STRIDE: SCENE TRANSITION VIA INTERPOLATIVE DIFFUSION ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video transition generation, the synthesis of coherent intermediate frames bridging a given start and end frame, is a core problem in generative video modeling. Existing diffusion and flow methods are not only computationally heavy but also rely on large pretrained image/video backbones. We present STRIDE (**S**cene **Tr**ansition via **I**nterpolative **D**iffusion **E**stimation), an inference-efficient bidirectional latent-diffusion framework whose 3D U-Net denoiser ($\sim$1.54B parameters) is customized and trained as *generative-prior-free*, that carries no transition or motion prior from any pretrained generator (a frozen VAE and CLIP serve only as fixed interfaces). STRIDE couples position-aware progressive masking, coupled bidirectional temporal attention with a forward/backward consistency loss, and multi-anchor CLIP conditioning, is trained under a four-phase motion-filtered curriculum with an annealed-fusion inference schedule. Operating in a compact 256^2 pixel and 32^2 latent space, it infers $[3.7$ to $43]\times$ faster than DynamiCrafter, SEINE, TVG, and Wan2.1-VACE-1.3B, at $\approx$ $[36$ to $181]\times$ lower per-clip compute than DynamiCrafter, SEINE, and TVG. Further, it attains the best inter-frame structural consistency (VSI, FSIM, and tMS-SSIM on all three benchmark datasets i.e., Morph-Bench, TC-Bench, and TransitBench and tPSNR on one of the three). Although SVD-KFI (Generative Inbetweening) wins the four perceptual/appearance-similarity metrics (CLIPSim, DreamSIM, LPIPS, EMD) on all three benchmarks, its inference is $\approx 49\times$ slower than STRIDE, with Wan and Framer occupying intermediate positions. Because all eight metrics are computed between consecutive frames, they measure inter-frame smoothness rather than motion fidelity. This makes the resulting frontier an operating-point trade-off between inter-frame structural consistency and perceptual/appearance similarity, rather than a quality ranking. In our human study (5 models $\times$ 48 clips $\times$ 3 raters $=$ 720 ratings), we observe high inter-rater agreement (Krippendorff's $\alpha = 0.84$–0.87). The Friedman omnibus test reveals significant differences in transition-smoothness ratings across models ($p < 10^{-14}$). Furthermore, Holm-corrected pairwise Wilcoxon tests confirm that STRIDE is significantly preferred over all four pretrained-prior baselines (DynamiCrafter, SEINE, TVG, Wan2.1-VACE-1.3B). Despite using no pretrained diffusion backbone, STRIDE achieves significantly smoother transitions than pretrained-prior baselines, at the cost of occasional reductions in intermediate-frame visual quality near the transition midpoint.

1 INTRODUCTION

Generating smooth and coherent transitions between two frames is a core capability for video editing, visual effects, and generative media, and its demand has grown as diffusion and flow-based video synthesis proliferate. Recent conditional generators (DynamiCrafter (Xing et al., 2024), SEINE (Chen et al., 2023), TVG (Zhang et al., 2025), MAVIN (Zhang et al., 2024a), VTG (Yang et al., 2025b)) are built on large pretrained image/video backbones and, on large content gaps, produce abrupt or crossfaded transitions (see figure 4). We hypothesize two contributing tendencies: (i) treating the endpoints asymmetrically by conditioning on one frame while using the second as a prior rather than enforcing both boundary constraints throughout denoising, and (ii) using static conditioning that does not adapt to the transition's semantic complexity. In our controlled study, these

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.

surface on two axes: i) temporal smoothness (gradual, continuous evolution of content start→end) and ii) commitment to a single coherent intermediate (see section 4.5).

We propose **STRIDE** (*Scene Transition via Interpolative Diffusion Estimation*), a bidirectional latent-diffusion framework for transition generation whose denoiser is trained *generative-prior-free*[1]. We make the following contributions:

1. **A generative-prior-free transition formulation.** A randomly-initialized bidirectional denoiser with an interpolative fused-noise forward process and a $\sin^2$ position-aware progressive masking curriculum that grades the task from easy near-boundary reconstruction to hard center interpolation (see sections 3.2, 3.3).
2. **Coupled bidirectional temporal attention.** A single shared-weight denoiser runs under causal and anti-causal masks each step and the two directional noise estimates are fused by position under a trainable consistency loss (see section 3.5).
3. **Multi-anchor CLIP conditioning.** Boundary-weighted, timestep-aware CLIP conditioning on visible anchors, trained under a four-phase motion-filtered curriculum (see sections 3.4, 3.7).
4. **A controlled characterization of the smoothness–fidelity tradeoff.** We show that a generative-prior-free denoiser reaches strong temporal smoothness but incurs a measurable mid-sequence fidelity/commitment penalty (the U-curve), quantified by automatic metrics, a five-method human study, ground-truth fidelity, and negative-result ablations (see sections 4.5, 5).

2 Related Work

2.1 Existing Video Transition Generation Models

Pretrained-prior transition generators. Most recent transition methods adapt a large pretrained video-diffusion backbone. **DynamiCrafter** (Xing et al., 2024) conditions on both endpoints via cross-attention but produces abrupt, identity-losing transitions across large gaps. **MAVIN** (Zhang et al., 2024a) infills action segments via consecutive noising and a Gaussian Filter Mixer (GFM), which we adopt at inference (see section 3.8). **SEINE** (Chen et al., 2023)'s stochastic random-mask diffusion can break object preservation when endpoints share a category. **TVG** (Zhang et al., 2025) is training-free but is upper-bounded by its base model. **VTG** (Yang et al., 2025b) is a unified image-to-video framework with no public source code, limiting direct comparison. **Wan2.1-VACE-1.3B** (Wan Team, 2025), an open-weight 1.3B-parameter video-conditional adapter, is the current parameter-matched pretrained-prior frontier and our fifth automatic-metric baseline (section 4.3).

Video frame interpolation and bounded generation. Classical VFI (Super-SloMo (Jiang et al., 2018), RIFE (Huang et al., 2022), FILM (Reda et al., 2022), AMT (Li et al., 2023), EMA-VFI (Zhang et al., 2023)) and diffusion VFI (LDMVFI (Danier et al., 2024), VIDIM (Jain et al., 2024)) excel at small inter-frame motion under short-range interpolation. Bounded generators—Time Reversal Fusion (Feng et al., 2024a), Generative Inbetweening (Wang et al., 2025b), Framer (Wang et al., 2025a) are our closest competition. However, each depends on a frozen image-to-video prior (Generative Inbetweening and Framer adapt Stable Video Diffusion (Blattmann et al., 2023)), inheriting its inference cost. Time Reversal Fusion's bidirectional fusion is a test-time heuristic over the frozen prior, not a learned mechanism, and Framer forgoes bidirectionality altogether. ViBiDSampler (Yang et al., 2025a) similarly exploits bidirectional diffusion sampling but modifies the sampling trajectory of a pretrained SVD backbone rather than training a randomly-initialized denoiser. Among these, STRIDE is the only method that is generative-prior-free *and* jointly trains a randomly-initialized bidirectional backbone with its directional prediction and fusion (table 4).

[1]The 3D U-Net denoiser ($\sim$1.54B parameters) is trained from randomly initialized weights. A frozen VAE (`stabilityai/sd-vae-ft-mse`) and frozen CLIP ViT-L/14 serve solely as fixed interfaces and carry no transition or motion prior. The frozen CLIP encoder does supply a semantic manifold from which interior guidance is derived via linear embedding interpolation (see section 3.4). We use *generative-prior-free* to mean the absence of any pretrained *generative* image/video backbone while the frozen VAE and CLIP serve only as representation interfaces. We run no architectural ablations and so do not claim to isolate individual components (see section 6).

3 METHOD

3.1 TRAINING-DATA CURATION

Motion-centric transitions are best learned from clips with genuine motion. Randomly sampled segments are often near-static, where the optimal interpolation collapses to a trivial crossfade between the two anchors rather than true intermediate motion. To avoid training the model on this degenerate solution, every natural-video clip passes a two-stage, `ffmpeg`-free, motion-aware pipeline (see figure 3), applied identically to videos and frame sequences except the Augmenter dataset (Das et al., 2026) (our quality anchor—a curated transition dataset with four-annotator ratings as we train on its subset with mean rating $>3.5/5$, and manually curate benchmark test splits to share no anchor pair with training). The driving statistic is a resolution-invariant *motion metric* i.e., mean *residual* Farnebäck flow (Farnebäck, 2003) after RANSAC partial-affine camera-motion removal, normalized by square root of the product of its height and width ($\sqrt{HW}$). Stage 1 keeps the top 25% of clips by 90th-percentile score while stage 2 extracts (in $O(n)$ via a prefix-sum sliding window) each retained clip's contiguous M-frame ($M \geq 65$) window of maximal cumulative motion, feeding the bucketed ($\{8, 16, 24, 32\}$) curriculum of section 3.7. **Vimeo-90K exception.** Vimeo-90K septuplets are fixed-length 7-frame sequences so they cannot yield a $M \geq 65$ window and therefore bypass Stage 2. Each septuplet is used whole in the $N{=}8$ training bucket. We replicate frame `im7` to form a synthetic eighth frame `im8`, and treat `im1`/`im8` as the boundary anchors and frames `im2`–`im6` as the interior targets.

3.2 PROBLEM FORMULATION

Let $\mathbf{F} = \{\mathbf{f}_i\}_{i=0}^{N-1} \subset \mathbb{R}^{3\times H\times W}$ be an N-frame sequence and $\mathbf{m} \in \{0,1\}^N$ a mask partitioning frames into visible *anchors* ($m_i = 0$; $m_0 = m_{N-1} = 0$ by construction) and *generated* frames ($m_i = 1$). The model learns $p_\theta(\{\mathbf{f}_i : m_i = 1\} \mid \{\mathbf{f}_i : m_i = 0\})$. Following LDM (Rombach et al., 2022), we work in the latent space of a frozen KL-VAE (`sd-vae-ft-mse`), $\mathbf{z}_i = s\,\mathcal{E}(\mathbf{f}_i)$ ($s = 0.18215$), giving $\mathbf{z} \in \mathbb{R}^{B\times 4\times N\times 32\times 32}$. We use a cosine $\bar{\alpha}_t$ schedule rescaled to *zero terminal SNR* (Lin et al., 2024) ($\bar{\alpha}_T = 0$), aligning train/sampling marginals at $t = T$ and removing terminal-step brightness drift. STRIDE (see figure 1) generates all masked intermediate frames from their two boundary anchors (and optional interior anchors). At each denoising step, the shared 3D U-Net is run once per temporal direction and the two predictions are fused into a single noise estimate.

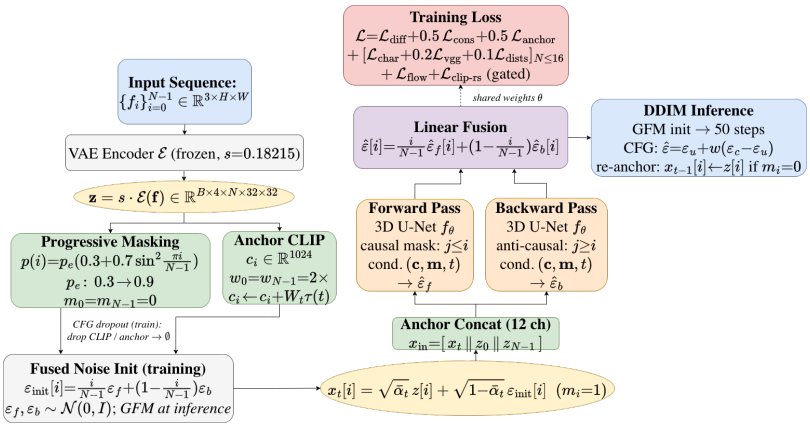


Figure 1: Overview of STRIDE. A frozen VAE encodes frames; a $\sin^2$ curriculum mask selects which to generate. Noised latents are concatenated with the two broadcast anchor latents (12-channel input), and anchors are CLIP-encoded with boundary-weighted cross-attention. The shared 3D U-Net runs a forward (causal) and backward (anti-causal) pass per step and the two noise predictions are linearly fused to compute the training objective or drive DDIM inference.

3.3 POSITION-AWARE PROGRESSIVE MASKING CURRICULUM

Our masking envelope concentrates on the hardest-to-interpolate interior frames while keeping boundary frames visible. Each interior frame $\mathbf{f}_i$ with $i \in \{1, \ldots, N{-}2\}$ is masked independently with probability:

$$p(i) = p_e \cdot \Big(0.3 + 0.7\sin^2\Big(\frac{\pi i}{N-1}\Big)\Big), \tag{1}$$

where the $\sin^2$ envelope peaks at $i = \lfloor N/2 \rfloor$. This makes the sinusoidal component vanish at the boundaries (leaving a residual base probability $0.3\,p_e$), but keep boundary frames unconditionally visible by construction ($m_0 = m_{N-1} = 0$). Independently, the *global* rate is warmed linearly, $p_e = p_{\text{init}} + (p_{\text{final}} - p_{\text{init}}) \min(e/E_{\text{warm}}, 1)$ (p_{init}=0.3, p_{final}=0.9, E_{warm}=5), together leading STRIDE from easy near-boundary reconstruction toward full-sequence interpolation.

3.4 Conditioning and Backbone

Multi-anchor CLIP conditioning. CLIP (ViT-L/14) (Radford et al., 2021) is applied only to anchor frames $\mathcal{A} = \{i : m_i = 0\}$. Masked-position embeddings are linearly interpolated between the nearest left $\mathbf{c}_l$ and right anchors $\mathbf{c}_r$: $\mathbf{c}_i = (1-\alpha)\mathbf{c}_l + \alpha\mathbf{c}_r$ with $\alpha = (i-l)/(r-l)$; since $\{0, N{-}1\} \subseteq \mathcal{A}$, every masked position is bracketed. In cross-attention, however, masked positions receive a $-\infty$ logit bias (see boundary weighting below), so *only the anchor embeddings serve as cross-attention keys/values* and generated frames attend to anchor tokens alone. The interpolated embeddings $\{\mathbf{c}_i\}$ are used solely by the CLIP relative-smoothness auxiliary loss (see section 3.6), not by the cross-attention.

Boundary weighting. Boundary frames receive $2\times$ cross-attention weight via additive log-space logit biases ($\log 2$ boundaries, $\log 1$ interior anchors, $-\infty$ masked positions), and the diffusion timestep is injected as a learned additive bias across all N context tokens (eq. (5), appendix A.3).

