


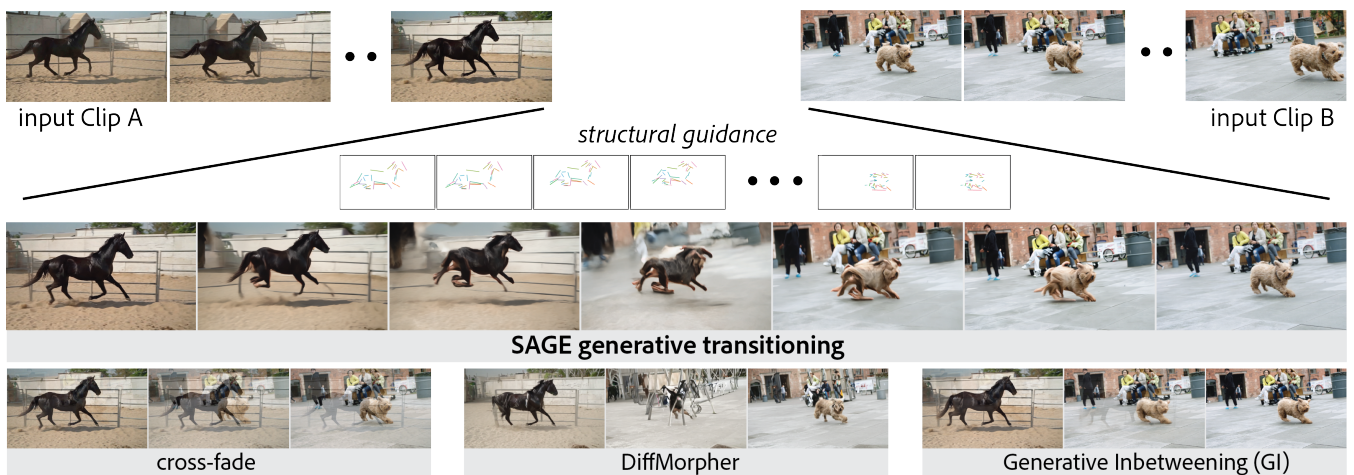


# SAGE: Structure-Aware Generative Video Transitions between Diverse Clips

Mia Kan<sup>1</sup>  Yilin Liu<sup>1,2</sup>  Niloy J. Mitra  <sup>1,3</sup><sup>1</sup>University College London <sup>2</sup>Autodesk Research <sup>3</sup>Adobe Research

**Figure 1:** SAGE: Structure-Aware Generative video transitions. Given two diverse clips (top),  $C_A$  and  $C_B$ , prior approaches (bottom) (e.g., cross-fade, DiffMorpher, Generative Inbetweening) often suffer from ghosting, structural collapse, or flicker. We introduce SAGE that extracts structural lines and motion cues, propagates them via B-spline trajectories to produce structural guidance. The guidance is then used to condition a pretrained diffusion model to synthesize temporally smooth with motion-coherent transitions (middle) in a zero-shot setting.

## Abstract

Video transitions aim to synthesize intermediate frames between two clips, but naïve approaches such as linear blending introduce artifacts that limit professional use or break temporal coherence. Traditional techniques (cross-fades, morphing, frame interpolation) and recent generative inbetweening methods can produce high-quality plausible intermediates, but they struggle with bridging diverse clips involving large temporal gaps or significant semantic differences, leaving a gap for content-aware and visually coherent transitions. We address this challenge by drawing on artistic workflows, distilling strategies such as aligning silhouettes and interpolating salient features to preserve structure and perceptual continuity. Building on these strategies, we propose SAGE (Structure-Aware Generative video transitions) as a simple yet effective zeroshot approach that combines structural guidance, provided via line maps and motion flow, with generative synthesis, enabling smooth, motion-consistent transitions without fine-tuning. Extensive experiments and comparison with current alternatives, namely [RKT\*22, ZCL\*24, ZZK\*24, JHM\*25, ZRW\*25], demonstrate that SAGE outperforms both classical and the latest generative baselines on quantitative metrics and user studies for producing transitions between diverse clips. The simple method effectively bypasses the need to acquire suitable training data, which is particularly difficult in our creative setting involving diverse clips. Code is available via the project page at <https://kan32501.github.io/sage.github.io/>.

