# REPRISE: IDENTITY-FAITHFUL HUMAN VIDEO GENERATION FROM IMAGE AND VIDEO REFERENCES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Subject-driven video generation has made substantial progress in synthesizing videos of a referenced person in novel scenes, yet existing methods typically condition on a single reference image and measure identity by facial similarity between the generated face and that same reference. A reference captures the person under one specific viewpoint, styling, lighting, and expression, and is thus a sample of an identity rather than the identity itself, so same-reference protocols reward copying such transient attributes rather than preserving the person. The issue becomes more severe in multi-person scenarios, where no token-level cue binds each subject to its own references, leading to subject confusion and attribute swapping. We propose **Reprise**, a multi-reference identity-preserving video generation framework in which each character is specified by multiple image and video references from different contexts, and identity is defined as what remains invariant across these contexts. We construct **Reprise-10K**, a corpus of 10K identities derived from film and television, in which each identity is paired with multiple reference images and video clips spanning distinct viewpoints, lighting, expressions, and styling. We further introduce **ID-RoPE**, which extends rotary position embedding with an additional identity coordinate carried by every token, grouping attention by character so that all references of a character are attended jointly rather than any single one being favored, and avoiding the copying shortcut present in existing reference-conditioned designs. A **cross-context identity reward** post-trains the model by scoring generated faces against held-out references of each identity, steering the model toward generating the person itself rather than copying the conditioning references. Extensive experiments demonstrate that Reprise improves identity preservation, multi-person subject binding, and naturalness, while reducing identity confusion and excessive copying of the conditioning references.

## 1 INTRODUCTION

Subject-driven video generation has made substantial progress in synthesizing videos of a referenced person in novel scenes, with applications in film production, advertising, and personalized media (Liu et al., 2025b; Deng et al., 2025; Li et al., 2025). The central challenge of the task is identity preservation, namely whether a viewer can recognize the same individual throughout the generated video despite changes in viewpoint, lighting, styling, and expression.

The prevailing practice extracts identity from a single reference image and injects it into a pretrained video backbone, either through dedicated identity encoders (Liu et al., 2025b; Li et al., 2025) or by concatenating reference latents into the generation sequence (Zhong et al., 2025; Xue et al., 2025a; Deng et al., 2025). Identity preservation is then measured by facial similarity between the generated face and the same reference (Yuan et al., 2025).

However, a reference image captures the person under one particular viewpoint, styling, lighting, and expression at a single moment; it is a sample of an identity rather than the identity itself. Measuring identity by similarity to the conditioning reference therefore rewards copying these transient attributes rather than preserving the person, and pretrained video backbones readily realize such copying, since video pretraining is dominated by temporal coherence. Figure 1 shows where this



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.




Figure 1: **Reference-conditioned video generation behaves like compositing, not generation.** Left: the single reference given to every system. Right: four scenes, one per column, with the prompt printed underneath. Top row: a different recent reference-conditioned generator per column, named in the corner of the panel. Bottom row: Reprise. The prior systems return the reference itself rather than the person in it, keeping its crop, frontal pose, and facial lighting whether the prompt asks for a snowstorm, a station platform, a wheat field, or a moving bus; the subject sits in front of the scene instead of inside it and stays rigid while the scene moves around it. Reprise renders the same identity at the scale, viewpoint, and illumination the scene calls for, and lets it move with the scene.

leads: current systems return the reference photograph re-framed but not re-rendered, carrying its crop, pose, and lighting into every prompted scene. The subject is composited in front of the scene rather than placed inside it, and it holds still while the scene moves, so the result is a high reference similarity attached to a person who never behaves like one. The issue compounds in multi-person generation, where no existing mechanism reliably binds each subject to its own references.

We argue that the key missing ingredient is a definition of identity that copying cannot satisfy. We define identity as the component of appearance and motion that remains invariant across a person's appearances in different contexts. Under this definition, a single image is insufficient in principle: the invariant component can only be identified from multiple references drawn from different contexts, and motion characteristics are observable only in video references. The definition also determines the measurement: identity should be scored against cross-context references of the same person rather than against the conditioning ones.

To this end, we present Reprise, a multi-reference identity-preserving video generation framework in which each character is specified by multiple image and video references from different contexts. We first construct Reprise-10K, a corpus of 10K identities derived from film and television, in which each identity is paired with multiple reference images and reference videos spanning distinct viewpoints, lighting, expressions, and styling, and targets range from single-character to multi-character scenes. Because the references of an identity come from different episodes and scenes rather than from the target clip, the person is decoupled from the context in which the references were captured, and copying any single reference is no longer sufficient to reconstruct the target.

Leveraging this corpus, Reprise introduces two designs tailored for identity-faithful multi-reference generation. First, **ID-RoPE** extends rotary position embedding with an additional identity coordinate carried by every token, so that tokens of the same character receive a fixed attention affinity given directly by their coordinates. Attention is thereby grouped by character with no learnable parameters and no attention masks, and each subject attends to all of its references jointly instead of favoring any single one. All references are further placed outside the temporal and spatial index range of the target, which suppresses the copy pathway acquired in pretraining.

Second, we post-train the model with a **cross-context identity reward**, which makes the invariant component an explicitly optimized quantity rather than a byproduct of reconstruction. Instead of comparing the generated face with the conditioning references, we score it against a pool of references of the same identity withheld from conditioning and drawn from mutually disjoint contexts, whose context-specific components cancel under pooling and leave only the identity component, so that reproducing any conditioning reference no longer attains a high reward. Since this reward is a

differentiable face-embedding similarity, we optimize it by reward gradient, and each generated face is matched to and scored against the pool of its own character.

For evaluation, we add a cross-context identity measure that scores each generated face against references of the same person from contexts the model has not seen, so that it reflects the person rather than any single photograph. Under this protocol, Reprise improves identity preservation, multi-person subject binding, and naturalness over single-reference and multi-reference baselines (Table 1), and ablations validate each design choice.

Our main contributions are summarized as follows:

- We construct **Reprise-10K**, a corpus of 10K identities derived from film and television through cross-context reference pairing, providing multiple image and video references per identity and covering targets from single-character to multi-character scenes.
- We propose **Reprise**, a multi-character reference-conditioned video generation framework that supports multiple reference images and multiple reference videos per character. Within it, we introduce **ID-RoPE**, an additional identity coordinate carried by every token, which groups attention by character so that each subject is bound jointly to all of its references, and keeps references positionally apart from the target to suppress the copy pathway.
- We design a **cross-context identity reward** for post-training, which replaces the conditioning references with a held-out cross-context pool as the comparison target so that context-specific attributes cancel and only the identity component is rewarded.