Anchor-latent concatenation. CLIP cross-attention supplies high-level semantics but not the precise pixel content of the endpoints. We therefore additionally condition the U-Net on the anchor *latents* by channel-wise concatenation ($\|$). The latents at positions 0 and $N{-}1$ are broadcast across all N temporal slots and concatenated with the noised latent to form a 12-channel input: $\mathbf{x}_t^{\text{in}} = [\,\mathbf{x}_t \,\|\, \text{br}(\mathbf{z}_0) \,\|\, \text{br}(\mathbf{z}_{N-1})\,] \in \mathbb{R}^{B\times 12\times N\times 32\times 32}$ (see eq. (6), appendix A.3), where $\text{br}(\cdot)$ denotes temporal broadcasting. This gives every frame direct, spatially-aligned access to both endpoints and is the primary anchor-appearance propagation path.

3D U-Net backbone. STRIDE uses a 3D U-Net with channel dimensions $[320, 640, 1280, 1280]$ (multipliers $[1, 2, 4, 4]$) and two `ResBlock3D` units per encoder/decoder level. Each level contains one `TransformerBlock3D` ordering `BidirectionalTemporalAttention` (causal/anti-causal mask, section 3.5), `SpatialSelfAttention` (per-frame), and `MultiAnchorCrossAttention` (CLIP conditioning, section 3.4), followed by a GELU feed-forward network ($4\times$ expansion). Each `ResBlock3D` block is conditioned on the sum of the timestep and a learned *direction* embedding ($\text{dir} = 1$ for forward and $\text{dir} = 0$ for backward). The spatial downsampling uses a stride $(1, 2, 2)$ in the `Conv3D` block, preserving the temporal dimension. The model has 1.54 B parameters, trained in `bfloat16` with gradient checkpointing.

3.5 Bidirectional Temporal Attention and Linear Fusion

Position-fused training noise. To keep the noise distribution consistent with the subsequent prediction fusion, we draw two independent samples and position-linearly fuse them before doing actual noising: $\boldsymbol{\varepsilon}_{\text{init}}[i] = \alpha_i\,\boldsymbol{\varepsilon}_{\text{fwd}}[i] + (1-\alpha_i)\,\boldsymbol{\varepsilon}_{\text{bwd}}[i]$ with $\alpha_i = i/(N{-}1)$ (see eq. (7), Appendix A.3), where $\boldsymbol{\varepsilon}_{\text{fwd}}, \boldsymbol{\varepsilon}_{\text{bwd}} \sim \mathcal{N}(0, I)$ are two independent noise tensors drawn *once per clip* (not per frame), $[\cdot][i]$ denotes the frame-i slice, and $\alpha_i = i/(N{-}1)$ is the normalized temporal position of frame i. This aligns the target with the fused directional predictions (see eq. (2)). Since the two Gaussians are independent, the fused noise has position-dependent variance $\alpha_i^2 + (1-\alpha_i)^2 \in [0.5, 1]$ that is minimized at the center. So the interior is mildly under-noised (we retain it for prediction/target consistency and revisit in section 5). Throughout this section, bare $\boldsymbol{\varepsilon}$ denotes a noise *sample* drawn from $\mathcal{N}(0, I)$ while hatted $\hat{\boldsymbol{\varepsilon}}$ denotes the model's *predicted* noise estimate. We add noise only to masked frames, $\mathbf{x}_t[i] = \sqrt{\bar{\alpha}_t}\,\mathbf{z}[i] + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}_{\text{init}}[i]$ for m_i=1 and $\mathbf{x}_t[i] = \mathbf{z}[i]$ otherwise (see eq. (8), appendix A.3).

Two masked passes with shared weights. Each denoising step invokes the 3D U-Net twice on the same input $\mathbf{x}_t^{\text{in}}$ with one shared set of $\mathbf{Q}/\mathbf{K}/\mathbf{V}$ weights and output projections, differing only in (i) the temporal-attention mask inside every `BidirectionalTemporalAttention` layer and (ii) the learned direction embedding added to the residual stream. The forward pass runs under the *causal* mask $\mathbf{M}^{\text{fwd}}[i, j]$=0 for $j \le i$ (else $-\infty$) and the backward pass under its transpose $\mathbf{M}^{\text{bwd}}$, giving $\hat{\boldsymbol{\varepsilon}}_{\text{fwd}} = f_\theta(\mathbf{x}_t^{\text{in}}, t, \mathbf{c}, \mathbf{m}, \text{dir}{=}1, \mathbf{M}^{\text{fwd}})$ and $\hat{\boldsymbol{\varepsilon}}_{\text{bwd}}$ analogously with $\text{dir}{=}0$ (see eqs. (9), (10),appendix A.3).

Position-linear fusion. We blend the predictions with a position-dependent weight $\alpha_i = \frac{i}{N-1}$:

$$\hat{\boldsymbol{\varepsilon}}_{\text{fused}}[i] = \alpha_i \, \hat{\boldsymbol{\varepsilon}}_{\text{fwd}}[i] + (1 - \alpha_i) \, \hat{\boldsymbol{\varepsilon}}_{\text{bwd}}[i]. \tag{2}$$

Frame 0 is governed entirely by the backward pass ($\alpha_0 = 0$) and frame $N-1$ by the forward pass ($\alpha_{N-1} = 1$). Critically, the blend weights α_i are chosen to match the noise-target weights of eq. (7). This keeps $\mathcal{L}_{\text{diff}}$ as well-defined mean square error (MSE) loss between aligned quantities. Anchor *appearance* is available to every frame at all positions via the 12-channel concatenation (see section 3.4), which our primary anchor-propagation path. The temporal-attention direction determines how interior frames share context with each other. The causal forward pass propagates information from anchor 0 outward, and the anti-causal backward pass propagates information from anchor $N-1$ inward, giving interior frames near each boundary access to the full opposite-side temporal context.

Classifier-free guidance. To enable guidance, we train the unconditional pathway with per-sample conditioning dropout (Ho & Salimans, 2022). The CLIP context $\mathbf{c}$ and the anchor channels are each zeroed with *total marginal* probability $p_{\text{clip}}{=}p_{\text{anchor}}{=}0.10$, of which $p_{\text{both}}{=}0.05$ is the *joint* probability that both are dropped in the same step (so CLIP-only, anchor-only, and joint dropout each occur with probability 0.05, and neither with 0.85 making the joint rate exceeds the 0.01 of independent dropout by design). At inference, we run a conditional and an unconditional pass and then extrapolate: $\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}_{\text{uncond}} + g_{\text{cfg}} \, (\hat{\boldsymbol{\varepsilon}}_{\text{cond}} - \hat{\boldsymbol{\varepsilon}}_{\text{uncond}})$, with guidance scale $g_{\text{cfg}} = 3.0$ (see eq. (12), appendix A.3). Each branch of eq. (12) is itself the fused estimate of eq. (2).

3.6 Training objectives

Diffusion loss. The primary objective is a masked MSE loss on the fused noise prediction, weighted by `Min-SNR-`γ (Hang et al., 2023) is $\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\boldsymbol{\varepsilon}}\big[w(t) \, \|\hat{\boldsymbol{\varepsilon}}_{\text{fused}} - \boldsymbol{\varepsilon}_{\text{init}}\|_2^2\big]$ over masked frames, with $w(t) = \min(\text{SNR}_t, \gamma)/\text{SNR}_t$ and $\gamma{=}5$ (see eq. (13), appendix A.3).

Bidirectional consistency loss. To enforce structural agreement between the two directional predictions, we apply a parabolic temporal weight that emphasizes the center frames, where the two predictions can diverge the most:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}\left[\frac{\sum_{i:m_i=1} 4\alpha_i(1{-}\alpha_i) \, \|\hat{\boldsymbol{\varepsilon}}_{\text{fwd}}[i] - \hat{\boldsymbol{\varepsilon}}_{\text{bwd}}[i]\|_2^2}{\sum_i m_i}\right]. \tag{3}$$

The factor $4\alpha_i(1 - \alpha_i)$ vanishes at the anchors and peaks at the center. This is the only term that explicitly penalizes directional disagreement as the direction-specific behavior itself is carried by the learned `DirectionEmbedding` and the causal/anti-causal masks.

Anchor reconstruction loss. Under ε-prediction and the zero-terminal-SNR schedule, anchor latents remain unnoised during training ($\mathbf{x}_t[i]{=}\mathbf{z}[i]$ for $m_i{=}0$) and are re-pinned at every *inference* step (see section 3.8), which keeps the terminal step from recovering signal from pure noise (Lin et al., 2024). We therefore recover the clean latent $\hat{\mathbf{z}}_0$ in closed form and penalize it at the anchor positions, $\mathcal{L}_{\text{anchor}} = \sum_{i:m_i=0} \|\hat{\mathbf{z}}_0[i] - \mathbf{z}[i]\|_2^2 \,/\, \sum_i (1 - m_i)$ (see eq. (15), appendix A.3). It acts as a *zero-noise consistency regularizer* that keeps the endpoints from drifting. Because anchor frames are not noised ($\mathbf{x}_t[i] = \mathbf{z}[i]$ for $m_i{=}0$), the inversion formula for $\hat{\mathbf{z}}_0$ involves a $\sqrt{\bar{\alpha}_t}$ denominator that is zero at $t = T$ (zero-terminal-SNR schedule). The $\text{clip}(\cdot, -1, 1)$ in eq. (15) bounds the result at this degenerate step; $t = T$ accounts for $1/T = 0.1\%$ of training steps and has no meaningful gradient through the clipped output, so the loss is well-behaved in practice.

Perceptual and auxiliary losses. Pixel/feature supervision requires costly VAE decoding, so every $K_{\text{dec}}{=}10$ steps (for $N \leq 16$) we decode $\hat{\mathbf{z}}_0$ for one random masked frame per item at 128^2 resolution under gradient checkpointing (VAE activations recomputed on the backward pass so gradients reach the U-Net through decoded RGB while the VAE stays frozen), applying $\mathcal{L}_{\text{perc}} = 0.10 \, \mathcal{L}_{\text{char}} + 0.20 \, \mathcal{L}_{\text{vgg}} + 0.10 \, \mathcal{L}_{\text{dists}}$ (Charbonnier (Barron, 2019), multiscale VGG-19 (Johnson et al., 2016), DISTS (Ding et al., 2020)). Two low-frequency auxiliaries (gated to $N \leq 16$) fire infrequently: 1) a RAFT flow-warp term (low-flow weighting) (Teed & Deng, 2020) ($\lambda_{\text{flow}}{=}0.10$, every 50 steps) and a 2) CLIP relative-smoothness term ($\lambda_{\text{clip}}{=}0.05$, every 100 steps). The gate bounds decoder memory as longer buckets ($N \in \{24, 32\}$) use the latent terms ($\mathcal{L}_{\text{diff}}, \mathcal{L}_{\text{cons}}, \mathcal{L}_{\text{anchor}}$) alone.

The total perceptual loss is: $\mathcal{L} = \mathcal{L}_{\text{diff}} + 0.5\,\mathcal{L}_{\text{cons}} + 0.5\,\mathcal{L}_{\text{anchor}}$ every step, plus the gated perceptual/auxiliary terms $\mathcal{L}_{\text{perc}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}}$ for $N \leq 16$ (see eq. (16), appendix A.3). Training uses `AdamW8bit` (Dettmers et al., 2021) with a cosine schedule warm-restarted at each phase boundary and `bfloat16` mixed precision. An `fp32` EMA shadow copy (decay 0.999) is used for validation and inference (check full hyperparameters in appendix A.14).

3.7 Four-Phase Training Curriculum

To prevent catastrophic forgetting while progressively exposing STRIDE to harder in-the-wild data, training proceeds in four phases over 45 epochs (see table 14). Augmenter and DAVIS are used at 100% throughout while the in-the-wild sources (Kinetics, LAVIB, WebVid) enter in Phases 2 to 3 and are *excluded* from Phase 4, a clean-data perceptual fine-tune at a warm-restarted $0.5\times$ learning rate. Training used 5 NVIDIA's A100-40GB GPUs at an effective batch $\approx$512, `bfloat16`, and `gradient checkpointing`, totaling $\approx$144 wall-clock hours ($\approx$722 GPU-hours).

3.8 Gaussian Filter Mixer Inference Initialization

Standard DDIM initializes from pure Gaussian noise, but the anchor latents are already known at inference, so noise-only initialization discards this information and introduces a train/test gap. We avoid this with *Gaussian Filter Mixer* (Zhang et al., 2024a), applied **only at inference**: $\mathbf{x}_T[i] = (1-\sigma(i))\,\mathbf{z}[i] + \sigma(i)\,\boldsymbol{\varepsilon}_i$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, I)$, $\sigma(i) = \sigma_0 \sin(\pi i/(N-1))$, $\sigma_0 = 0.5$ (eq. (17), appendix A.3). Anchors (m_i=0) supply their known $\mathbf{z}[i]$ directly; interior frames use the linear interpolation of the boundary anchor latents instead. This *endpoint-anchored* initialization keeps boundary-adjacent frames near their known anchors while giving interior frames more noise toward the center; combined with the midpoint-minimal noise variance of section 3.5, it partially seeds the midpoint ghosting of section 5.

DDIM denoising loop. We run 50 DDIM (Song et al., 2021) steps (T=1000), re-clamping anchor latents after each update and restoring boundary pixels from input images post-decode to remove VAE artifacts.

Annealed-fusion sampling. The position-linear fusion (2) averages the two estimates at every step. We need this coupled mean early when global structure forms, but late in sampling it leaves interior frames a soft average rather than committing to one side. We therefore *anneal* the fusion with s indexing the S=50 DDIM steps and decay horizon K_{AF}=18. A cosine schedule $\lambda_s = \frac{1}{2}(1 + \cos(\pi \min(s, K_{\text{AF}})/K_{\text{AF}}))$ decays from 1 to 0, and we replace eq. (2) by

$$\hat{\boldsymbol{\varepsilon}}^{\text{AF}}[i] = \lambda_s\,\hat{\boldsymbol{\varepsilon}}_{\text{fused}}[i] + (1-\lambda_s)\,\hat{\boldsymbol{\varepsilon}}_{\text{commit}}[i], \tag{4}$$

where $\hat{\boldsymbol{\varepsilon}}_{\text{commit}}[i]$ is $\hat{\boldsymbol{\varepsilon}}_{\text{fwd}}[i]$ if $\alpha_i > 0.5$ else $\hat{\boldsymbol{\varepsilon}}_{\text{bwd}}[i]$ (each interior frame committed to its nearer endpoint). Sampling thus begins from the coupled mean and cosine-anneals toward per-frame directional commitment, reducing the mid-sequence blend without a hard midpoint seam. We denote this *STRIDE (AF)* and select it empirically against the coupled-mean sampler and two stochastic variants as in section 4.4.