## CCS Concepts

• **Computing methodologies** → **Image manipulation; Computer vision; Machine learning;**

## 1. Introduction

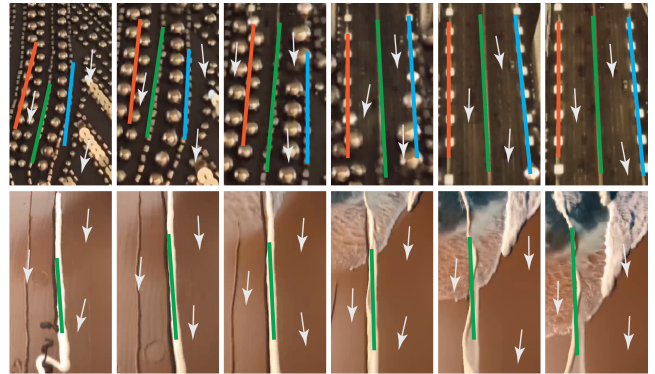
*Video transition* refers to the task of synthesizing intermediate frames to seamlessly bridge two video clips. Such transitions are essential in editing, storytelling, and generative media, enabling fluid scene changes without distracting the viewer. Naïve strategies, such as linear blending in pixel or latent space, often introduce flickering, ghosting, or spurious objects, breaking temporal coherence and making them unsuitable for professional workflows.

More advanced methods, including morphing [BN92, Wol98] and frame interpolation [RKT\*22, HZH\*22, BLM\*19], treat transitions as interpolation between the start and end frames only. Recent generative inbetweening methods [JGZ\*24, ZRW\*25, ZCZ\*25], designed specifically for video, demonstrate that modern video generation models, when trained with suitable data, can hallucinate plausible intermediates. However, these approaches assume small temporal gaps and closely aligned semantics, and they often fail when clips differ significantly in content and/or style – video pairs we refer to as *diverse clips*. Thus, current methods lack a way to generate transitions that are both *content-aware* and *visually coherent* across diverse scenarios, especially across significantly different clip pairs – a scenario that is engaging because of its creative possibilities.

We address this gap by targeting the challenging setting of diverse clips that differ in style, structure, and/or semantics. Skilled artists can craft compelling transitions in such cases, but rely on manual design tailored to specific pairs of clips (see Figure 2 and project page). Note that, as an additional challenge, we have very limited examples of such transitioning effects (e.g., in social media posts or creative content), and hence cannot finetune or retrain a generative model to directly produce such effects. Instead, we distill their heuristics into guiding strategies: first, detecting and aligning line features and silhouettes across clips, and second, interpolating salient features (e.g., feature lines and structural outlines) to anchor transitions in intermediate frames, as these provide smooth structural, semantic, and motion transitions.

We propose Structure-Aware Generative vidEo transitions (SAGE), a method that fuses *structural guidance* with generative synthesis. Given a pair of video clips  $\{C_A, C_B\}$ , we first extract line maps and optical flow for the final frame of clip  $C_A$  and the initial frame of clip  $C_B$ . From these, we detect, match, and interpolate to produce intermediate line structures that capture both geometry and motion cues. Specifically, we demonstrate that suitably encoding detected linear structures, using their centers and slopes, along with the detected motion flows, allows us to establish better quality matches as well as produce more fluid motion-aware structural guidance. The extracted control structures can then be used to condition a pretrained generative inbetweening model [ZRW\*25], producing temporally smooth and semantically consistent transitions, in a zero-shot fashion. Unlike prior methods, SAGE unifies geometric guidance with generative synthesis *without* requiring fine-tuning, and makes use of both object appearance and motion while directly leveraging pretrained generative models.