## 2 Related Work

**Subject-driven human video generation.** Existing systems bind identity from photographs, taking one portrait or several images per subject and injecting them through cross-modal alignment, masked guidance, frozen MLLM states, or lightweight plug-and-play adapters (Liu et al., 2025b; Deng et al., 2025; Li et al., 2025; Zhang et al., 2025; Xing et al., 2026; Girish et al., 2025; Xue et al., 2025a). Concat-ID (Zhong et al., 2025) is the direct ancestor of the latent-concatenation conditioning we build on. Video-reference systems concatenate heterogeneous clips, compress them into slot tokens, or attend over full reference sequences, all in single-subject settings (Wang et al., 2026a; Lai et al., 2026; Liang et al., 2026; Meng et al., 2026; Xu et al., 2026; Guo et al., 2026; Zhang et al., 2026). A separate line uses video as a *driving signal*, as in motion transfer and portrait animation (Hu et al., 2024; Guo et al., 2024) and in VINO's motion and camera cloning (Chen et al., 2026), which copies trajectories rather than informing identity; our reference clips serve the opposite role. We target the multi-subject regime, and our static-repeat control (App. E) tests whether video references carry identity information beyond extra appearance views.

**RL post-training for generation.** Policy-gradient (Black et al., 2023; Liu et al., 2025a; Xue et al., 2025b) and preference (Wallace et al., 2023; Huang et al., 2026) families dominate, and reward-gradient methods (Xu et al., 2023; Clark et al., 2023; Prabhudesai et al., 2024; Yuan et al., 2023; Shen et al., 2025) form the sample-efficient tier. Multi-human identity RL rewards clips *globally* (Meng et al., 2025; Pan et al., 2025). The copy-paste pathology of same-source identity rewards is documented (Yuan et al., 2025; Chen et al., 2025); existing remedies pool frames of the same clip (Shen et al., 2025) or replace the scorer with a learned preference model (Meng et al., 2025), whereas our reward keeps the scorer and moves the comparison target to references withheld from conditioning and drawn from mutually disjoint contexts.

## 3 Method

### 3.1 Setup and Notation

Reprise generates a target video of $N$ characters, each specified by reference images and clips from different contexts (Figure 2). Text, reference, and target tokens form one sequence for the frozen Wan2.1 DiT (Wan Team, 2025; Zhong et al., 2025); target tokens attend to references but not vice versa, and a single LoRA is the only trainable component.

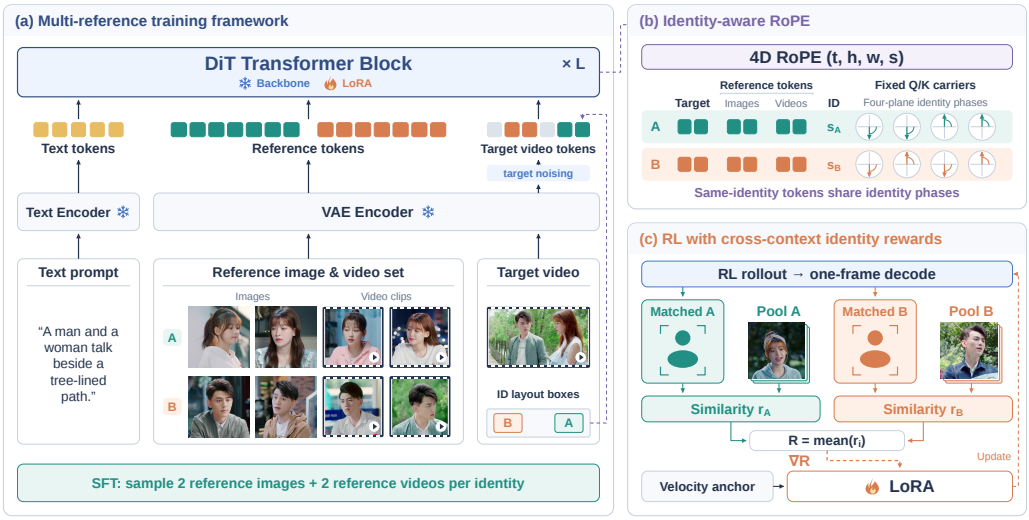



Figure 2: **Overview of Reprise.** (a) Multi-reference training: text, reference, and target tokens form one sequence for the frozen DiT with a single LoRA. (b) Identity-aware RoPE: every token carries a fourth coordinate $s$ as a four-plane phase code on fixed carriers, so the target region and all references of one character share a code. (c) Cross-context identity reward: generated faces are scored against pools of references withheld from conditioning and drawn from disjoint contexts, with the gradient taken through one decoded frame and the final denoising step, and a velocity anchor to the supervised model.

Rotary position embedding factorizes each attention head ($d_h$=128) into 64 complex planes: plane $p$ rotates queries and keys by an angular frequency $\omega_p$ applied to one coordinate $c_p(x)$ of token $x$, and the pre-softmax logit between tokens $x$ and $y$ decomposes as

$$\ell(x,y) = \frac{1}{\sqrt{d_h}} \sum_{p=1}^{64} \mathrm{Re}\Big[\, q_p(x)\, \overline{k_p(y)}\, e^{\,i\,\omega_p (c_p(x) - c_p(y))} \Big] = \ell_{\mathrm{cont}}(x,y) + B(x,y), \tag{1}$$

where $\ell_{\mathrm{cont}}$ and $B$ restrict the sum to the planes outside and inside a designated subset $\mathcal{C}$, respectively. Because the sum ranges over disjoint planes, reassigning the planes in $\mathcal{C}$ changes the logit only through the additive term $B$, without a bias tensor and without modifying the attention kernel. The frequency schedule and the $(22, 21, 21)$ split of the 64 planes over $(t, h, w)$ are given in App. C.

### 3.2 Multi-Reference Conditioning

Each reference, image or clip, is encoded by the same VAE as the target and concatenated into the attention sequence at full spatial resolution, so that the target reads all references of all characters within one self-attention. Video pretraining gives the backbone a strong prior for reading a token from the same spatial position in a neighboring frame; placing reference tokens at indices the target also occupies makes that prior directly applicable, since reproducing the token at the matching index becomes an inexpensive way to reduce the training loss, and a model that takes this route reproduces its references instead of generating the person they depict. Our placement therefore keeps every reference token outside the index range of the target. A target token at frame $f$, row $r$, and column $c$, and the $j$-th frame of reference asset $a$, receive

$$(t,h,w)_{\mathrm{target}} = (f,\ r,\ c), \qquad (t,h,w)_{a,j} = (F + \delta j,\ r + H,\ c), \tag{2}$$

with $F$=21 target latent frames, stride $\delta$=0.25, row shift $H$=30, and $j = 0, \ldots, L_a - 1$ for an $L_a$-frame asset ($L_a$=1 for a reference image); column indices are unchanged.