4 Experiments

4.1 Experimental Setup

We evaluate STRIDE against four pretrained-prior baselines: 1) DynamiCrafter, 2) SEINE, 3) TVG and 4) Wan2.1-VACE-1.3B over three transition benchmarks i.e., MorphBench (Zhang et al., 2024b), TC-Bench (Feng et al., 2024b), and TransitBench. We additionally include two bidirectional keyframe-interpolation methods namely, **SVD-KFI** (Wang et al., 2025b), and **Framer** (Wang et al., 2025a) in the automated 8-metric tables, each adapted to the two-anchor protocol via endpoint-reversal. We exclude both from the human study, since the endpoint-reversal adaptation confounds the model's own generative quality with artifacts introduced by adapting it to a protocol it was not designed for. Thus it makes human ratings an unreliable basis for comparison. Adding to the list, we exclude Time Reversal Fusion as it is an inference-time heuristic over a frozen prior with no matched two-anchor checkpoint.

Disjoint train/test split. The curated training subset shares sources with two test benchmarks (395 clips from MorphBench, 539 from TC-Bench, none from TransitBench). We preclude leakage by manually curating the MorphBench/TC-Bench test splits so that *no start-end anchor pair appears in*

training, and additionally remove any test clip whose endpoint pair is near-identical (CLIP-cosine) to a training pair. TransitBench is externally leakage-free by construction. We further ran an automated source-level audit (name-normalized source-key matching plus exact-content SHA-1 hashing of each test clip's start/end anchor against every training frame) across all 153,658 training clips (Vimeo-90K, Augmenter, UCF-101, Kinetics-400, WebVid10M, LAVIB, DAVIS) and found zero source-key or content-hash overlap with any of the 90/55/82 MorphBench/TC-Bench/TransitBench test clips. All methods are scored on the same 90/55/82 clips.

Protocol. Each method is evaluated at its own output resolution (see table 5) as we do not resize for the main tables (the resolution-fairness study of section 4.4 does). Every method produces *16 frames* per clip and we draw a *single stochastic sample* per clip. STRIDE generates the 14 interior frames with 50 DDIM steps, guidance $g_{\text{cfg}} = 3.0$, and native-resolution output (generated frames Lanczos-upsampled (Duchon, 1979) to source resolution, anchors reloaded losslessly). The held-out fidelity study (see section 4.4) instead uses the $N{=}7$ Vimeo-90K septuplet protocol (two anchors, five real interiors as ground truth). All automatic-metric evaluation and the human study use the Stage-A EMA checkpoint (a lightly smoothness-fine-tuned variant of `phase_4_end`, epoch 45) with annealed-fusion inference (STRIDE (AF) (see appendix A.1 for checkpoint lineage).

4.2 Evaluation Metrics

We report eight metrics from one shared harness: CLIPSim (Radford et al., 2021), VSI (Zhang et al., 2014), FSIM (Zhang et al., 2011), EMD (Rubner et al., 2000), DreamSim (Fu et al., 2023), LPIPS (Zhang et al., 2018), tPSNR, and tMS-SSIM (Wang et al., 2003). *All eight are no-reference and computed between consecutive generated frames* (i vs. $i{+}1$), measuring inter-frame *consistency*. We group them into two families i.e., *structural consistency* (tPSNR, tMS-SSIM, VSI, FSIM) and *perceptual/appearance similarity* (CLIPSim, LPIPS, DreamSim, EMD). Both of which reward inter-frame smoothness; neither measures motion fidelity or semantic diversity (a model emitting near-identical frames scores well on all eight). Two table conventions: VSI and FSIM are stored as $1 -$ similarity (*lower is better*); the "t" prefix marks consecutive-frame use of PSNR/MS-SSIM (remaining conventions in appendix A.4). Because all eight reward smoothness, they do not expose mid-sequence ghosting (see the U-curve (section 5) for that analysis.

4.3 Quality Results

Full per-benchmark tables and the Pareto frontier plot are in appendix A.4 (figure 5). **STRIDE attains the best inter-frame structural consistency: VSI, FSIM, and tMS-SSIM on all three benchmarks, and tPSNR on one of three (MorphBench)** while Framer attains the highest tPSNR on TC-Bench and SVD-KFI on TransitBench. SVD-KFI (Wang et al., 2025b) instead achieves the strongest perceptual/appearance similarity, winning CLIPSim, DreamSIM, LPIPS, and EMD on all three benchmarks at $\approx 49\times$ STRIDE's wall-clock cost. Wan and Framer occupy intermediate positions while DC and SEINE are dominated on both metric families. Because all eight metrics are computed between consecutive frames, they should not be read as motion-fidelity measures (see section 4.2) and the split between the two families is an operating-point difference, not a quality ranking.

4.4 Further Quantitative Analyses

Resolution fairness. Re-scoring the four fixed-square-resolution baselines (STRIDE, SEINE, TVG, DynamiCrafter) at a common 256^2 and 512^2 (the 512^2 disadvantages STRIDE with exclusions in appendix A.5), STRIDE's advantages (VSI, FSIM, tPSNR, tMS-SSIM) persist by wide margins on all three benchmarks with the appearance ranking preserved, so they are not a 256^2 artifact.

Fidelity to ground truth. Against the *real* interiors of the held-out Vimeo-90K septuplet split (7,732 clips), STRIDE attains LPIPS 0.330, PSNR 19.9 dB, MS-SSIM 0.728, DreamSim 0.400, degrading monotonically boundary→midpoint (see appendix A.7). The forward and backward predictions disagree by a mean relative ℓ_2 of 0.110 averaged over interior frames on the held-out split, showing that the two directional branches produce systematically different noise estimates, so their fusion is non-redundant.

Sampler selection. On a separate 48-clip pool (TVS-48), annealed fusion (AF) attains the best smoothness (4.08 vs. 3.48) and average (Holm-Wilcoxon $p{=}0.034$ vs. baseline), tying on ghosting/commitment while both stochastic variants regress. So we adopt AF.

4.5 Comparative Human Study

Since no-reference metrics can't express holistic human preference or mid-sequence ghosting, we run a controlled subjective study of STRIDE (AF) against DynamiCrafter, SEINE, TVG, and Wan2.1-VACE-1.3B. To let readers judge the transitions directly, we provide an anonymous supplementary gallery. [2]

Design. CTB-48 (48 clips, stratified by difficulty and disjoint from the TVS-48 sampler-selection pool) yields a fully-crossed within-subjects design ($48\times5\times3=720$ clip-method-rater evaluations, each scored on three Likert dimensions for $2{,}160$ dimension-level scores(see appendix A.13). Raters give three 1–5 Likert scores for each video shown in randomized order with method identity hidden: **Smoothness**, **Ghosting** (absence of superimposed content), and **Semantic Commitment** (single coherent intermediate).

Statistical analysis. Responses are ordinal and repeated-measures, so we use Krippendorff's α (Krippendorff, 2011), $\mathrm{ICC}(2,k)$, and Kendall's W for inter-rater reliability, a Friedman omnibus per dimension, and Holm-corrected pairwise Wilcoxon tests (Holm, 1979) (rank-biserial r) where significant, summarized by critical-difference (CD) diagrams (Demšar, 2006) (full tables in appendix A.11). Inter-rater reliability is *high* on every dimension ($\alpha \in [0.84, 0.87]$, $\mathrm{ICC}(2,k) \in [0.94, 0.95]$, Kendall $W \in [0.83, 0.90]$).

Results. Figure 2 and table 2 (appendix A.2) give the outcome. **STRIDE is preferred for smoothness by a large margin** (mean 3.92 vs. $\leq$2.92; Friedman $\chi^2(4)=73.12$, $p=5\times10^{-15}$; matched-pairs rank-biserial $r=0.74$–0.99 across the four baselines, with $r\geq0.97$ over DC/SEINE/TVG and $r=0.74$ over the closer Wan). **Wan2.1-VACE-1.3B is significantly best on commitment** (mean 3.11 vs. $\leq$2.40; $p<0.001$ vs. all others) while **STRIDE, DC, SEINE, and TVG are not significantly different after Holm correction** (all 2.21–2.40). **On ghosting, STRIDE ranks below all four baselines** (mean 2.24 vs. 2.98–3.20) which forms the subjective counterpart of the U-curve. Full CD diagrams and pairwise tables are in appendix A.11.

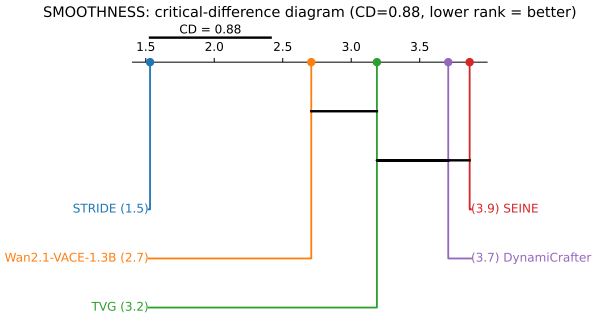


Figure 2: Critical-difference diagram for *smoothness* ($N=48$, five methods. A bar groups methods not significantly different, Holm-corrected Wilcoxon, table 21). STRIDE (AF) clears all four baselines beyond the critical difference (all pairwise $p\ll0.001$)

4.6 Efficiency

Table 5 compares complexity and runtime. STRIDE generates a 16-frame clip in 11 s at 256^2, $\mathbf{3.7}\times$ faster than Wan (624^2, 40 s) and 6.7–43$\times$ faster than the 768^2 baselines. Pixel-normalized GFLOPs (col. [h]) yield a $\mathbf{4}$–$\mathbf{19}\times$ compute advantage. STRIDE's bucketed curriculum also supports variable-length output ($N\in\{8,16,24,32\}$) from a single checkpoint, unlike the fixed-length baselines (see appendix table 3 for runtime/quality tradeoff across N on a 10-clip MorphBench subsample).

5 Limitations

STRIDE's central limitation is *mid-sequence mean-regression*: frames near the temporal center relax toward a smooth crossfade blend of the two anchors rather than tracking true intermediate motion,

[2] Anonymous supplementary videos: `https://anonstride2026.github.io/stride-transitions/`. Each row plays the same start→end transition rendered by STRIDE and four baselines (DynamiCrafter, SEINE, TVG, Wan2.1-VACE) from identical boundary frames. These 8 rows illustrate section 4.5's smoothness/commitment findings. Full 48-clip results, including losses, are in table 20.

producing soft ghosting on hard transitions. The boundaries stay crisp; this is a graceful loss of high-frequency mid-sequence detail, not structural collapse.

The U-curve: measurement, source, and outlook. The U-curve (see fig. 7) makes the limitation measurable: per-frame pixel standard deviation is near unity at the boundaries and dips toward the center, deeper dips indicating stronger mean-regression that is undetectable by the consecutive-frame metrics of section 4.2. We hypothesize it arises from a generative-prior-free ε-prediction objective whose conditional mean tends toward a blurred average, compounded by the fused-noise target (section 3.5) whose variance $\alpha_i^2+(1-\alpha_i)^2$ is minimized to 0.5 at the center; a pretrained prior would supply the missing high-frequency content (Feng et al., 2024a). Two inference choices further compound this: the GFM seed already carries 50% crossfade blend at the midpoint, and $\mathcal{L}_{\text{cons}}$'s parabolic peak maximally penalizes directional disagreement where commitment is hardest.

Negative results: fine-tuning and loss ablation. A warm-started fine-tune ladder (refer appendix A.10) retained one gain that the stronger per-frame supervision reduced flicker but left the ghost unchanged. Variance-preserving rescaling of $\boldsymbol{\varepsilon}_{\text{init}}[i]$ was absorbed in $\approx$100 steps with no U-curve change, confirming a mode error independent of noise scale. A six-arm loss ablation (see table 17) found that *removing* Min-SNR-γ (γ=5) uniquely improves mid-sequence fidelity (U-curve $\downarrow$ on $3/4$ metrics, $p<0.001$). The regularizer suppresses high-SNR steps where midpoint structure is hardest to learn. We additionally tested replacing GFM's crossfade-seeded initialization with pure Gaussian noise (anchors still pinned and `no_gfm`, evaluated on the same n=45 DAVIS gt16 gate as table 17). This collapses overall fidelity on all four metrics (LPIPS $+0.220$, PSNR -5.75 dB, MS-SSIM -0.360, DreamSim $+0.383$; all Holm-corrected $p < 0.001$) and even worsens PSNR's own `u_depth` ($+1.174$, p=0.001). This apparent narrowing on the remaining three `u_depth` metrics reflects uniformly degraded output rather than reduced mid-sequence commitment. GFM initialization is therefore necessary for coherent generation, not merely a source of the ghost.

Comparison caveats. Our benchmark comparison carries two caveats we state explicitly. First, baselines are generated at their configured square resolution, which for DynamiCrafter (768^2; native 320×512) exceeds its native training resolution; the resolution-fairness re-score (appendix A.5) equalizes the *measurement* but not the *generation* resolution, so we restrict resolution-fair claims to the structural-consistency metrics (which persist at both 256^2 and 512^2) and the wall-clock and pixel-normalized compute of table 5, and leave native-resolution baseline generation to a revision. Second, all benchmark metrics are consecutive-frame and no-reference, so comparative motion fidelity against the baselines is not established (the ground-truth study of section 4.4 reports STRIDE in isolation).

