-feature distributions of their generated clips and corresponding ground-truth target clips, then report the unweighted mean across the 119 subjects. For $\mathrm{FVD}_{\mathrm{test\,set}}$, we pool video features from all 581 generated clips across subjects for each configuration. We compute a single FVD between this distribution and the pooled video-feature distribution of the entire real test set, containing 70,893 clips from 128 subjects. Both FVD variants are sensitive to appearance and motion, and neither isolates visual quality alone. Sync-C uses the original SyncNet v2 (LSE-C) evaluator with S3FD face detection (SHA256 prefixes `961e8696f888` and `d54a87c2b754`, respectively). We resample the first 4.8125 seconds to 25 frames per second and resize the tracked face crops to 224×224 using the evaluator's asymmetric cropping pipeline. Audio is taken from the original driving WAV, converted to mono 16 kHz PCM, and represented by 13-dimensional MFCCs with a 10 ms hop. Each comparison window contains five video frames and twenty MFCC frames, covering 200 ms. The evaluator searches offsets from -15 to $+15$ video frames (±0.6 seconds) and computes confidence as the median offset-distance minus the minimum distance, without temporally correcting the video. The final Sync-C score is the unweighted mean of the 581 clip-level scores.

D Representation Analysis Details

Evaluation cohort and features. The analysis in Section 4.5 uses a separate cohort of 527 reference clips from 30 identities excluded from our training data, rather than the target clips used for the generation comparisons. For both frozen checkpoints, we encode the same original, unmasked videos and compute the 1408-dimensional mean over all final-layer tokens, including background locations. We then ℓ_2-normalize each clip representation. The resulting features precede the learned

identity encoder and generator, and all quantitative metrics are computed in this normalized feature space rather than in PCA coordinates.

Uniformity and same-person alignment. Following Wang & Isola (2020), we compute uniformity U as the natural logarithm of the mean of $\exp(-t\|x_i - x_j\|_2^2)$ over all unordered pairs of distinct clips, where x_i is a normalized clip representation and $t{=}2$. This includes both same-person and different-person pairs, with self-pairs excluded and equal weight per pair. Same-person alignment A is the mean of $\|x_i - x_j\|_2^\alpha$ over distinct clips of the same identity, with $\alpha{=}2$ and equal weight per same-person pair. These same-person pairs are evaluation positives, not the training pairs used by V-JEPA. Uniformity and alignment are evaluation statistics, not training objectives.

PCA visualization. Figure 5 uses the same five test clips from each of 12 selected subjects in both panels. The clip representations are the normalized all-token means described above, with PCA fitted separately for each encoder and projected onto two dimensions. This selected visualization is qualitative rather than a summary of the full quantitative cohort. Colors identify subjects, and shaded polygons outline their clip groups rather than confidence regions.

E User Study

Protocol. We collected 240 pairwise comparison responses from 24 unique participants across videos generated using 12 distinct source audios, with approximately 60 comparisons against each baseline. Each trial presented the ground-truth target clip and two generated videos, one from Non-Direct + Ours and one from Non-Direct Video Reference (FramePack), MimicTalk, OmniTalker, or TAVR. The A/B presentation order was randomized. Each response contained three required A/B judgments, with no tie option. Appearance similarity asks which generated subject more closely resembles the person in the ground-truth video, considering facial identity and other stable appearance characteristics. Behavioral similarity asks which subject's expressions and movements more closely resemble those of the person in the ground-truth video, considering facial expressions, head movements, gestures, and movement style rather than matching exact actions or timing. Overall preference asks which generated video is preferred, considering appearance and behavioral similarity, visual quality, and how well lip movements match the speech.

Table 7: User-study preference for Non-Direct + Ours over each baseline, rounded to whole percentages. FramePack denotes Non-Direct Video Reference.

Compared against	Appearance	Behavior	Overall preference
FramePack	58%	74%	73%
MimicTalk	94%	94%	97%
OmniTalker	64%	73%	72%
TAVR	97%	97%	90%

Results. Table 7 reports the percentage of preferences for Non-Direct + Ours over each baseline, rounded to whole percentages. Participants preferred the combined configuration on all three criteria against every evaluated baseline. Against Non-Direct Video Reference, the preference was more pronounced for behavioral similarity and overall preference than for appearance similarity, supporting the value of identity conditioning beyond pixel-level reference input.

F Additional Related Work

Motion Transfer and Customization. AnimateAnyone (Hu, 2024) and MagicAnimate (Xu et al., 2024c) animate images from pose sequences, while LivePortrait (Guo et al., 2024) and Wan-Animate (Cheng et al., 2025) transfer motion from driving videos. DreamVideo (Wei et al., 2024b) and MotionDirector (Zhao et al., 2024) learn reusable subject or motion concepts but require per-concept optimization. Our method instead generates audio-driven motion conditioned on unaligned reference clips, without a prescribed motion trajectory or subject-specific optimization.