All assets start at $t$=$F$, so the time coordinate encodes frame order within an asset and nothing else; equal-length assets receive identical coordinates, so asset order carries no identity information and all grouping is left to §3.3. The stride is fractional so that no reference frame lands on an index the target occupies. The row shift $H$=30 makes the two spatial ranges disjoint. A reference image contributes 1,024 tokens and a 17-frame clip 17,408.

### 3.3 ID-RoPE: A Per-Token Identity Coordinate

The placement of §3.2 states where a token sits but not which character it belongs to: positional coordinates describe geometry alone, so two tokens that depict different characters at the same spatial position are indistinguishable under $(t, h, w)$. ID-RoPE extends the coordinate of every token to $(t, h, w, s)$, where $s{=}0$ denotes background and $s{=}n$ the $n$-th character, and contributes to the attention logit a term that depends on two tokens only through whether they carry the same identity, realized through the existing rotary path with no learned parameters. The $P{=}4$ slowest temporal planes are reassigned to the new axis, turning the $(22, 21, 21)$ split into $(18, 21, 21, 4)$, and on those planes the query and key content is replaced by a fixed unit-modulus carrier whose phase encodes the identity coordinate:

$$q_p(x) \leftarrow m_q^{(l)}\, e^{\,i\,\phi_p(s_x)}, \qquad k_p(y) \leftarrow m_k^{(l)}\, e^{\,i\,\phi_p(s_y)}, \qquad p \in \mathcal{C}, \tag{3}$$

applied after the per-head RMS normalization of queries and keys, with no temporal rotation on $\mathcal{C}$. The codes $u(s)$ form an orthogonal family, $\mathrm{Re}\langle u(s), u(s')\rangle = P \cdot \mathbf{1}[s{=}s']$ (construction in App. D), so the identity-axis contribution takes closed form:

$$B(x, y) = \frac{m_q^{(l)} m_k^{(l)}}{\sqrt{d_h}} \sum_{p \in \mathcal{C}} \cos(\phi_p(s_x) - \phi_p(s_y)) = \beta_l \cdot \mathbf{1}[\, s_x{=}s_y \,], \qquad \beta_l = \frac{P\, m_q^{(l)} m_k^{(l)}}{\sqrt{d_h}}, \tag{4}$$

that is, $\beta_l$ on every pair sharing a coordinate, including background with background, and exactly zero both across characters and between any character and the background, leaving those pairs entirely to the content term. On pairs that share a coordinate the term adds a uniform $+\beta_l$ to every logit, lifting each reference set as a whole by $e^{\beta_l}$ before normalization. $B$ is invariant to code-row permutations. The amplitudes $m_q^{(l)}, m_k^{(l)}$ are fixed rather than learned, each matching the RMS energy of the content it replaces; the calibration is given in Eq. (9) of App. D.

**Identity coordinates for reference and target tokens.** The bias of Eq. (4) is pairwise, so both sides must carry informative coordinates: if every target token kept the background coordinate, the term would be identical for every target token and carry no information about which character it belongs to. We therefore assign the coordinate at token granularity. Reference assets are subject-centered crops and are coded whole, every token of an asset carrying the coordinate of the character it depicts. Target tokens take their coordinate from a per-frame bounding box for each character, $s{=}n$ inside the box of character $n$, and $s{=}0$ elsewhere and wherever boxes overlap; text tokens carry the background coordinate.

A character's reference pool holds more assets than one step can feed, and a fixed choice would supervise reproduction of the chosen assets, so at each step we resample two reference images and two reference clips per character, together with a fresh code-row assignment; the LoRA is trained with the base flow-matching objective restricted to target tokens, with no auxiliary losses.

### 3.4 Cross-Context Identity Reward

**The comparison target.** Flow matching asks whether a frame is reproduced, not whether a face is the person it should depict, so the supervised stage permits the invariance that the corpus and the coordinates make learnable without ever requiring it; post-training makes it the optimized quantity. We treat a face embedding as the sum of a component invariant to the recording and a component specific to the viewpoint, lighting, and styling under which the face was captured. Scoring a generated face against the conditioning references assigns the same credit to reproducing that context as to reproducing the person, and the context is the cheaper of the two, because it is present in the input, so a higher score under this protocol means a better copy rather than a better likeness.

Same-clip pooling (Shen et al., 2025) cancels pose and expression but not the clip's lighting and styling, so the context component cancels only when sources are mutually disjoint. We therefore score each generated face against a pool of references of the same identity, withheld from conditioning and drawn from contexts disjoint from one another and from the target: the context components then point in inconsistent directions and average toward zero, and reproducing any conditioning reference no longer attains a high reward. The reward pool, the conditioning references, and the evaluation references are drawn from pairwise disjoint context partitions fixed before training.

**Optimization.** Because the reward is a differentiable embedding similarity, we optimize it by reward gradient (Clark et al., 2023; Prabhudesai et al., 2024). A rollout is sampled under the training conditioning distribution, the computation graph is kept for the final denoising step only (Clark et al., 2023), which is where facial detail is resolved (Shen et al., 2025), and one uniformly sampled frame is decoded differentiably (Prabhudesai et al., 2024); faces detected on that frame define differentiable aligned crops through which the gradient reaches the pixels (Shen et al., 2025). Each detected face is matched to one of the characters conditioning the rollout by the cosine between its embedding and the pool vectors, with the assignment computed by the Hungarian algorithm and held constant (Cheng et al., 2025); the rollout solver, the crop construction, and the detection safeguards are specified in App. A. The stage optimizes

$$R(x) = \frac{1}{|\mathcal{M}|} \sum_{(k,n)\in\mathcal{M}} \cos\Big(f\big(\mathrm{crop}_k(x)\big),\ p_n\Big), \qquad p_n = \tfrac{1}{M}\widehat{\textstyle\sum_m e\big(r_m^{(n)}\big)},$$
$$\max_\theta\ \mathbb{E}\big[R(x_\theta)\big]\ -\ \lambda\,\mathbb{E}\,\big\|v_\theta(z_t,t,\mathrm{cond}) - v_{\mathrm{SFT}}(z_t,t,\mathrm{cond})\big\|^2, \tag{5}$$

where $\widehat{\cdot}$ denotes $\ell_2$ normalization, $f$ and $e$ are the same face embedding applied to generated crops and to the withheld references $r_m^{(n)}$, and $\mathcal{M}$ is the matching. The anchor penalizes velocity drift from the supervised model (Shen et al., 2025) and closes the shortcut of improving the face score by suppressing motion. The reward optimizes likeness only; binding each character to its references and to its region in the frame is the work of the identity coordinate of §3.3, which the reward does not read.