6 Conclusion and Future Work

We presented STRIDE, a 1.54B *generative-prior-free* diffusion model for video transition generation in a compact 256^2 pixel space (32^2 latent). Across all three benchmarks (MorphBench, TC-Bench, TransitBench), STRIDE occupies a high temporal-smoothness operating point as it substantially improves consecutive-frame consistency (tPSNR, tMS-SSIM, VSI, FSIM) and human-rated smoothness while sacrificing intermediate-frame fidelity and semantic commitment near the transition midpoint. STRIDE anchors the structural-consistency end while SVD-KFI (Wang et al., 2025b) anchors the perceptual-similarity end, winning CLIPSim, DreamSIM, LPIPS, and EMD (all consecutive-frame appearance-similarity metrics) on all three benchmarks at $\approx49\times$ STRIDE's wall-clock inference cost with Wan2.1-VACE-1.3B and Framer occupying an intermediate position. This frontier is a generative-prior-free vs. pretrained-prior operating-point difference, not a simple quality ranking. In our five-method human study ($5\times48\times3$=720 clip-method-rater evaluations; Krippendorff α=0.84–0.87), STRIDE is significantly preferred for transition smoothness over all four baselines (Holm-corrected Wilcoxon, r=0.74–0.99); on semantic commitment, Wan leads significantly while STRIDE, DynamiCrafter, SEINE, and TVG are not significantly different after Holm correction. At 3.7–43$\times$ lower wall-clock inference latency and $\approx$36–181$\times$ lower per-clip compute than DynamiCrafter, SEINE, and TVG (3.7$\times$ over Wan), STRIDE demonstrates that pretrained generative priors are not necessary for highly smooth bounded transitions but that the absence of such a prior exposes a previously hidden midpoint-fidelity failure mode that consecutive-frame metrics miss. This cost is the mid-sequence ghosting of section 5 that is unresolved by our fine-tune ladder. In our future work, we target to remove this U-curve via pretrained-generative-prior and a *boundary-pinned stochastic bridge* with anchor-pinned marginals.

AI USE STATEMENT

We used LLM (Claude) to support drafting (e.g. paraphrasing for clarity) and formatting during manuscript preparation. We also used it for initial prototype code which we then revised into the final implementation. All generated text was reviewed and verified by all authors. Further on, all scientific ideas, methodological choices, analyses, interpretations, and conclusions presented in this paper are solely those of the authors. In particular, the training and evaluation data and human-study ratings were not generated by an LLM as they were collected and prepared by the authors and human raters. The authors take full responsibility for the correctness, originality, and final form of the manuscript and all reported results.

ETHICS STATEMENT

This work includes a small subjective study in which three adult volunteers (excluding authors) rated short generated video clips. Participation was voluntary. Raters gave informed consent after receiving a written description of the task and rubric, were free to stop at any time, and were not asked for any personal information. Only anonymized ordinal ratings are reported (see appendix A.13). We also note the dual-use risk of video-transition and morphing models, which could in principle be misused to fabricate misleading footage of real people (some public benchmarks contain faces). STRIDE targets short, bounded-keyframe transitions rather than identity synthesis, operates at a low 256^2 resolution, and carries an honestly-reported mid-transition ghosting limitation (see section 5). We release it as a research baseline and discourage any deceptive or non-consensual use.

REPRODUCIBILITY STATEMENT

All results (automatic metrics, GT fidelity, and human study) use a single Stage-A EMA checkpoint (light smoothness fine-tune of `phase_4_end`, epoch 45) under a fully specified inference schedule (50 DDIM steps, guidance g_{cfg}=3.0, GFM σ_0=0.5, annealed-fusion horizon K_{AF}=18, eq. (4)) and a shared eight-metric harness (section 4.2) applied identically to every method, including the officially released baselines at default settings (see section 4.1). The training recipe, four-phase curriculum, and motion-aware data curation are given in sections 3.6–3.1 and table 14. The human-study design, protocol, and statistical pipeline are detailed in appendices A.11 and A.13, with an anonymous video gallery linked from section 4.5.

REFERENCES

Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4331–4339, 2019.

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.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1472–1480, 2024.

Abhirup Das, Nishant Singh, and Anubha Gupta. Smooth or jarring? evaluating video transitions with transisense and vt-bench. In *International Conference on Pattern Recognition*, pp. 310–326. Springer, 2026.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2021.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Claude E Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology and Climatology*, 18(8):1016–1022, 1979.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pp. 363–370. Springer, 2003.

Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, pp. 378–395. Springer, 2024a.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024b.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7441–7451, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European conference on computer vision*, pp. 624–642. Springer, 2022.

Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7341–7351, 2024.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9000–9008, 2018.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. *Departmental Papers (ASC), University of Pennsylvania*, 2011.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5404–5411, 2024.

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

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pp. 250–266. Springer, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

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

Wan Team. Wan: Open and advanced large-scale video generative models, 2025.

Wen Wang, Qiuyu Wang, Kecheng Zheng, Hao Ouyang, Zhekai Chen, Biao Gong, Hao Chen, Yujun Shen, and Chunhua Shen. Framer: Interactive frame interpolation. In *International Conference on Learning Representations*, volume 2025, pp. 51601–51628, 2025a.

Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steve Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 30594–30609, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4bbdef62653d8088717640e7660a1ebb-Paper-Conference.pdf.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thirty-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Serin Yang, Taesung Kwon, and Jong Chul Ye. Vibidsampler: Enhancing video interpolation using bidirectional diffusion sampler. In *International Conference on Learning Representations (ICLR)*, 2025a.

Zuhao Yang, Jiahui Zhang, Yingchen Yu, Shijian Lu, and Song Bai. Versatile transition generation with image-to-video diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16981–16990, 2025b.

Bowen Zhang, Xiaofei Xie, Haotian Lu, Na Ma, Tianlin Li, and Qing Guo. Mavin: Multi-action video generation with diffusion models via transition video infilling. *arXiv preprint arXiv:2405.18003*, 2024a.

Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5682–5692, 2023.

Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10):4270–4281, 2014.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Rui Zhang, Yaosen Chen, Yuegen Liu, Wei Wang, Xuming Wen, and Hongxia Wang. Tvg: A training-free transition video generation method with diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

A Supplementary Material

A.1 Deferred Main-Text Floats and Protocol Details

The following items are referenced from the main text and deferred here for brevity.

Configuration lineage (section 4.1). All reported results i.e., automatic metrics, GT fidelity, and the human study use a single Stage-A EMA checkpoint (a light smoothness fine-tune of `phase_4_end`, epoch 45) that we call as STRIDE (AF). Stage-A is a short 5-epoch fine-tune, warm-started (weights and EMA) from `phase_4_end` at a warm-restarted learning rate (5×10^{-5}, $\eta_{\min}=10^{-5}$, weight decay 0.01, EMA decay 0.99, Min-SNR $\gamma=5$), on clean data only (Vimeo-90K, UCF-101, Augmenter, DAVIS. The in-the-wild Kinetics/WebVid/LAVIB sources are excluded, as in frozen Phase 4). It adds denser per-frame RGB/perceptual supervision (secondary decode every 3 steps rather than 10) and a 25% static-background oversampling weight, both targeting flicker rather than the benchmark metrics. The checkpoint was retained using a 30-clip flicker diagnostic probe (detrended flicker index, CLIP-RS, and Ghost Index (see appendix A.10), and the annealed-fusion sampler was selected on the separate TVS-48 pool (see section 4.4). Both selection pools are disjoint from the MorphBench/TC-Bench/TransitBench test clips, so no scored benchmark clip informed model or sampler selection. The fine-tuning ablation ladder (see appendix A.10) is warm-started from the same base `phase_4_end` checkpoint as a separate, diagnostic-only exploratory branch that backs no reported result.

Table 1: Training datasets, clip counts used by STRIDE, and their role across the four-phase curriculum (Section 3.7). A subset of VT-Bench (curated, mean rating $>3.5/5$, called Augmenter) and DAVIS are used at full proportion in every phase. Phase 4 excludes the in-the-wild sources (Kinetics, LAVIB, WebVid). Counts are on-disk pools after motion filtering, not full published sizes.

Dataset	# clips	Domain / role	Phase(s)
Augmenter	7,117	Rated transitions (anchor)	1–4
Vimeo-90K	∼36,000	VFI septuplets	1–4
DAVIS 2017	46	Object boundaries	1–4
UCF-101	6,627	Sports / human motion	1–4
Kinetics-400	39,125	Human actions	2–3
LAVIB	15,000	Large-scale web VFI	2–3
WebVid-10M	49,743	In-the-wild web video	3

Baselines, checkpoints, and resolutions. We run each baseline from its official public release with default settings. DynamiCrafter uses the `dynamicrafter_512_interp_v1` interpolation checkpoint (whose native resolution is 320×512), which we run at a square 768×768 to match our benchmark configuration and the 768^2 we quote in the table 5. As with SEINE, this places DynamiCrafter above its native training resolution, i.e., at an off-distribution square resolution, which the resolution-fairness re-score (see section 4.4) subsequently controls. As DynamiCrafter is text-conditioned, we supply a fixed caption "`smooth transition`" so the result is driven by the two image endpoints rather than a scene-specific prompt. TVG is training-free as it reuses the same pretrained DynamiCrafter weights as its generative backbone, but replaces DynamiCrafter's DDIM

sampling with a Gaussian-Process regression over the latent trajectory to interpolate between the two boundary anchors. Because the backbone is unchanged, TVG shares DynamiCrafter's 768^2 operating point. We report its inference time *excluding* the optional LPIPS-based per-frame selection pass, because that pass is an application-level post-filter absent from the other methods and would otherwise make the runtime comparison inconsistent (including it does not change the reported quality metrics, which are computed on the selected frames). SEINE's public checkpoint is a LaVie/Latte-based model released at a 320×512 base resolution. We generate its transitions directly with SEINE's own masked-sampling pipeline at 784^2 (the `image_size` of its sampling configuration) as the frames are produced by SEINE's diffusion U-Net and VAE with *no* super-resolution or upsampling stage, so the output is genuinely generated rather than upscaled. Because 784^2 is above SEINE's released base resolution, this is an *off-distribution* resolution, controlled for by the resolution-fairness re-score (see section 4.4).

Disjoint train/test split. Section 4.1 gives the leakage-control methodology and clip counts (395/539/0 source overlap, 90/55/82 scored test clips. The one point worth adding is that the 90/55/82 figures are the exact sample sizes underlying every per-metric mean and standard deviation reported in section 4.3 and appendix A.4. It is specifically the resolution-fairness re-score (see section 4.4) and not the leakage control that accounts for why raw scores are not comparable across the three benchmarks.

A.2 Comparative Human Study: Summary Table

The main-text summary of the comparative human study (see section 4.5) is reproduced here while the per-dimension full statistics follow in appendix A.11.

Table 2: Comparative human study (CTB-48, 48 clips $\times$ 5 methods $\times$ 3 raters $=$ 720 clip-method-rater evaluations). Cells are mean(std) (1–5, higher = better); **bold** marks the best method per dimension. Friedman omnibus significant on all four dimensions ($p\leq2\times10^{-3}$). Holm-corrected pairwise Wilcoxon confirms STRIDE best on *Smoothness* (over all four), SEINE best on *Ghosting* (over STRIDE), Wan best on *Commit.* (over all four others); *Average*: Wan and STRIDE rank first/second. Full median [IQR], ranks, and pairwise tests in appendix A.11.

Method	Smooth. ↑	Ghost. ↑	Commit. ↑	Avg. ↑
STRIDE (AF)	**3.92(0.60)**	2.24(0.62)	2.21(0.64)	2.79(0.49)
DynamiCrafter	1.99(1.02)	2.98(1.18)	2.33(1.00)	2.43(0.89)
SEINE	1.92(1.06)	**3.20(1.20)**	2.23(1.29)	2.45(0.98)
TVG	2.40(1.19)	2.99(1.30)	2.40(1.09)	2.60(1.06)
Wan2.1-VACE-1.3B	2.92(1.28)	3.07(1.12)	**3.11(1.15)**	**3.03(1.06)**

A.3 Full Method Details

For completeness, we collect here the displayed equations that are summarized inline in section 3 while the main-text prose and the three novel equations (fusion see eq. (2), annealed fusion see eq. (4), and consistency see eq. (3)) are unchanged.

Boundary weighting (section 3.4).

$$w_i = \begin{cases} 2 & i \in \{0, N-1\}, \\ 1 & i \in \mathcal{A} \setminus \{0, N-1\}, \\ 0 & m_i = 1. \end{cases} \tag{5}$$

Anchor-latent concatenation (section 3.4).

$$\mathbf{x}_t^{\text{in}} = \left[\, \mathbf{x}_t \,\|\, \text{br}(\mathbf{z}_0) \,\|\, \text{br}(\mathbf{z}_{N-1}) \,\right] \in \mathbb{R}^{B\times12\times N\times32\times32}, \tag{6}$$

Position-fused training noise (section 3.5).

$$\boldsymbol{\varepsilon}_{\text{init}}[i] = \frac{i}{N-1}\,\boldsymbol{\varepsilon}_{\text{fwd}}[i] + \left(1 - \frac{i}{N-1}\right)\boldsymbol{\varepsilon}_{\text{bwd}}[i], \tag{7}$$

Masked noising (section 3.5).

$$\mathbf{x}_t[i] = \begin{cases} \sqrt{\bar{\alpha}_t}\,\mathbf{z}[i] + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}_{\text{init}}[i] & m_i = 1, \\ \mathbf{z}[i] & m_i = 0. \end{cases} \tag{8}$$

Directional attention masks (section 3.5).

$$\mathbf{M}^{\text{fwd}}[i,j] = \begin{cases} 0 & j \le i \\ -\infty & j > i \end{cases} \quad \mathbf{M}^{\text{bwd}}[i,j] = \begin{cases} 0 & j \ge i \\ -\infty & j < i \end{cases} \tag{9}$$

Forward/backward passes (section 3.5).

$$\hat{\boldsymbol{\varepsilon}}_{\text{fwd}} = f_\theta\left(\mathbf{x}_t^{\text{in}}, t, \mathbf{c}, \mathbf{m}; \text{dir} = 1, \mathbf{M}^{\text{fwd}}\right) \tag{10}$$

$$\hat{\boldsymbol{\varepsilon}}_{\text{bwd}} = f_\theta\left(\mathbf{x}_t^{\text{in}}, t, \mathbf{c}, \mathbf{m}; \text{dir} = 0, \mathbf{M}^{\text{bwd}}\right) \tag{11}$$

Classifier-free guidance (section 3.5).

$$\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}_{\text{uncond}} + g_{\text{cfg}}\left(\hat{\boldsymbol{\varepsilon}}_{\text{cond}} - \hat{\boldsymbol{\varepsilon}}_{\text{uncond}}\right), \tag{12}$$

Diffusion loss (section 3.6).

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\boldsymbol{\varepsilon}}\left[w(t) \cdot \frac{\sum_{i:m_i=1} \|\hat{\boldsymbol{\varepsilon}}_{\text{fused}}[i] - \boldsymbol{\varepsilon}_{\text{init}}[i]\|_2^2}{\sum_i m_i} \right], \tag{13}$$

$$w(t) = \frac{\min(\text{SNR}_t,\, \gamma)}{\text{SNR}_t}, \quad \text{SNR}_t = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}, \quad \gamma = 5.$$

Anchor reconstruction loss (section 3.6).

$$\hat{\mathbf{z}}_0 = \text{clip}\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\boldsymbol{\varepsilon}}_{\text{fused}}}{\sqrt{\bar{\alpha}_t}},\, -1,\, 1 \right), \tag{14}$$

$$\mathcal{L}_{\text{anchor}} = \frac{\sum_{i:m_i=0} \|\hat{\mathbf{z}}_0[i] - \mathbf{z}[i]\|_2^2}{\sum_i (1-m_i)}. \tag{15}$$

At $t = T$ with $\bar{\alpha}_T = 0$, the denominator $\sqrt{\bar{\alpha}_T} = 0$ is degenerate; the clip bounds the output to $[-1, 1]$ so the loss remains finite. Since anchor frames are not noised ($\mathbf{x}_T[i] = \mathbf{z}[i]$ for m_i=0), the unclamped ratio would be $(\mathbf{z}[i] - \hat{\boldsymbol{\varepsilon}}_{\text{fused}}[i])/0$; the clamp replaces this with ± 1, yielding a trivial constant loss at $t = T$. This degenerate step occurs with probability $1/T = 0.001$ and contributes no informative gradient.

Total objective (section 3.6).

$$\begin{aligned} \mathcal{L} = & \underbrace{\mathcal{L}_{\text{diff}} + 0.5\,\mathcal{L}_{\text{cons}} + 0.5\,\mathcal{L}_{\text{anchor}}}_{\text{every step}} \\ & + \underbrace{\mathcal{L}_{\text{perc}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}}}_{\text{gated, } N \le 16}. \end{aligned} \tag{16}$$

GFM initialization (section 3.8).

$$\mathbf{x}_T[i] = \left(1 - \sigma(i)\right)\mathbf{z}[i] + \sigma(i)\,\boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, I), \tag{17}$$

At inference, $\mathbf{z}[i]$ is the known anchor latent for m_i=0, and the linear interpolation $\mathbf{z}[i] = (1-\alpha_i)\mathbf{z}[0] + \alpha_i\mathbf{z}[N-1]$ for interior frames m_i=1 (true latent unavailable), with $\alpha_i = i/(N-1)$.

A.4 Native-Resolution Quality Tables

Remaining metric conventions (section 4.2). CLIPSim is computed with CLIP ViT-B/32, not the ViT-L/14 used for conditioning, and EMD is a 16-bin blue-channel histogram distance. Tables 6, 7, and 8 give the full per-benchmark numbers behind the pattern summarized in section 4.3.

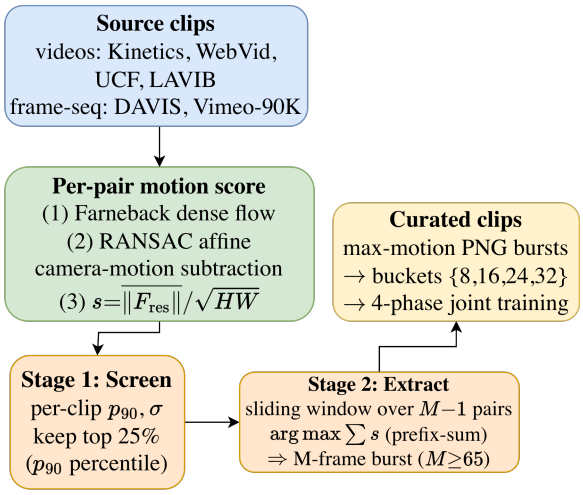


Figure 3: Motion-aware training-data curation (see section 3.1). The motion metric shown in the flowchart is $s = \|\mathbf{F}_{\text{res}}\|/\sqrt{HW}$, where $\mathbf{F}_{\text{res}}$ is the mean residual Farnebäck optical-flow magnitude after RANSAC camera-motion removal, normalized by image resolution $\sqrt{HW}$. Stage 1 screens whole clips by their p_{90} motion score and keeps the top 25%, and Stage 2 extracts the contiguous M-frame window ($M \geq 65$) of maximal cumulative motion from each retained clip. The curated bursts feed the bucketed four-phase training curriculum.

Table 3: Frame-count (N) sweep on a fixed 10-clip MorphBench subsample (STRIDE, Stage-A EMA, AF sampler, 50 DDIM steps, g_{cfg}=3.0). Runtime is mean wall-clock inference time per clip; tMS-SSIM is the same consecutive-frame metric as table 6. Larger N trades runtime for smoother transitions.

N	Runtime (s/clip)	tMS-SSIM ↑
8	11.75	0.856
16	13.49	0.932
24	16.23	0.950
32	19.35	0.955

A.5 Per-Benchmark Resolution-Fairness Tables

Section 4.4 gives the headline conclusion that the structural-fidelity wins are not a 256^2 artifact. Here, we lay the protocol and full numbers behind it. The valid control equalizes the *measurement*, not the *generation*, resolution (forcing the baselines to generate at 256^2 would cripple them out of distribution). We therefore keep every method's native output and *re-score* the four fixed-square-resolution methods after resizing each frame with an identical filter to a common square resolution, re-running the unchanged eight-metric protocol at 256^2 and at the more conservative 512^2 (which disadvantages STRIDE, the upsampled method, while the baselines are downsampled). Wan2.1-VACE-1.3B is excluded from these re-score tables as its native resolution varies by benchmark (624^2 for MorphBench/TransitBench, 464×832 for TC-Bench), making a single common-resolution re-score protocol ill-defined. SVD-KFI and Framer (added to the benchmark suite after this resolution-fairness control was run) are not yet included in this re-score and we plan to extend it to all seven methods as a revision. The resolution-fair comparison for Wan, SVD-KFI, and Framer is the wall-clock and pixel-normalized FLOP figures already reported in section 4.6 and table 5. Tables 9–11 give the full eight-metric re-score for the four fixed-resolution methods. Empirically, MorphBench tMS-SSIM is $0.94/0.91$ at $256/512$ vs. $0.85/0.81$ for the best baseline. The appearance-similarity ranking is likewise preserved (TVG retains CLIPSim, EMD, DreamSim, LPIPS; STRIDE ties LPIPS at 256^2 on MorphBench and rises modestly at 512^2, as expected for an upsampled source, with no perceptual claim). On hard transitions (see figure 4), the baselines tend to produce a translucent double-exposure of both endpoint contents superimposed at the midpoint. STRIDE avoids that superimposition artifact as its midpoint instead mean-regresses toward a smooth crossfade blend of the



Figure 4: Qualitative comparison on a representative hard transition (MorphBench, Musk→Obama). Columns sample normalized time $t \in [0, 1]$; the first and last columns are the given boundary frames. STRIDE keeps the endpoints pinned and the sequence temporally stable, but the interior mean-regresses toward an anchor blend that is visible at t=0.5 as a crossfade rather than a true semantic morph. This is the U-curve limitation quantified in Fig. 7; we report it rather than hide it.

anchors (the U-curve limitation described in section 5), which is perceptually distinct from the baselines structural double-exposure. Because this control equalizes measurement but not generation resolution, we do not claim the *appearance*-similarity comparison is resolution-fair; the resolution-fair conclusions are the structural-consistency wins (which hold at both 256^2 and 512^2) and the wall-clock and pixel-normalized compute of table 5 (see the comparison caveat in section 5).

Table 4: Where STRIDE sits among boundary-conditioned / bidirectional transition methods. "Prior" = depends on a pretrained image/video generative backbone; "Bidir." = uses a forward/backward mechanism; "Trained" = the bidirectional behavior is learned (vs. a test-time heuristic). STRIDE is the only entry that is generative-prior-free *and* trains its bidirectional fusion.

Method	Prior	Bidir.	Trained
TRF (Feng et al., 2024a)	✓	✓	–
SVD-KFI (Wang et al., 2025b)	✓	✓	✓
Framer (Wang et al., 2025a)	✓	–	✓
SEINE (Chen et al., 2023)	✓	–	✓
DynamiCrafter (Xing et al., 2024)	✓	–	✓
TVG (Zhang et al., 2025)	✓	✓	–
Wan2.1-VACE (Wan Team, 2025)	✓	–	✓
STRIDE (ours)	–	✓	✓

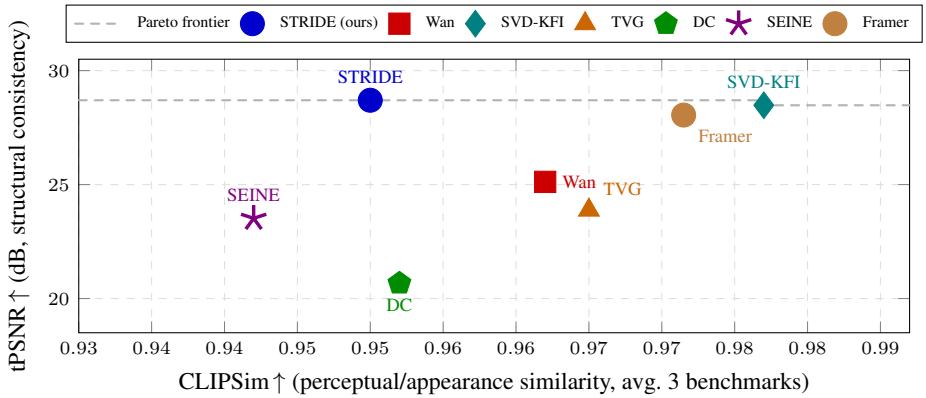


Figure 5: Pareto frontier (dashed line) between inter-frame structural consistency (tPSNR ↑) and perceptual/appearance similarity (CLIPSim ↑), averaged over MorphBench, TC-Bench, and TransitBench. Both axes are consecutive-frame smoothness measures, not motion-fidelity or semantic-diversity measures. STRIDE anchors the high-consistency end; SVD-KFI anchors the perceptual-similarity end at ≈49× STRIDE's wall-clock cost while Wan and Framer are intermediate. DC and SEINE are dominated on both axes.

A.6 Distribution-Level Fidelity (FVD Control)

Table 12 reports the FVD control discussed in section 5. A linear crossfade is pushed through STRIDE's exact pipeline with the diffusion removed (VAE encode–decode, anchors restored). The $+1.3$ FVD added by the VAE round-trip localizes STRIDE's gap to the generative process rather than the frozen autoencoder. This is the small-motion Vimeo-90K regime, which is *not* the large-motion transition setting STRIDE targets.

A.7 Fidelity to Ground Truth Results

Section 4.4 gives the headline fidelity numbers while here we detail out the protocol and the edge-to-midpoint breakdown behind them. We score STRIDE's generated interiors, with no retraining, against the *real* interior frames of the official held-out Vimeo-90K septuplet test split (7,732 of the 7,824 official clips are scored, the remainder dropped for missing or unreadable frames). *Every* metric degrades monotonically from the boundaries to the midpoint (edge→mid: PSNR -1.10 dB, LPIPS $+0.094$, MS-SSIM -0.090, DreamSim $+0.113$). Table 13 and Fig. 6 give the full per-position fidelity with every full-reference metric getting worst at the midpoint. We note that no pretrained-prior baseline is evaluated under this protocol: DynamiCrafter, SEINE, and TVG are all conditioned on single-endpoint image prompts and generating them under the Vimeo-90K two-anchor septuplet protocol (five interior frames between fixed boundaries) would require adapting their conditioning to a held-out-test distribution. This adaptation would conflate model quality with protocol compliance and we therefore report STRIDE's absolute numbers as evidence of the U-curve shape rather than a comparative fidelity claim.

Table 5: Performance metrics of STRIDE compared to state-of-the-art transition/interpolation baselines. Lower is better for inference time, GFLOPs, and model size. **Best** value per column in bold. *GFLOPs/fwd* is a single U-Net forward at each method's native resolution; *GFLOPs/clip* multiplies by the number of forwards per 16-frame clip (STRIDE 200 = 4 per DDIM step over 50 steps; DynamiCrafter and TVG 100 = 2 per step over 50; SEINE 500 = 2 per step over 250) and is the compute figure comparable across rows. Even accounting for its $4\times$ bidirectional-CFG overhead, STRIDE's per-clip compute is $\approx$ 36–181$\times$ below the 768^2 baselines; STRIDE is $3.7\times$ faster than the parameter-matched Wan2.1-VACE-1.3B (40 s/clip).

Model	**Eval. Res.**[a] (px)	**Inf. Time**[b] (s/video)	**GFLOPs**[c] (/fwd.)	**GFLOPs**[d] (/clip)	**GFLOPs**[h] (/px·clip)	**Params** (M)	**Size**[e] (GiB)
STRIDE (ours)	256^2	**11**	**384**	**76,800**	**1.2**	1536	2.86
DynamiCrafter	768^2	74	51,167	5,116,700	8.7	1439	2.68
TVG[f,i]	768^2	471	27,948	2,794,800	4.7	1439	2.68
SEINE	784^2	281	27,810	13,905,000	22.6	**909**	**1.68**
Wan2.1-VACE-1.3B[g]	624^2	40	–	–	–	1300	–
SVD-KFI[j]	512^2	540	–	–	–	$\sim$1500	–
Framer[k]	512^2	59	–	–	–	$\sim$2230	–

[a] Output resolution at which each model is run here; STRIDE operates in a 256^2 pixel space (32^2 latent). DynamiCrafter/TVG (native 320×512) and SEINE (native 320×512) are run at the larger square resolutions shown, i.e. above their release resolutions (section 4.1).

[b] Wall-clock seconds to generate one 16-frame clip on a single NVIDIA A100-SXM4-40GB GPU, averaged over the benchmark clips after a warm-up run, at each method's default sampler and step count: STRIDE/DynamiCrafter/TVG use 50 DDIM steps, SEINE 250 DDPM steps.

[c] FLOPs for a single U-Net forward at the model's native resolution. *Not* directly comparable across rows because native resolution differs; STRIDE's low count reflects its 256^2 working space, not solely a smaller network (it has the most parameters).

[d] Per-16-frame-clip FLOPs = GFLOPs/fwd $\times$ forwards/clip (see caption); this is the compute figure that accounts for STRIDE's larger sampling schedule.

[e] Parameter count and on-disk model size reported in fp16 (GiB, 1024^3 bytes).

[f] TVG inference time excludes its optional LPIPS-based per-frame selection pass (section 4.1).