# 4 Experiments

## 4.1 Training Data: the Reprise-10K Corpus

To support identity-faithful multi-reference human video generation, we construct Reprise-10K from long-form television and film, where a fixed cast recurs across many scenes under changing conditions and scenes are often populated by several characters at once. Each sample pairs one target clip with several reference images and several reference clips, every reference tied to a specific character in the target, so the corpus supervises identity preservation and multi-character binding directly. Its organizing principle is cross-context pairing: no reference comes from the target clip, which would let the model reconstruct the target by copying pose, background, and layout, and no two references of one character come from adjacent shots, whose styling and lighting are shared.

Specifically, we ingest three consecutive episodes of each series, or the whole of each film, cut them at scene boundaries into single-shot clips, and detect, embed, and associate faces into per-shot tracks. The tracks are clustered jointly across all shots of a title into character-level identities, which is the step that gives each identity several contexts, and a recurrence test keeps the main cast and discards extras and crowds. After quality gating, a character's references are drawn across clusters of its candidates in face-embedding space, so that they differ from one another in viewpoint and expression rather than repeat one look; reference clips are screened by head motion and by facial landmark change, so that those admitted carry head rotation and expression change. For multi-character targets the procedure runs independently per character.

Reprise-10K comprises 10,000 identities with three to twelve reference images and three to twelve reference clips per character slot, reference sets spanning three to twenty-four distinct shots, and targets divided into single-character and two- to three-character co-frame clips. App. B gives the full pipeline.

## 4.2 Setup

**Implementation details.** Reprise trains a LoRA at both 1.3B and 14B, and post-training starts from the supervised checkpoint and updates the same LoRA. Targets are $81$ frames at $480 \times 832$, reference images are $512 \times 512$ crops, and reference clips are $65$ frames. Ablations use the 1.3B model under identical settings unless stated otherwise, and hyperparameters are listed in App. A.

**Benchmark.** No existing benchmark supplies cross-context references of the same identity, so we construct a test set of $100$ identities collected from the web, celebrities and ordinary individuals alike.

Table 1: Comparison on the cross-context test set, reported in the metric groups of Xu et al. (2026) with one column added to the identity group. **FaceSim** scores against the conditioning references; **CCIS** against references of the same identity from other contexts, given to no model as input. Systems are ordered by overall performance across the five columns. Best per column in **bold**.

| Method | FaceSim↑ | CCIS↑ | MotionSmooth↑ | Natural↑ | X-CLIP↑ |
|---|---|---|---|---|---|
| RefAlign-1.3B | 0.508 | 0.371 | 0.946 | 3.49 | 0.311 |
| Phantom-1.3B | 0.492 | 0.361 | 0.925 | 3.67 | 0.329 |
| VACE-1.3B | 0.483 | 0.372 | 0.959 | 3.58 | 0.304 |
| MAGREF | 0.507 | 0.383 | 0.932 | 3.61 | 0.327 |
| Reprise-1.3B (Ours) | 0.514 | 0.395 | 0.949 | 3.65 | 0.328 |
| VACE-14B | 0.526 | 0.409 | 0.941 | 3.69 | 0.315 |
| VINO | 0.511 | 0.396 | 0.939 | 3.72 | 0.331 |
| RefAlign-14B | 0.552 | 0.391 | 0.950 | 3.55 | 0.318 |
| BindWeave | 0.517 | 0.412 | 0.934 | 3.63 | 0.338 |
| Phantom-14B | 0.499 | 0.404 | 0.948 | 3.79 | 0.326 |
| SkyReels-V3 | 0.504 | 0.390 | 0.956 | 3.81 | 0.333 |
| Reprise-14B (Ours) | **0.589** | **0.508** | **0.964** | **3.92** | **0.347** |

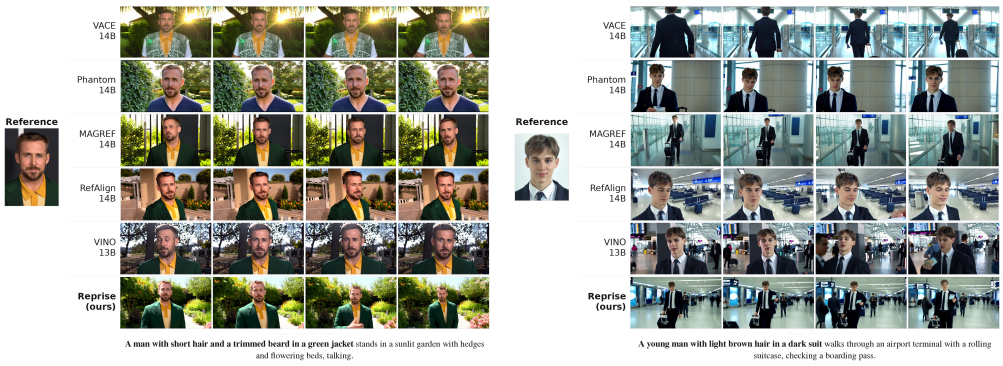
It consists of $150$ cases: $100$ single-subject cases and $50$ multi-subject cases formed by combining these identities, $30$ with two characters and $20$ with three. Each character is given two reference images and two reference clips, each from a different context; each case is paired with one natural-language prompt and generated under three random seeds.

**Metrics and baselines.** For identity preservation we report two columns that share one estimator and differ only in the references they score against. FaceSim (Yuan et al., 2024; Kim et al., 2022) compares each generated face with the references supplied to the model, and CCIS compares it with references of the same identity drawn from other contexts, which are given to no model as input. In both columns the references of an identity are first pooled into a single $\ell_2$-normalized identity vector, exactly as in Eq. (5); every detected face in the generated clip is then scored against that vector by cosine similarity in the face-embedding space, and the column averages those scores over the frames of the clip. A model that reproduces the conditioning pixels raises FaceSim without raising CCIS; the two metrics measure identity-reference consistency from different angles. In multi-subject cases each detected face is scored against the references of the character whose conditioning region it occupies. We further report Motion Smoothness from VBench (Huang et al., 2024; Li et al., 2023), NaturalScore from OpenS2V-Eval (Yuan et al., 2025) for perceptual naturalness on a five-point scale, and X-CLIP (Ma et al., 2022) for prompt following.