[g] Wan2.1-VACE-1.3B (Wan Team, 2025) uses a pretrained video prior; GFLOPs not computed (proprietary attention pattern). Native resolution varies by benchmark: 624^2 for MorphBench/TransitBench, 464×832 for TC-Bench.

[h] Per-pixel-per-clip compute = GFLOPs/clip $\div$ native output pixels (res^2). Normalizes for the spatial-resolution difference across methods; STRIDE's 4–19$\times$ advantage over the 768^2/784^2 baselines on this column is the resolution-fair compute comparison.

[i] TVG's GFLOPs/fwd (27,948) is lower than DynamiCrafter's (51,167) despite sharing the same 3D U-Net backbone at identical 768^2 resolution. The discrepancy arises from profiling scope: DynamiCrafter's forward is profiled with its full text-conditioned CLIP cross-attention path active (we supply the fixed caption "smooth transition"); TVG's GP-based sampling bypasses text conditioning, so its forward is profiled without that cross-attention branch. Both figures are from `thop` profiling on the same GPU.

[j] SVD-KFI (Wang et al., 2025b) (Generative Inbetweening) is based on SVD-XT ($\sim$1.5B params); GFLOPs not computed. The 512^2 output is resized from the model's 1024×576 native resolution; runtime is $\approx$9 min/clip on a single A100-40GB. See section 4.1 for the two-anchor protocol adaptation.

[k] Framer (Wang et al., 2025a) uses a ControlNet-augmented SVD-XT backbone ($\sim$1.53B UNet + $\sim$0.70B ControlNet $\approx$ $\sim$2.23B params, estimated from released fp32 checkpoint sizes). GFLOPs not computed (dual UNet+ControlNet forward per step). Generates natively at 1024×576, resized back to each clip's native input resolution as with SVD-KFI so $\approx$59 s/clip pooled mean across all three benchmarks on a single A100-40GB.

A.8 Four-Phase Training Schedule

Table 14 gives the per-epoch per-dataset clip counts across the four-phase curriculum of section 3.7.

A.9 Sampler-Selection Ablation

Section 4.4 shows STRIDE's AF configuration comes best on smoothness/average, ties on ghosting/commitment) while table 15 gives the full per-dimension ratings behind it, over four inference configurations of the same frozen mode. The coupled-mean baseline (see eq. (2)), annealed fusion (AF, see eq. (4)), a stochastic DDIM variant (SDE), and AF+SDE, each rated under the identical 3-rater protocol and dimensions as the main comparative study. The per-dimension breakdown clarifies *why* AF is preferred over the stochastic variants: SDE and AF+SDE lose specifically on ghosting and commitment, not on smoothness, so their added stochasticity is not compensating for anything AF lacks.

Table 6: Quality metrics (mean, standard deviation in brackets) on the *MorphBench* benchmark. ↑ / ↓ indicate that higher / lower is better. **DC** = DynamiCrafter. VSI and FSIM are reported as $1 -$ similarity, so lower is better. Best value per row in **bold**.

Metric	Dir.	STRIDE	SEINE	TVG	DC	Wan	SVD-KFI	Framer
CLIPSim	↑	0.964(0.012)	0.95(0.03)	0.97(0.018)	0.96(0.02)	0.972(0.022)	**0.984(0.011)**	0.978(0.016)
VSI	↓	**0.018(0.006)**	0.054(0.03)	0.04(0.02)	0.077(0.04)	0.042(0.026)	0.034(0.023)	0.033(0.025)
FSIM	↓	**0.065(0.022)**	0.15(0.09)	0.122(0.06)	0.2(0.09)	0.127(0.071)	0.103(0.070)	0.100(0.067)
tPSNR	↑	**29.10(2.48)**	24.5(6.5)	23.5(4.53)	20(6.35)	25.148(5.370)	27.03(7.09)	27.40(6.22)
tMS-SSIM	↑	**0.919(0.025)**	0.65(0.26)	0.81(0.11)	0.56(0.22)	0.773(0.150)	0.858(0.105)	0.848(0.127)
EMD	↓	0.276(0.101)	0.254(0.21)	0.23(0.14)	0.24(0.149)	0.209(0.169)	**0.141(0.103)**	0.177(0.118)
DreamSim	↓	0.044(0.012)	0.08(0.06)	0.03(0.02)	0.058(0.03)	0.043(0.039)	**0.023(0.018)**	0.032(0.023)
LPIPS	↓	0.198(0.033)	0.29(0.15)	0.19(0.07)	0.33(0.13)	0.209(0.100)	**0.144(0.081)**	0.152(0.088)

Table 7: Quality metrics on the *TC-Bench* benchmark (mean, standard deviation in brackets). ↑ / ↓ indicate that higher / lower is better. **DC** = DynamiCrafter. VSI and FSIM are reported as $1 -$ similarity, so lower is better. Best value per row in **bold**.

Metric	Dir.	STRIDE	SEINE	TVG	DC	Wan	SVD-KFI	Framer
CLIPSim	↑	0.949(0.016)	0.96(0.03)	0.98(0.011)	0.97(0.02)	0.981(0.014)	**0.985(0.012)**	0.985(0.012)
VSI	↓	**0.018(0.010)**	0.052(0.052)	0.04(0.03)	0.055(0.053)	0.029(0.028)	0.025(0.029)	0.023(0.027)
FSIM	↓	**0.075(0.031)**	0.13(0.11)	0.11(0.09)	0.14(0.11)	0.084(0.072)	0.080(0.082)	0.076(0.079)
tPSNR	↑	31.12(3.92)	25.9(8.6)	26.55(7.43)	25.4(9.4)	28.335(6.040)	30.90(10.08)	**31.85(9.49)**
tMS-SSIM	↑	**0.925(0.036)**	0.72(0.28)	0.84(0.14)	0.72(0.24)	0.876(0.133)	0.910(0.113)	0.912(0.115)
EMD	↓	0.206(0.055)	0.24(0.41)	0.13(0.1)	0.133(0.135)	0.077(0.068)	**0.069(0.052)**	0.080(0.084)
DreamSim	↓	0.056(0.017)	0.07(0.08)	0.024(0.01)	0.047(0.038)	0.025(0.022)	**0.015(0.014)**	0.017(0.019)
LPIPS	↓	0.183(0.045)	0.23(0.16)	0.16(0.08)	0.232(0.14)	0.129(0.087)	**0.097(0.081)**	0.098(0.086)

A.10 Fine-Tuning Ablation Ladder

Section 5 narrates and interprets each stage of the ladder; Table 16 gives the full numeric results, and this section specifies only the evaluation protocol and proxies behind it. Each stage warm-starts from the trained model (EMA decay 0.99, one variable per stage) and is evaluated on a fixed 30-clip probe set against three paired proxies: 1) the detrended background-flicker index FLICK (the temporal standard deviation of the static background after removing a per-pixel linear luminance trend, so that legitimate global transitions are not miscounted as flicker; lower is better) 2) the CLIP-space smoothness roughness CLIP-RS $= \text{mean}\,|\Delta^2 \cos(\mathbf{f}_t, \mathbf{f}_{t+1})|$ (lower is smoother) 3) and the Ghost Index of appendix A.12 (endpoint-filtered at 0.05, 23 of 30 clips; lower is more committed). Deltas are paired medians against each stage's warm-start reference (A vs. the base model; B and C1 vs. A), with p-values from the Wilcoxon signed-rank test on per-clip differences.

Boundary transition jitter. All 48 TVS-48 renders show a second, localized artifact independent of the U-curve. The frame-to-frame pixel change at the anchor→first-generated and last-generated→anchor transitions is 3 to 7× higher than in the smooth interior. The cause is the hard pixel-restoration of the input anchor frames after each decode step. So generated frame 1 need not match the real frame 0 in high-frequency texture, leaving a visible jolt at the sequence endpoints. None of the fine-tune stages above targets this transition; annealed fusion or a boundary-conditioned bridge prior are natural future directions.

Table 17 gives the full per-arm results for the six-arm loss ablation ladder mentioned in section 5. Each arm removes one component from the Stage-A EMA canonical checkpoint and is evaluated on 45 DAVIS gt16 scenes (n=45) against the same control using paired Wilcoxon signed-rank tests with Holm–Bonferroni correction across arms. Our **Primary gate** is U-curve depth (`u_depth`) on four metrics (LPIPS, PSNR, MS-SSIM, DreamSim) with higher = worse ghosting. Further our **Secondary gate** is overall fidelity on the same four metrics. A run passes the gate only if it improves $\geq 3/4$ u_depth metrics with no fidelity regression. Table 18 gives the full gradient-budget breakdown for the gradient analysis referenced in section 5. Gradient norms $\|\nabla_\theta \mathcal{L}_k\|$ are computed per loss term k on a fixed 111-step diagnostic window (flow-warp and CLIP-RS at their respective gated cadences: 22 and 11 steps), using the canonical Stage-A EMA checkpoint.

A.11 Comparative Human Study: Full Statistics

This section reports the complete analysis behind section 4.5 i.e., inter-rater reliability and the Friedman omnibus per dimension (see table 19), the descriptive median [IQR] and Friedman average ranks (see table 20), the Holm-corrected pairwise Wilcoxon tests of STRIDE against each baseline

Table 8: Quality metrics on the *TransitBench* benchmark (mean, standard deviation in brackets). ↑ / ↓ indicate that higher / lower is better. **DC** = DynamiCrafter. VSI and FSIM are reported as 1 − similarity, so lower is better. Best value per row in **bold**.

Metric	Dir.	STRIDE	SEINE	TVG	DC	Wan	SVD-KFI	Framer
CLIPSim	↑	0.953(0.013)	0.93(0.027)	0.96(0.01)	0.94(0.019)	0.948(0.020)	**0.976(0.016)**	0.967(0.018)
VSI	↓	**0.030(0.007)**	0.1(0.05)	0.05(0.02)	0.13(0.06)	0.075(0.039)	0.042(0.033)	0.050(0.034)
FSIM	↓	**0.095(0.021)**	0.24(0.09)	0.12(0.05)	0.28(0.12)	0.186(0.093)	0.108(0.082)	0.125(0.071)
tPSNR	↑	25.87(1.77)	20.16(5.98)	21.6(3.4)	16.6(6.2)	21.879(5.835)	**27.52(8.51)**	24.91(5.63)
tMS-SSIM	↑	**0.881(0.030)**	0.47(0.22)	0.82(0.08)	0.42(0.25)	0.654(0.193)	0.846(0.117)	0.812(0.121)
EMD	↓	0.392(0.154)	0.4(0.21)	0.38(0.2)	0.39(0.19)	0.460(0.293)	**0.233(0.153)**	0.349(0.191)
DreamSim	↓	0.066(0.017)	0.11(0.05)	0.05(0.02)	0.087(0.03)	0.097(0.041)	**0.036(0.025)**	0.055(0.029)
LPIPS	↓	0.229(0.027)	0.388(0.13)	0.18(0.05)	0.4(0.15)	0.276(0.127)	**0.144(0.087)**	0.169(0.076)

Table 9: Quality metrics on *MorphBench*, all methods re-scored at a common measurement resolution (mean, standard deviation in brackets). ↑ / ↓ indicate that higher / lower is better. **STRIDE** is ours, **DC** = DynamiCrafter. VSI and FSIM are reported as 1 − similarity, so lower is better. Best value per row *within each resolution block* in **bold** with tied best values bold for all tied methods.

		256 × 256				512 × 512			
Metric	Dir.	STRIDE	SEINE	TVG	DC	STRIDE	SEINE	TVG	DC
CLIPSim	↑	0.96(0.011)	0.95(0.03)	**0.97(0.018)**	0.96(0.02)	0.96(0.011)	0.95(0.03)	**0.97(0.017)**	0.96(0.02)
VSI	↓	**0.02(0.006)**	0.054(0.03)	0.04(0.02)	0.07(0.04)	**0.018(0.006)**	0.054(0.03)	0.04(0.02)	0.07(0.04)
FSIM	↓	**0.068(0.02)**	0.15(0.09)	0.12(0.066)	0.2(0.09)	**0.068(0.02)**	0.15(0.09)	0.12(0.066)	0.2(0.09)
tPSNR	↑	**29.15(2.25)**	25(6.98)	24.04(4.73)	20.67(6.73)	**28.78(2.28)**	24.71(6.75)	23.5(4.56)	20.36(6.54)
tMS-SSIM	↑	**0.94(0.014)**	0.69(0.25)	0.85(0.1)	0.6(0.21)	**0.91(0.024)**	0.66(0.26)	0.81(0.11)	0.56(0.22)
EMD	↓	0.27(0.1)	0.258(0.21)	**0.237(0.14)**	0.243(0.148)	0.27(0.1)	0.258(0.21)	**0.23(0.14)**	0.24(0.14)
DreamSim	↓	0.045(0.011)	0.08(0.06)	**0.03(0.02)**	0.058(0.033)	0.045(0.011)	0.08(0.06)	**0.03(0.02)**	0.058(0.033)
LPIPS	↓	**0.14(0.024)**	0.23(0.16)	**0.14(0.07)**	0.26(0.128)	0.2(0.028)	0.27(0.16)	**0.19(0.08)**	0.311(0.13)

Table 10: Quality metrics on *TC-Bench*, all methods re-scored at a common measurement resolution (mean, standard deviation in brackets). ↑ / ↓ indicate that higher / lower is better. **STRIDE** is ours, **DC** = DynamiCrafter. VSI and FSIM are reported as 1 − similarity, so lower is better. Best value per row *within each resolution block* in **bold** with tied best values bold for all tied methods.