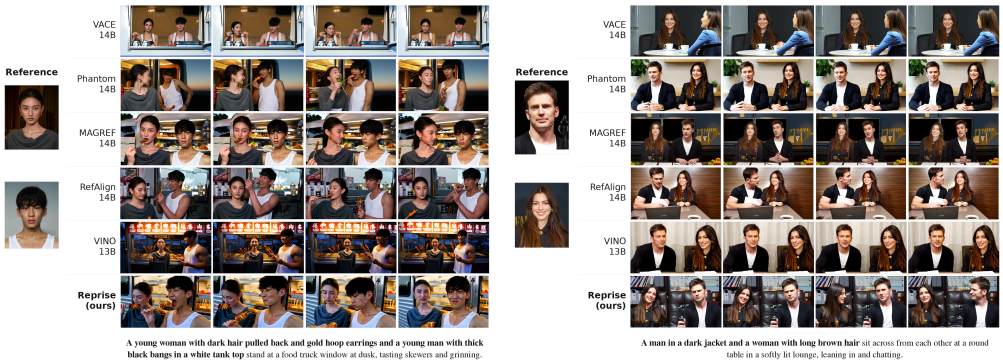
We compare Reprise with strong open-source subject-to-video baselines: VACE (Jiang et al., 2025), Phantom (Liu et al., 2025b), RefAlign (Wang et al., 2026b), MAGREF (Deng et al., 2025), BindWeave (Li et al., 2025), VINO (Chen et al., 2026), and SkyReels-V3 (SkyReels Team, Skywork AI, 2026). For VACE, Phantom, and RefAlign we report both the 1.3B and the 14B variant, to assess performance across model scales. VACE unifies reference-conditioned video creation, Phantom aligns image and video with text, RefAlign aligns reference and generation representations, MAGREF disentangles subjects by masked guidance, BindWeave adds a multimodal understanding model, and VINO conditions on interleaved omni-modal context.

### 4.3 Quantitative Comparison

Table 1 compares Reprise with subject-to-video baselines along identity, motion, naturalness, and prompt following. RefAlign-14B is the strongest baseline on FaceSim at $0.552$, ahead of the next system by $0.026$, which is the quantity its alignment objective contains, and on references it never saw it falls to $0.391$, below four baselines it outranked on FaceSim; the same reversal holds for RefAlign-1.3B and, inside our own model, for the reward-target ablation of Table 3. Phantom-14B and BindWeave, which do not optimize reference similarity directly, sit lower on FaceSim and higher on CCIS. Reprise-14B is first on both columns, $0.589$ and $0.508$, and it is the only system for which the two columns agree, which is what the cross-context measure of §4.2 is designed to isolate. On the multi-subject subset the separation is larger, Reprise-14B at $0.487$ against $0.372$ for the strongest




(a) Single-subject comparison.




(b) Multi-subject comparison.

Figure 3: **Qualitative comparison.** Each panel shows a conditioning reference (left), five open-source baselines, and Reprise (bottom row), with four evenly-spaced frames per method. **(a)** Single-subject. **(b)** Multi-subject, with two characters of the same gender and similar age.

baseline, which is the multi-subject binding claim of §3.3 appearing inside the identity columns; at 1.3B the same subset is level with that baseline at 0.370 (§4.5).

The remaining columns say what the identity gains cost. Reprise-14B is first on smoothness at 0.964, on naturalness at 3.92, and on prompt following at 0.347, so the identity gains are not paid for in temporal coherence or text alignment; naturalness separates copying from generation (Yuan et al., 2025), and the two RefAlign scales are lowest there at 3.49 and 3.55, including the baseline that leads FaceSim. Finally, Reprise-1.3B improves on RefAlign-1.3B, the 1.3B baseline that leads FaceSim, in both identity columns, 0.514 against 0.508 and 0.395 against 0.371, with the held-out margin four times the conditioning one; its CCIS sits inside the band of the 14B baselines, 0.383 to 0.412, so at one tenth the parameters the held-out identity reaches the level of the larger systems, while its FaceSim stays below the baselines that lead that column.

### 4.4 Qualitative Comparison

In the single-subject cases of Figure 3(a), the baselines carry the conditioning reference's lighting and crop into every prompted scene; under large head rotation the distinctive facial geometry fades toward a generic mean. In the multi-subject cases (b), characters swap positions or converge on a blend of both references, most often when they are of the same gender and similar age, where appearance similarity alone cannot decide which reference belongs to whom. Reprise holds facial structure through rotation, derives its lighting from the prompted context rather than the reference, and keeps the character–reference binding stable across the clip.

Table 2: **Identity coordinate** (1.3B). *w/o ID-RoPE* returns the four planes to $t$ and drops $s$; *w/o token-level $s$* codes only the references, giving every target token the background coordinate; *w/o carrier* keeps per-token $s$ but rotates its phase into unmodified content. Smooth is MotionSmooth; best CCIS in **bold**; the two right columns are controls.

| Variant | FaceSim↑ | CCIS↑ | Smooth↑ | X-CLIP↑ |
|---|---|---|---|---|
| w/o ID-RoPE | 0.469 | 0.346 | 0.950 | 0.332 |
| w/o token-level $s$ | 0.492 | 0.366 | 0.949 | 0.335 |
| w/o carrier | 0.497 | 0.375 | 0.950 | 0.330 |
| Full | 0.514 | **0.395** | 0.949 | 0.328 |

Table 3: **Reward target** (1.3B). All rows start from the shared supervised checkpoint (*SFT*) under one post-training budget and differ only in the reward, that is, in what a generated face is scored against; *Full* is the same run as *Full* in Table 2, whose caption also explains the columns.

| Variant | FaceSim↑ | CCIS↑ | Smooth↑ | X-CLIP↑ |
|---|---|---|---|---|
| SFT | 0.502 | 0.363 | 0.948 | 0.325 |
| + reference reward | 0.527 | 0.355 | 0.938 | 0.308 |
| + same-clip pool reward | 0.521 | 0.374 | 0.943 | 0.321 |
| Full (cross-context reward) | 0.514 | **0.395** | 0.949 | 0.328 |

### 4.5 Ablation Study

**Reference conditioning.** Adding references from fresh contexts lowers FaceSim monotonically while raising CCIS, and a token-matched control that replaces each reference clip by its own first frame repeated returns under a third of the CCIS the clips contribute; App. E reports the four conditioning settings, this control, and the default of two images and two clips per character.

**The identity coordinate.** In Table 2, removing the axis lowers CCIS from $0.395$ to $0.346$ and FaceSim from $0.514$ to $0.469$, and the loss is confined to the multi-subject subset, where CCIS falls from $0.370$ to $0.224$ while the single-subject subset holds at $0.407$, since with one character there is no competing reference set for a target region to be bound to. Coding only the reference side recovers the multi-subject subset to $0.284$: the reference pool stays partitioned, but every target token receives an identical row of biases and none is pointed at any character, which is what a coordinate assigned per shot would give, the target being a single shot. Keeping the per-token assignment and rotating the identity phase into unmodified content recovers further, to $0.310$, the bias then being content-dependent so that cross-identity pairs no longer cancel; App. D derives both. Both control columns stay flat across the table, motion smoothness within $0.001$ and X-CLIP within $0.007$, so every change the coordinate produces is an attribution change.

**The reward target.** In Table 3, rewarding similarity to the conditioning references raises FaceSim from $0.502$ to $0.527$ while CCIS falls below the supervised floor to $0.355$ and X-CLIP drops to $0.308$, so the optimization pressure transfers the conditioning scene into the output in place of the prompted one; motion smoothness falls to $0.938$, the lowest in the table, because the reward observes one frame at a time and contains no term coupling consecutive frames, so the transferred appearance is imposed inconsistently along the clip. A pool drawn from the same clip recovers part of the gap, $0.374$ CCIS, since pose and expression cancel under the pool but the clip's lighting and styling do not, as §3.4 predicts. The cross-context pool reaches $0.395$, $+0.032$ over the floor, with FaceSim held at $0.514$ and both controls at or above it; it is the only reward variant whose held-out gain exceeds its conditioning gain, and two thirds of that gain ($+0.021$) come from the disjointness of the pool rather than from pooling per se, since the same-clip pool realizes only $+0.011$. Removing the velocity anchor gains $+0.003$ CCIS for $-0.011$ motion smoothness, so the anchor holds the post-trained velocity field to the supervised one at a negligible identity cost.

## 5 Conclusion

We presented Reprise, a framework for generating people who resemble themselves rather than the photographs used to specify them. It rests on a definition of identity as what remains invariant across a person's appearances in different contexts, and that single definition fixes all three parts of the system: Reprise-10K supplies references from contexts disjoint from the target, ID-RoPE binds each character jointly to all of its references, and the cross-context reward makes the invariant component an optimized quantity rather than a byproduct of reconstruction. The same definition determines the measurement, and reporting reference similarity next to cross-context identity is what makes the difference between copying and generating visible at all.

## AI USE STATEMENT

Generative AI tools were used for language editing of this manuscript. All content is the responsibility of the authors.

## REFERENCES


Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Junyi Chen, Tong He, Zhoujie Fu, et al. Vino: A unified visual generator with interleaved omnimodal context. *arXiv preprint arXiv:2601.02358*, 2026.

Zhuowei Chen, Bingchuan Li, Tianxiang Ma, et al. Phantom-data: Towards a general subject-consistent video generation dataset. *arXiv preprint arXiv:2506.18851*, 2025.

Yufeng Cheng, Wenxu Wu, Shaojin Wu, et al. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Yufan Deng, Yuanyang Yin, Xun Guo, et al. Magref: Masked guidance for any-reference video generation with subject disentanglement. *arXiv preprint arXiv:2505.23742*, 2025.

Sharath Girish, Viacheslav Ivanov, Tsai-Shien Chen, et al. Alchemint: Fine-grained temporal control for multi-reference consistent video generation. *arXiv preprint arXiv:2512.10943*, 2025.

Jianzhu Guo et al. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

Zujin Guo et al. Generate your talking avatar from video reference. *arXiv preprint arXiv:2604.27918*, 2026.

Li Hu et al. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. arXiv:2311.17117.

Ziqi Huang, Yinan He, Jiashuo Yu, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. arXiv:2311.17982.

Zitong Huang, Kaidong Zhang, Yukang Ding, et al. Mind the generative details: Direct localized detail preference optimization for video diffusion models. *arXiv preprint arXiv:2601.04068*, 2026.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, et al. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. *arXiv preprint arXiv:2204.00964*, 2022.

Yixuan Lai, He Wang, Kun Zhou, et al. Slot-id: Identity-preserving video generation from reference videos via slot-based temporal identity encoding. *arXiv preprint arXiv:2601.01352*, 2026.

Zhaoyang Li, Dongjun Qian, Kai Su, et al. Bindweave: Subject-consistent video generation via cross-modal integration. *arXiv preprint arXiv:2510.00438*, 2025.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *CVPR*, 2023.

Benjamin Liang, Ce Chen, Desmond Lin, et al. Avatar v: Scaling video-reference avatar video generation. *arXiv preprint arXiv:2606.13872*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, et al. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Lijie Liu, Tianxiang Ma, Bingchuan Li, et al. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025b.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, et al. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022.

Xiangyu Meng, Zixian Zhang, Zhenghao Zhang, et al. Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning. *arXiv preprint arXiv:2510.14256*, 2025.

Zijie Meng, Jiwen Liu, Yufei Liu, et al. Argus: Stacked multi-view identity mosaic injection for subject-preserving video generation. *arXiv preprint arXiv:2606.11670*, 2026.

Panwang Pan, Jingjing Zhao, Yuchen Lin, et al. Id-crafter: Vlm-grounded online rl for compositional multi-subject video generation. *arXiv preprint arXiv:2511.00511*, 2025.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, et al. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Liao Shen, Wentao Jiang, Yiran Zhu, et al. Identity-preserving image-to-video generation via reward-guided optimization. *arXiv preprint arXiv:2510.14255*, 2025.

SkyReels Team, Skywork AI. Skyreels-v3 technique report. *arXiv preprint arXiv:2601.17323*, 2026.

Bram Wallace, Meihua Dang, Rafael Rafailov, et al. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

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

Jiahao Wang, Hualian Sheng, Sijia Cai, et al. Anyid: Ultra-fidelity universal identity-preserving video generation from any visual references. *arXiv preprint arXiv:2603.25188*, 2026a.

Lei Wang, YuXin Song, Ge Wu, et al. Refalign: Representation alignment for reference-to-video generation. *arXiv preprint arXiv:2603.25743*, 2026b.

Jiazheng Xing, Fei Du, Hangjie Yuan, et al. Lumosx: Relate any identities with their attributes for personalized video generation. *arXiv preprint arXiv:2603.20192*, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, et al. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

Yulong Xu, Xinyue Liu, Shujuan Li, et al. Vera: Identity-faithful human subject-to-video generation. *arXiv preprint arXiv:2607.20247*, 2026.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, et al. Stand-in: A lightweight and plug-and-play identity control for video generation. *arXiv preprint arXiv:2508.07901*, 2025a.

Zeyue Xue, Jie Wu, Yu Gao, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025b.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, et al. Instructvideo: Instructing video diffusion models with human feedback. *arXiv preprint arXiv:2312.12490*, 2023.

Shenghai Yuan, Jinfa Huang, Xianyi He, et al. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024.

Shenghai Yuan, Xianyi He, Yufan Deng, et al. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Pengze Zhang, Yanze Wu, Mengtian Li, et al. Omnitransfer: All-in-one framework for spatio-temporal video transfer. *arXiv preprint arXiv:2601.14250*, 2026.

Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, et al. Kaleido: Open-sourced multi-subject reference video generation model. *arXiv preprint arXiv:2510.18573*, 2025.

Yong Zhong, Zhuoyi Yang, Jiayan Teng, et al. Concat-id: Towards universal identity-preserving video synthesis. *arXiv preprint arXiv:2503.14151*, 2025.