		256 × 256				512 × 512			
Metric	Dir.	STRIDE	SEINE	TVG	DC	STRIDE	SEINE	TVG	DC
CLIPSim	↑	0.95(0.01)	0.96(0.035)	**0.98(0.011)**	0.97(0.02)	0.95(0.01)	0.96(0.035)	**0.98(0.011)**	0.97(0.02)
VSI	↓	**0.017(0.008)**	0.05(0.05)	0.04(0.035)	0.05(0.05)	**0.017(0.008)**	0.05(0.05)	0.04(0.035)	0.05(0.05)
FSIM	↓	**0.06(0.022)**	0.13(0.11)	0.11(0.09)	0.137(0.11)	**0.06(0.023)**	0.13(0.11)	0.11(0.09)	0.14(0.11)
tPSNR	↑	**31.6(3.58)**	26.82(9.14)	26.9(7.5)	26.14(9.7)	**30.77(3.77)**	26.3(8.8)	26.55(7.43)	25.65(9.6)
tMS-SSIM	↑	**0.95(0.018)**	0.75(0.3)	0.88(0.11)	0.77(0.23)	**0.92(0.032)**	0.73(0.29)	0.84(0.14)	0.73(0.25)
EMD	↓	0.2(0.055)	0.24(0.41)	**0.128(0.11)**	0.13(0.13)	0.2(0.055)	0.24(0.41)	**0.128(0.1)**	0.13(0.13)
DreamSim	↓	0.055(0.015)	0.07(0.08)	**0.02(0.02)**	0.046(0.037)	0.056(0.01)	0.07(0.08)	**0.024(0.02)**	0.047(0.03)
LPIPS	↓	0.13(0.03)	0.18(0.17)	**0.11(0.07)**	0.166(0.129)	0.186(0.037)	0.21(0.17)	**0.16(0.08)**	0.2(0.14)

Table 11: Quality metrics on *TransitBench*, all methods re-scored at a common measurement resolution (mean, standard deviation in brackets). ↑ / ↓ indicate that higher / lower is better. **STRIDE** is ours, **DC** = DynamiCrafter. VSI and FSIM are reported as 1 − similarity, so lower is better. Best value per row *within each resolution block* in **bold** with tied best values bold for all tied methods.

		256 × 256				512 × 512			
Metric	Dir.	STRIDE	SEINE	TVG	DC	STRIDE	SEINE	TVG	DC
CLIPSim	↑	0.95(0.01)	0.93(0.02)	**0.96(0.01)**	0.94(0.02)	0.95(0.012)	0.93(0.02)	**0.96(0.01)**	0.94(0.02)
VSI	↓	**0.03(0.0076)**	0.1(0.04)	0.05(0.02)	0.13(0.06)	**0.03(0.007)**	0.1(0.04)	0.05(0.02)	0.13(0.06)
FSIM	↓	**0.09(0.02)**	0.24(0.09)	0.124(0.05)	0.28(0.12)	**0.097(0.02)**	0.24(0.09)	0.124(0.05)	0.28(0.12)
tPSNR	↑	**26.17(1.74)**	20.8(6.4)	22.8(3.88)	17.5(7.2)	**25.76(1.78)**	20.45(6.19)	21.6(3.4)	17.06(6.6)
tMS-SSIM	↑	**0.92(0.017)**	0.54(0.19)	0.85(0.08)	0.49(0.24)	**0.87(0.028)**	0.49(0.211)	0.82(0.08)	0.44(0.25)
EMD	↓	0.38(0.16)	0.4(0.21)	**0.375(0.2)**	0.38(0.19)	0.38(0.16)	0.4(0.21)	**0.375(0.2)**	0.38(0.19)
DreamSim	↓	0.066(0.016)	0.11(0.05)	**0.054(0.02)**	0.08(0.03)	0.066(0.01)	0.11(0.05)	**0.054(0.02)**	0.08(0.03)
LPIPS	↓	0.17(0.018)	0.33(0.12)	**0.14(0.048)**	0.33(0.13)	0.236(0.026)	0.37(0.13)	**0.18(0.05)**	0.37(0.15)

(see table 21), and the critical-difference diagrams for the three dimensions not shown in the main text (check fig. 8, 9, and 10).

Interior frames	FVD ↓
Linear crossfade (no VAE)	90.0
Crossfade → VAE round-trip	91.3
STRIDE (VAE + diffusion)	463.9

Table 12: Distribution-level fidelity on the held-out Vimeo-90K test split, computed with canonical I3D features (Unterthiner et al., 2018) on a fixed 988-clip subset (identical clips across all three rows; a subset is used because FVD needs only enough samples for a stable distribution estimate, not the full 7,732-clip fidelity set of Table 13). A linear crossfade is decomposed by pushing it through STRIDE's pipeline *minus* the diffusion (VAE encode-decode, anchors restored). The VAE round-trip adds only $+1.3$ FVD, so the gap to STRIDE is attributable to the generative process, not the frozen autoencoder.

Metric	Overall	Edge	Midpoint	Degr.
LPIPS ↓	0.330	0.291	0.386	0.094
PSNR ↑ (dB)	19.88	20.36	19.26	1.10
MS-SSIM ↑	0.728	0.772	0.682	0.090
DreamSim ↓	0.400	0.350	0.463	0.113

Table 13: Fidelity to *real* ground-truth interiors on the held-out Vimeo-90K test split (7,732 clips). *Edge* averages the boundary-adjacent interpolated positions ($\alpha \approx 0, 1$); *Midpoint* is $\alpha \approx 0.5$ (frame im4). *Degr.* is the magnitude of degradation from edge to midpoint (always in the worsening direction). It is positive for every metric, i.e. fidelity is consistently worst at the center.

Table 14: Four-phase joint training schedule (45 epochs). The Augmenter subset and DAVIS are always 100% included. Counts are ≈ per-epoch training-split sizes; "–" denotes a dataset not used in that phase. Phase 4 is a clean-data fine-tune that excludes the in-the-wild sources.

Phase	**Epochs**	**Vimeo-90K**	**UCF-101**	**Kinetics**	**LAVIB**	**WebVid**
1 — VFI Foundation	1–8	10k	2k	–	–	–
2 — Expansion	9–15	22k	5.3k	5k	4k	–
3 — Full Joint	16–35	100% (∼36k)	5.3k	12k	10k	7k
4 — Perceptual Polish	36–45	100% (∼36k)	5.3k	–	–	–

Table 15: Sampler-selection ablation on the TVS-48 benchmark (48 clips; distinct from the CTB-48 set of the main study) (mean rating, 1–5, higher = better). **Bold** marks the best configuration per dimension. AF is best on smoothness and average and ties the baseline (no significant difference) on ghosting and commitment; the stochastic variants (SDE, AF+SDE) degrade both. Full pipeline as in Appendix A.11.

Config	**Smooth.** ↑	**Ghost.** ↑	**Commit.** ↑	**Avg.** ↑
Baseline (mean)	3.48	**2.97**	**3.15**	3.20
AF (ours)	**4.08**	2.94	3.06	**3.36**
SDE	3.77	2.48	2.66	2.97
AF+SDE	3.85	2.44	2.65	2.98

A.12 Automatic Ghost Index: Negative Control

We test whether an automatic per-video *Ghost Index* (GI) can substitute for the human ghosting judgment. For each interior frame i we compute, in CLIP ViT-L/14 space, $\mathrm{GI}_i = 1 - |\cos(\mathbf{f}_i, \mathbf{f}_{\mathrm{start}}) - \cos(\mathbf{f}_i, \mathbf{f}_{\mathrm{end}})|$ and average over all interior frames (*lower* = less ghosting and a $50/50$ crossfade $\rightarrow 1$). The GI is a property of the generated video and hence rater-independent. Because it saturates near 1.0 when the two anchors are near-identical (endpoint distance $1 - \cos(\mathbf{f}_{\mathrm{start}}, \mathbf{f}_{\mathrm{end}})$ small), we correlate it with the human scores only on the 167 of 240 (clip, method) pairs whose endpoint distance is ≥ 0.05. A valid ghosting proxy would give Spearman $\rho < 0$ (Index lower = better vs. human higher = better). Instead every correlation is non-negative (see table 22). The Index is confounded by method since abrupt or held-frame baselines score *low* because their frames snap decisively to one endpoint, which the metric rewards even though raters judge those transitions worse. The automatic proxy therefore cannot rank architectures, motivating the subjective study.

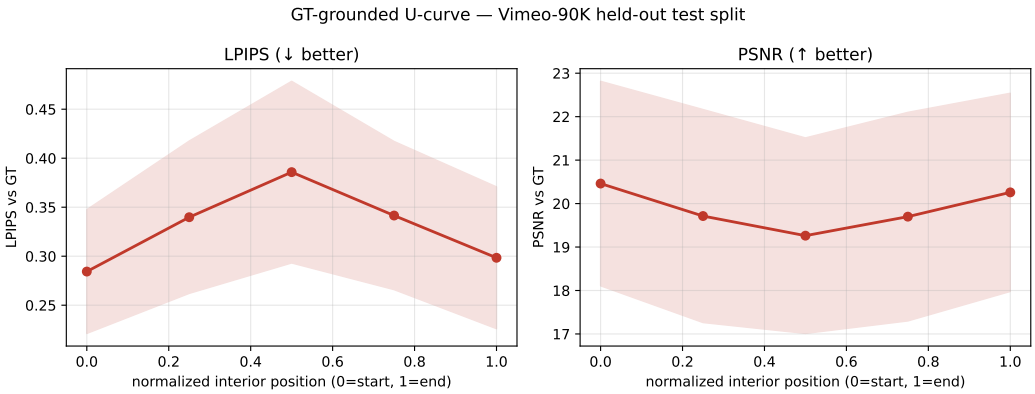


Figure 6: Per-frame fidelity to ground truth as a function of normalized interpolation position on the held-out Vimeo-90K test split (7,732 clips). All four metrics form a symmetric U (error rises toward the center / quality falls), worst at the midpoint $\alpha \approx 0.5$. This is the mean-regression behavior discussed in section 5, measured directly against real interior frames.

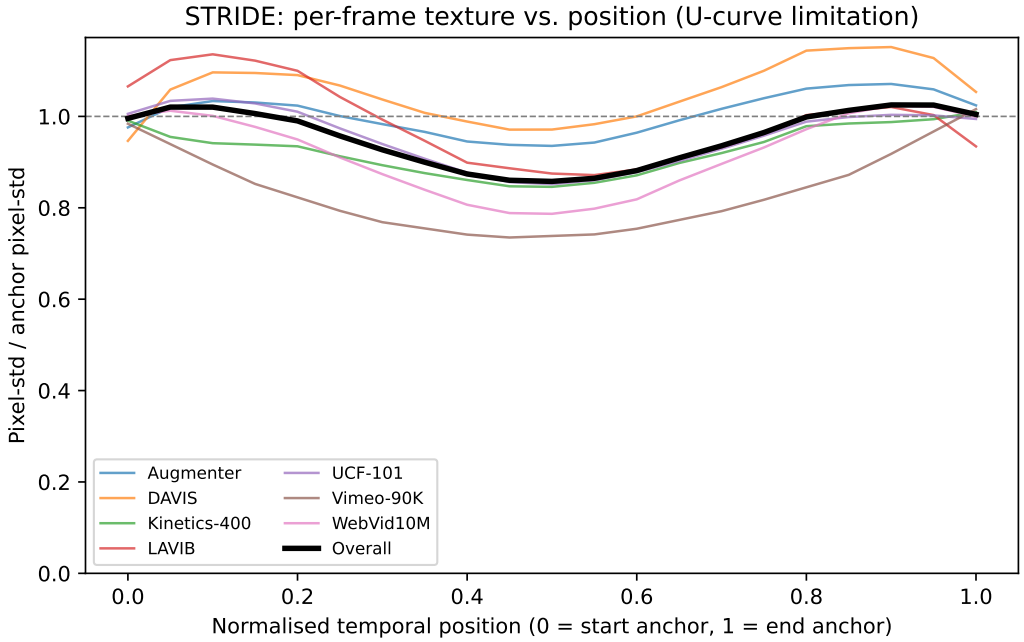


Figure 7: The U-curve. Per-frame pixel standard deviation, normalized by the mean anchor standard deviation and resampled onto a common temporal axis (0: start anchor, 1: end anchor), averaged per benchmark. A value of 1 indicates anchor-level texture; the characteristic dip toward the center ($\alpha \approx 0.5$) quantifies the mean-regression wherein the generated mid-frames carry measurably less high-frequency detail than the anchors. The dip is consistent across datasets, identifying it as a systematic property of the model rather than a content-specific artifact.

A.13 Human-Study Protocol Details

Raters and consent. The study used three raters (excluding authors) who evaluated all 240 videos independently, with no communication during rating and no knowledge of which method produced any given video (method identity was withheld and the on-screen ordering randomized per rater). Their identities are omitted here to preserve double-blind review and will be acknowledged in the camera-ready version. All three participated voluntarily and gave informed consent for the use of their anonymized ordinal ratings in this research. Before beginning they received a written descrip-

Stage	Lever	FLICK ↓	CLIP-RS ↓	GI ↓	Outcome
A	static reweight + RGB sup.	-9%	held	–	**kept**
B	flow-adaptive warp loss	-4.5%	$+6.2\%$ (ns)	–	rejected
C1	relax fwd/bwd coupling	$+9.5\%$	ns	-0.004	rejected

Table 16: Fine-tuning ladder on the frozen model (paired medians vs. warm-start reference; Wilcoxon signed-rank on per-clip differences, n=30; GI at n=23 after the endpoint filter). Stage A reduces flicker at no smoothness cost and is retained. Stage B gives only a marginal flicker reduction with no significant smoothness change ($+6.2\%$, $17/30$ clips rougher, p=0.45, concentrated on moving-subject clips). Stage C1 leaves ghosting statistically unchanged (ΔGI median -0.004, p=0.20) while significantly worsening flicker ($p=8\times10^{-4}$, smoothness p=0.15). No loss-level intervention moves the ghost, consistent with the fused-noise variance analysis of section 3.5. A separate six-arm loss ablation (see section 5) found that removing Min-SNR-γ (γ=5) is the only change that reduces U-curve depth on $3/4$ metrics ($p<0.001$) without regressing mean fidelity.

Table 17: Six-arm loss ablation ladder (n=45 DAVIS gt16 scenes, paired Wilcoxon+Holm). `u_depth` = edge-to-midpoint degradation (higher is worse ghosting). ↓/↑ with *** / ** / * = Holm-corrected $p<0.001$ / 0.01 / 0.05; ns = not significant. **Gate**: arm passes only if $\geq$ 3 u_depth metrics improve with zero fidelity regression (all-ns on overall fidelity). Only `no_minsnr` passes.

Arm	**Ablated component**	**u_depth direction (ghosting gate)**				**Gate**
		LPIPS	PSNR	MS-SSIM	DreamSim	
`no_minsnr`	Remove Min-SNR-γ (γ=5)	↓***	ns	↓***	↓***	**PASS**
`no_clip_rs`	Remove CLIP-RS train loss	ns	ns	↑*	ns	FAIL
`no_vgg`	Remove VGG perceptual loss	↓***	↑***	↑***	ns	FAIL
`no_perceptual`	Remove full perceptual block	↓ns	↓***	ns	↑**	FAIL
`no_flow`	Remove flow-warp loss	ns	ns	↑**	↑*	FAIL
`no_anchor`	Remove $\mathcal{L}_{\text{anchor}}$ loss	ns	ns	ns	ns	FAIL (flat)

Table 18: Gradient-budget analysis. Gradient norm per loss term (median over diagnostic steps) and its share of the total update budget. `clip_rs` dominates at 41.8% of the update norm; $\mathcal{L}_{\text{cons}}$ is decorative (0.1%). Shares do not add to 100% because terms fire at different cadences.