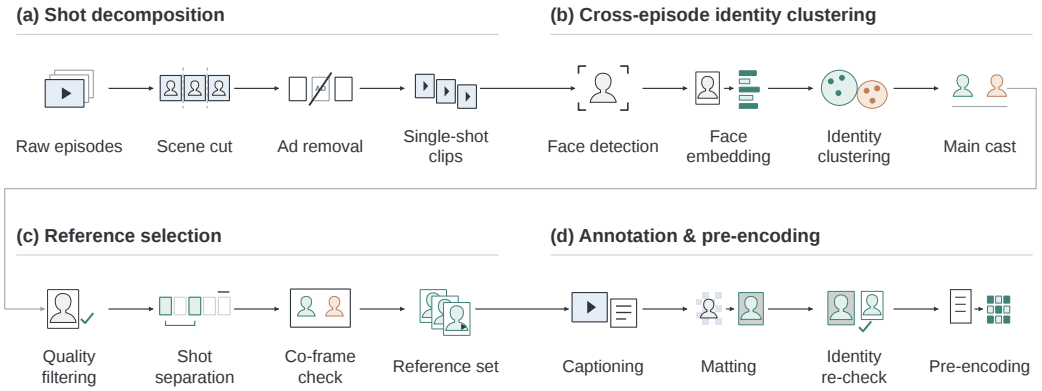



Figure 4: **The Reprise-10K construction pipeline.** Stage 1 decomposes raw episodes into single-shot clips by scene cutting and advertisement removal. Stage 2 detects and embeds faces in every clip, clusters the resulting tracks into identities jointly over all shots of all three episodes, and retains the main cast. Stage 3 selects each character's reference set under quality filtering, the shot-separation constraint, and the co-frame check. Stage 4 captions the material, mattes the references, re-verifies identity on the delivered crop, and pre-encodes every asset.

## A Implementation details

The backbone is the frozen Wan2.1 1.3B DiT. A single LoRA of rank 128 and scaling $\alpha$=128 is inserted into every linear projection (queries, keys, values, and output) of the self-attention in all 30 blocks; no other parameters are trained. Both the supervised stage and the reward stage share the same adapter.

**Supervised stage.** Training runs at 480p (480×832, 81 frames) with a global batch size of 16 and gradient accumulation over 2 micro-batches. The optimizer is AdamW ($\beta_1$=0.9, $\beta_2$=0.999, weight decay 0.01) at learning rate $5\times10^{-5}$ with linear warmup over the first 500 steps and cosine decay to $5\times10^{-6}$. At each step, two reference images and two reference clips per character are resampled together with a fresh code-row assignment, and the loss is the flow-matching objective restricted to target tokens. Training runs for 10,000 steps in fp16 on 24 NVIDIA H800 GPUs.

**Reward stage.** Rollouts are sampled at the same 480p format with a 20-step Euler solver, shorter than the 50-step inference schedule, and no classifier-free guidance. The computation graph is retained for the final denoising step only; one frame is uniformly sampled and decoded differentiably through the VAE decoder. The face detector (RetinaFace, confidence threshold 0.5) runs outside the computation graph; detected keypoints, treated as constants, define differentiable aligned crops of 112×112 for the face embedding (AdaFace-IR101). Pool vectors are precomputed once before training and frozen. The optimizer is AdamW at learning rate $10^{-5}$, gradient clipping 1.0, anchor weight $\lambda$=1, and the stage runs for 200 updates at global batch size 8. Detection safeguards: unmatched faces contribute nothing; a character whose face is not detected is dropped from the mean rather than scored zero, since a zero would inject a gradient toward erasing the face; and a rollout with no detected face is discarded.

## B Corpus construction

Figure 4 summarizes the pipeline; the paragraphs below specify each stage.

**Why television and film footage.** The definition of identity in §1 asks for training data with three properties that existing subject-to-video corpora do not provide together. The same person must recur under genuinely different viewpoints, lighting, styling, and expression, since otherwise the task degenerates into reproducing the reference and no metric can separate the two. Several identities must appear in the same frame, since otherwise the identity coordinate of §3.3 has nothing to resolve

and within-frame attribution has no training signal at all. And the references of one person must differ from each other, since otherwise a second reference carries no information the first did not. Long-form television and film satisfy the first two by construction, because a fixed cast recurs across many scenes under changing production conditions and those scenes are populated, and the third is obtained by active selection. We therefore build Reprise-10K from television series and feature films.

**Identity extraction.** We ingest three consecutive episodes of each series, or the whole of each film, above a resolution gate and cut them at scene boundaries into single-shot clips. Faces are detected in every clip, embedded, and associated into per-shot tracks. Tracks are then clustered into identities jointly over all shots of all three episodes rather than within any single clip, and it is this cross-episode step that creates multiple contexts per identity in the first place. A recurrence test retains the main cast, whose members reappear across shots with sufficient frontal and large-face evidence, and discards extras, posters, and background crowds. For each retained identity we export reference images and reference clips from its tracks, target clips taken from shots disjoint from every reference, and per-frame boxes for every character in the target, which are the source of the target-side identity coordinate.

**Reference selection.** Quality gates cover face size, sharpness, occlusion, and eye visibility, and are measured on the delivered crop rather than on the source frame. Two further constraints act on the reference set rather than on the individual reference. Yaw is bucketed and sampled round robin, which is a distribution constraint and not a gate, since profile views are not discarded but are precisely the shape evidence that a frontal view cannot supply. Within a bucket, references are chosen by greedy farthest-point selection in the embedding space, each new reference minimizing its largest cosine to those already chosen and required to fall below a threshold; two references of one character can otherwise reach a cosine of $0.965$, at which point the second is close to redundant. Reference clips are screened for dynamics by head-motion range, by the variation of inner-face landmarks after the rigid head-pose component has been removed, and by the rate of change of both, with upper bounds that exclude motion blur; the three gates are set as quantiles of the corpus distribution rather than as absolute values. Clips that pass carry expression change and head rotation that a still image cannot supply.

**Context separation and background removal.** A structural gate enforces the property the corpus exists for. Every reference comes from a shot disjoint from the target's, and no two references of one character come from shots less than two apart within an episode, because shot-reverse-shot neighbours share styling and lighting and would reintroduce the confound under the appearance of separation. A subject is admitted only if at least three reference images and three reference clips survive every gate, and a group only if every subject in it qualifies. Admitted references are matted and composited on constant grey. Matting serves two purposes: references from one scene share a background, which lets a model infer which reference belongs to which character without using the identity coordinate at all, and that background is itself carried into the target. Grey rather than white is used because a white field correlates with studio and overexposed footage and introduces a domain prior into the generation.