Term	**n steps**	**Median $\|\nabla_\theta \mathcal{L}\|$**	**Budget share**
$\mathcal{L}_{\text{diff}}$	111	4.0×10^{-1}	6.9%
$\mathcal{L}_{\text{cons}}$	111	6.9×10^{-3}	0.1%
$\mathcal{L}_{\text{anchor}}$	111	3.0×10^{-1}	5.1%
Charbonnier (pixel)	111	5.9×10^{-1}	10.0%
VGG	111	7.7×10^{-1}	13.1%
DISTS	111	7.1×10^{-1}	12.0%
Flow-warp	22	6.4×10^{-1}	11.0%
CLIP-RS train	11	2.5×10^{0}	41.8%

tion of the task and the rating rubric, were told they could stop at any time, and no personally identifying information was collected. The task presented only short generated video clips and carried no risk beyond ordinary screen viewing.

Interface and procedure. Each rater viewed one video at a time in a fixed-size player supporting replay, and entered three 1 to 5 Likert scores before advancing. The 240 videos were presented in a randomized order per rater. Ratings were recorded in a spreadsheet with one row per (clip, method) and a 1 to 5 drop-down per dimension.

Rubric wording. Raters were given the following definitions (higher = better): 1) **Smoothness** — "the transition plays as continuous motion without flicker, stutter, or abrupt jumps", 2) **Ghosting** — "interior frames are free of superimposed / double-exposure / crossfaded content" (a translucent blend of both endpoints scores low), and 3) **Semantic Commitment** — "each interior frame depicts a single coherent intermediate scene rather than an ambiguous mixture of the two endpoints". The per-clip average is computed post-hoc as the mean of the three and is not rated directly.

Table 19: Inter-rater reliability and Friedman omnibus statistics per dimension ($N = 48$ clips, $k = 5$ methods, 3 raters). The **Kendall** W column is the *inter-rater* concordance (rater agreement), *not* the Friedman effect size; the omnibus effect size is $W_{\text{eff}} = \chi^2/[N(k-1)]$ (0.38, 0.12, 0.15, 0.09 for Smoothness, Ghosting, Sem. Commit., Average respectively).

Dimension	**Kripp.** α	**ICC**	**Kendall** W	**Friedman** χ^2	p
Smoothness	0.836	0.938	0.840	73.12	5×10^{-15}
Ghosting	0.866	0.949	0.854	22.78	1×10^{-4}
Sem. Commit.	0.861	0.949	0.833	29.04	8×10^{-6}
Average	0.865	0.944	0.899	16.93	2×10^{-3}

Table 20: Human-study descriptives: median [IQR] of the 1–5 ratings and Friedman average rank (1=best, computed as $(k+1)-\bar{r}_{\text{Friedman}}$ where k=5). **Bold** marks the best (lowest) average rank per dimension.

Dimension	**Method**	**Median [IQR]**	**Avg. rank** $\downarrow$
Smoothness	STRIDE (AF)	4.0 [4.0, 4.0]	**1.53**
	Wan2.1-VACE-1.3B	3.0 [2.0, 4.0]	2.71
	TVG	2.0 [1.0, 3.0]	3.19
	DynamiCrafter	2.0 [1.0, 3.0]	3.71
	SEINE	2.0 [1.0, 3.0]	3.87
Ghosting	SEINE	3.0 [2.0, 4.0]	**2.53**
	Wan2.1-VACE-1.3B	3.0 [2.0, 4.0]	2.80
	TVG	3.0 [2.0, 4.0]	2.84
	DynamiCrafter	3.0 [2.0, 4.0]	2.95
	STRIDE (AF)	2.0 [2.0, 3.0]	3.88
Sem. Commit.	Wan2.1-VACE-1.3B	3.0 [2.0, 4.0]	**2.02**
	DynamiCrafter	2.0 [1.0, 3.0]	3.06
	TVG	2.0 [2.0, 3.0]	3.09
	STRIDE (AF)	2.0 [2.0, 3.0]	3.31
	SEINE	2.0 [1.0, 3.0]	3.51
Average	Wan2.1-VACE-1.3B	3.00 [2.33, 3.67]	**2.42**
	STRIDE (AF)	2.67 [2.67, 3.00]	2.65
	TVG	2.33 [1.92, 3.33]	3.05
	SEINE	2.33 [1.67, 3.08]	3.43
	DynamiCrafter	2.33 [1.67, 3.00]	3.46

CTB-48 pool composition. The main-text Design paragraph (see section 4.5) gives the CTB-48 category breakdown (15 clips each for object morph, camera motion, and scene change, plus 3 small-appearance controls). Here we specify only how that stratification was carried out. Sampling used even difficulty quartiles within each category, so the pool spans the easy-to-hard range rather than over-weighting the hardest clips, with the small-appearance-change category serving as an easy control (near-identical endpoints where all methods should succeed). Difficulty within each category is estimated from endpoint CLIP/LPIPS/DreamSim distances and camera-compensated flow statistics.

A.14 Reproducibility Statement

This section consolidates the optimizer- and schedule-level detail underlying the summary in the main-text Reproducibility Statement. Training (see section 3.6) uses `AdamW8bit` (Dettmers et al., 2021) (weight decay 0.01) under a cosine schedule (1000-step warm-up, $10^{-5} \rightarrow 10^{-4}$, $\eta_{\min}$=10^{-5}), warm-restarted at each phase boundary with factors $[1.0, 0.3, 0.3, 0.5]$, gradient clipping at 1.0, and `bfloat16` mixed precision via `Hugging Face Accelerate`. The held-out fidelity and FVD evaluations use the official Vimeo-90K `sep_testlist` split. Upon acceptance, we will release the model weights and the complete training, inference, evaluation, and human-study-analysis code to enable exact reproduction.

Table 21: Pairwise Holm-corrected Wilcoxon signed-rank tests of STRIDE (AF) against each baseline, with matched-pairs rank-biserial r (sign relative to STRIDE: $r>0$ favors STRIDE). STRIDE is significantly best on smoothness over all four baselines and significantly worst on ghosting; Wan is significantly best on semantic commitment over all four others; no average pair involving STRIDE survives Holm correction.

Dimension	**STRIDE vs.**	p_{Holm}	r
Smoothness	DynamiCrafter	1.3×10^{-7}	$+0.97$
	SEINE	5.1×10^{-8}	$+0.99$
	TVG	1.1×10^{-7}	$+0.98$
	Wan2.1-VACE-1.3B	2.9×10^{-4}	$+0.74$
Ghosting	DynamiCrafter	4.9×10^{-3}	-0.62
	SEINE	2.5×10^{-4}	-0.74
	TVG	7.5×10^{-3}	-0.57
	Wan2.1-VACE-1.3B	1.7×10^{-3}	-0.71
Sem. Commit.	DynamiCrafter	1.00	-0.11
	SEINE	1.00	$+0.01$
	TVG	1.00	-0.18
	Wan2.1-VACE-1.3B	5.7×10^{-4}	-0.72
Average	DynamiCrafter	0.22	$+0.35$
	SEINE	0.14	$+0.38$
	TVG	1.00	$+0.20$
	Wan2.1-VACE-1.3B	1.00	-0.21

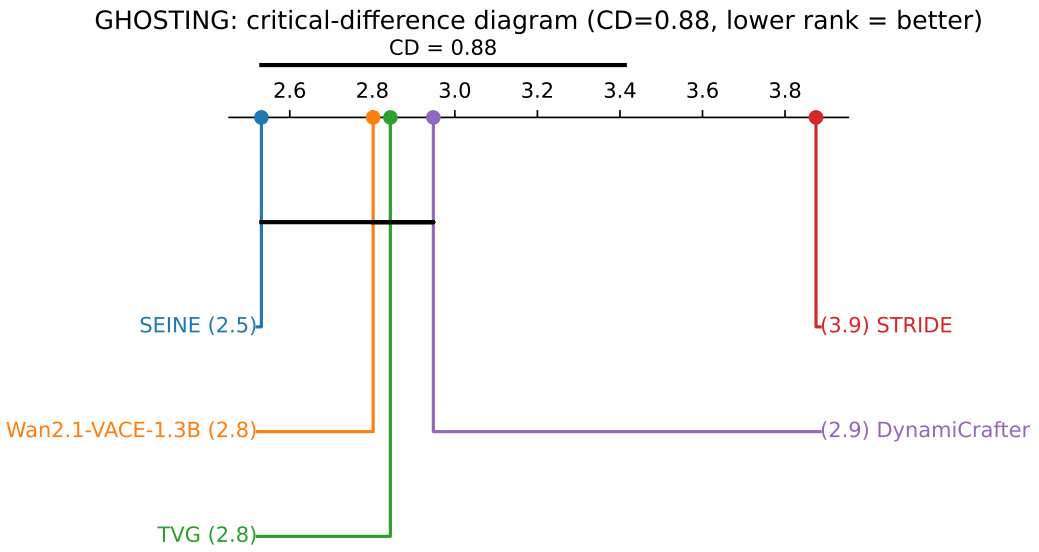


Figure 8: Critical-difference diagram for ghosting (lower average rank is better, $N=48$, five methods). STRIDE is significantly worst on ghosting against all four baselines; all four Holm-corrected pairwise p-values are ≤ 0.008 (see table 21).

Table 22: Spearman correlation of the automatic Ghost Index with the human scores across methods ($n=167$ (clip, method) pairs with distinguishable endpoints). A valid proxy needs $\rho < 0$; the observed signs are the wrong direction, so the Index does not transfer across architectures.

Human dimension	**Spearman** ρ	p
Ghosting	$+0.110$	0.207
Sem. Commitment	$+0.298$	5×10^{-4}
Average	$+0.331$	1×10^{-4}

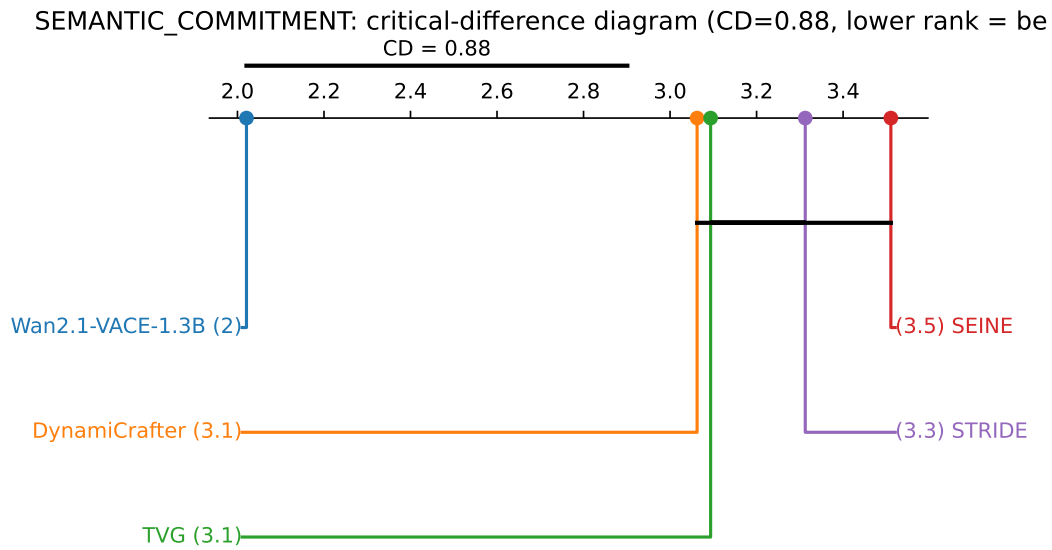


Figure 9: Critical-difference diagram for semantic commitment (lower average rank is better, N=48, five methods). Wan2.1-VACE-1.3B is significantly best, separated from all four others (Holm-corrected $p<0.001$); the remaining four are not significantly different after Holm correction.

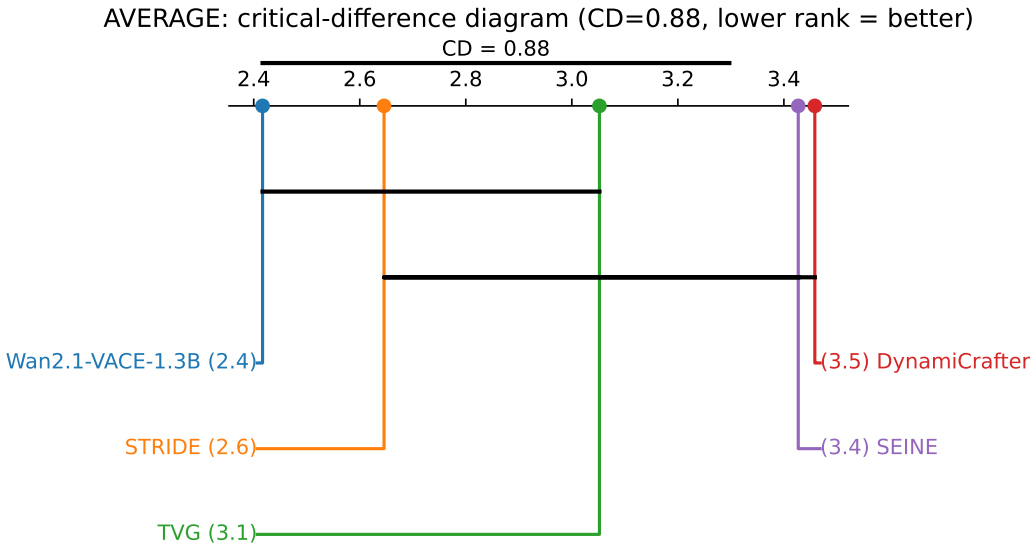


Figure 10: Critical-difference diagram for the per-clip average (lower average rank is better, N=48, five methods). Wan2.1-VACE-1.3B and STRIDE (AF) rank first and second; DC, SEINE, and TVG are significantly below Wan on average (Holm-corrected $p\leq 0.024$), but STRIDE vs. DC/SEINE/TVG and all other pairs are not significant after Holm correction.