**Scale and scope.** Reprise-10K comprises 10,000 identities. Every character slot carries at least three and at most twelve reference images and likewise at least three and at most twelve reference clips, its reference set spans between three and twenty-four distinct shots, and the large majority of slots draw references from more than one episode. Targets divide into single-character clips and co-frame clips of two or three characters. Three properties bound what the corpus can support. Context is not an annotated field; it is operationalized as shot disjointness together with cross-episode spread, so the claim that references differ in styling and lighting follows from how they were selected and is not separately measured. The reference side is a face crop and not a full body, so the corpus supervises facial identity and facial dynamics, not gait or physique. The cap of three co-framed characters is a property of the export configuration and not of the method, and the three-character subset is small enough that we report it separately rather than pooling it.

## C PRELIMINARIES

The backbone is the Wan2.1 1.3B text-to-video DiT (Wan Team, 2025), a flow-matching model with a causal VAE that compresses video $8\times$ spatially and $4\times$ temporally. An 81-frame $480\times832$ target is reduced by the VAE to $21\times60\times104$ latent cells, the temporal extent being $1+(81-1)/4=21$, and the DiT patch embedding of size $(1,2,2)$ then merges each $2\times2$ spatial block, giving latent tokens of shape $(T,H_l,W_l)=(21,30,52)$. Each of the 30 DiT blocks applies full softmax self-attention with head dimension $d_h$=128, factorized into 64 complex planes by rotary position embedding (RoPE). The 64 planes are split $(22,21,21)$ over the temporal, height, and width axes. Within each axis group, plane $p$ is assigned angular frequency

$$\omega_p=\theta^{-2(p-1)/D},\qquad \theta=10{,}000, \tag{6}$$

where $D$ is the number of real dimensions in the group ($D$=44 for $t$, 42 for $h$ and $w$), so that the group holds $D/2$ complex planes. Planes are ordered so that frequency decreases with index within each group: plane 1 of the temporal group has the highest temporal frequency, $\omega_1$=1, and plane 22 has the lowest. ID-RoPE reassigns the $P$=4 slowest temporal planes (indices 19–22 of the temporal group) to the identity axis, turning the split into $(18,21,21,4)$. These planes have $\omega_p\le\omega_{19}=\theta^{-36/44}\approx5.34\times10^{-4}$, so across the 20 frame intervals of the 21-frame target the accumulated phase is at most $20\omega_{19}\approx0.011$ rad; the temporal resolution they carry is negligible, and their reassignment does not observably affect temporal coherence.

## D THE IDENTITY COORDINATE IN FULL

**Code construction.** Let $H$ be a $2P\times2P$ real Hadamard matrix ($2P$=8) with $HH^\top=2P\,I$ and entries $\pm1$. Row $s$ is folded into $P$=4 complex code planes by pairing adjacent columns:

$$u_p(s)=\tfrac{1}{\sqrt{2}}\big(H_{s,2p-1}+i\,H_{s,2p}\big),\qquad p=1,\dots,P. \tag{7}$$

Each component has unit modulus, $|u_p(s)|=1$, satisfying the carrier form of Eq. (3), and the orthogonality of $H$ gives

$$\mathrm{Re}\langle u(s),u(s')\rangle=\sum_{p=1}^{P}\mathrm{Re}\big[\overline{u_p(s)}\,u_p(s')\big]=\tfrac{1}{2}\sum_{q=1}^{2P}H_{s,q}\,H_{s',q}=P\cdot\mathbf{1}[s{=}s'], \tag{8}$$

which is the identity used to obtain Eq. (4). The all-ones row of $H$ is assigned to $s$=0 (background); a global phase rotation $u(s)\mapsto e^{-i\pi/4}u(s)$ makes $u_p(0)=1$ for all $p$ without changing $\mathrm{Re}\langle\cdot,\cdot\rangle$. The remaining $2P-1=7$ rows are available for characters, so the capacity is $K\le2P=8$ groups (one background plus seven characters). Beyond this bound the code vectors are no longer orthogonal and the indicator in Eq. (4) breaks: cross-identity bias is no longer exactly zero.

**Carrier amplitude calibration.** The carrier amplitude is set by energy conservation:

$$m_q^{(l)}=\Big(\tfrac{1}{P}\textstyle\sum_{p\in\mathcal{C}}\mathbb{E}\,\big|q_p^{(l)}\big|^2\Big)^{1/2}, \tag{9}$$

and analogously for $m_k^{(l)}$, each matching the mean per-plane energy of the content it replaces, measured once on the frozen backbone over a small batch and then fixed (60 constants over 30 blocks); the axis adds no learnable parameters.

**The two degraded coordinates of Table 2.** Coding the reference side alone leaves a character's references attending to each other at $+\beta_l$ and to other characters at exactly zero, so the reference pool stays partitioned, but every target token carries $s$=0 and Eq. (4) then returns zero on all reference keys and $\beta_l$ on all target keys, identically for every target token, so no target token is pointed at any character. A coordinate assigned per shot reduces to this case: the target is a single shot, so a per-shot coordinate is constant over the frame and $\Delta s$ vanishes for every in-frame pair, which leaves the assignment inside a frame undetermined however the reference side is organized. Rotating the identity phase into unmodified content instead of replacing it gives a bias of $\mathrm{Re}[q_p^{(l)}\overline{k_p^{(l)}}e^{i(\phi_p(s_x)-\phi_p(s_y))}]$ rather than the constant of Eq. (4): same-identity pairs lose the uniform lift that raises a reference set as a whole, and cross-identity pairs scale with content instead of cancelling, so two characters of similar appearance retain a path between them.

## E REFERENCE-CONDITIONING ABLATION

Table 4: Reference conditioning (1.3B). All rows draw references from mutually disjoint contexts, disjoint from the target; the last row is the default. Row 4 matches row 5 in assets and tokens but repeats each clip's first frame, isolating motion from token budget. Tokens as in §3.2. Best per identity column in **bold**; MotionSmooth is a control.

| References per identity | Assets | Ref. tokens | FaceSim↑ | CCIS↑ | MotionSmooth↑ |
|---|---|---|---|---|---|
| 1 image | 1 | 1,024 | **0.561** | 0.322 | 0.951 |
| 2 images | 2 | 2,048 | 0.538 | 0.358 | 0.950 |
| 2 images + 1 clip | 3 | 19,456 | 0.521 | 0.381 | 0.951 |
| 2 images + 2 static-repeat clips | 4 | 36,864 | 0.527 | 0.369 | 0.957 |
| 2 images + 2 clips | 4 | 36,864 | 0.514 | **0.395** | 0.949 |

Across rows 1–3 and 5 of Table 4, FaceSim falls monotonically from $0.561$ to $0.514$ while CCIS rises from $0.322$ to $0.395$: a single reference is the easiest thing to reproduce and the weakest evidence of an invariant, and each added asset from a fresh context is a weaker candidate for copying and a stronger estimate of the person. Row 4, the token-matched static-repeat control, returns CCIS to $0.369$, close to the two-image level, so the gain of row 5 is attributable to the motion its clips carry rather than to the extra tokens.