We assess SAGE across varied video transitions, benchmarking against classical interpolation (i.e., cross-fade) and state-of-the-art generative baselines (FILM [RKT\*22], SEINE [CWZ\*23],



**Figure 2: Artist-designed transitions.** Two artist-crafted transitions illustrate the heuristics that inspire SAGE; full sequences are provided on the project page. (i) Structural anchoring: silhouettes and edges are aligned across clips to prevent scene collapse, as highlighted by the matching colored lines. (ii) Motion continuity: dominant flows such as camera pans are preserved to ensure fluid evolution, as indicated by the white arrows. (iii) Layered blending: foreground objects morph while backgrounds fade, reducing ghosting and clutter (not depicted here). These principles motivate our design of structure- and motion-aware generative transitions.

DiffMorpher [ZZX\*24], TVG [ZCL\*24], Generative Inbetweening [ZRW\*25] and VACE [JHM\*25]). Both quantitative results and a user study show that our method produces smoother, more natural transitions in the context of diverse clips. See Figure 1 and the project page for some examples.

**Novelty and contributions.** While our method is simple and leverages a pretrained generative inbetweening, the novelty lies in how these components are orchestrated to address a previously unexplored problem: *content-aware video transitions across diverse clips in a zero-shot setting*. Unlike prior frame interpolation or generative inbetweening methods, which assume consistent semantics and small temporal gaps, we explicitly distill artist-inspired heuristics into a principled design. Specifically, we (i) introduce *hierarchical structural anchoring*, where salient line structures are extracted, normalized, and matched in a layerwise manner to avoid background dominance; (ii) propose *motion-aware B-spline propagation*, which couples local line evolution with global foreground trajectories, mitigating the trajectory crossings and incoherent motion seen in naive interpolation; and (iii) demonstrate that these structural priors can condition a pretrained diffusion-based inbetweening model to achieve smooth transitions without task-specific fine-tuning. This synthesis of structural guidance with generative synthesis is unique, enabling transitions that are both temporally coherent and semantically adaptive despite the absence of curated training data.

## 2. Related Works

**Traditional video transitions.** Video editing has long relied on handcrafted transitions such as cross-fades, wipes, and dissolves. These approaches, being procedural, are computationally simple

and widely supported in editing software, but they are insensitive to scene content and therefore limited in realism and adaptability. Beyond preauthored transition functions, morphing techniques [BN92, Wol98] represented an early attempt to interpolate structural correspondences between frames, but typically required manual keypoint alignment or feature specification, making them impractical for general-purpose video transitions. With the advent of generative models and access to suitable datasets, DiffMorpher [ZZX\*24] set a new state-of-the-art by proposing diffusion-based generative morphing between images, providing a foundation for content-aware blending (see Section 5 for a comparison), though it does not leverage motion information present in video clips.

**Video frame interpolation.** Frame interpolation methods aim to generate intermediate frames between temporally adjacent inputs. Classical approaches estimated optical flow to warp pixels, while modern deep networks learn to predict motion or directly synthesize frames. Representative works include DAIN [BLM\*19], which leverages depth-aware warping; RIFE [HZH\*22], which predicts intermediate optical flows in real-time; and FILM [RKT\*22], which targets large motion using multi-scale fusion with perceptual constraints (see recent survey [KRK\*25]). While these methods achieve high-quality results for short-term interpolation, they typically assume that input frames share the same scene and semantics, and consequently fail when applied to transitions between diverse clips that differ substantially in appearance, motion, or style. Extensions such as SEINE [CWZ\*23], which generates long videos with smooth and creative transitions between shot-level clips, and VACE [JHM\*25], which unifies video generation and editing tasks, broaden applicability but still operate primarily from keyframes and do not exploit motion cues for diverse clip-to-clip transitions. In Section 5 and the project page, we provide comparisons with FILM, SEINE, and VACE.

**Generative inbetweening.** Recent progress in diffusion-based generative models has enabled more powerful video synthesis. In particular, video diffusion models can effectively hallucinate plausible intermediate content beyond deterministic flow warping. For example, Jain et al. [JGZ\*24] propose a diffusion framework for video interpolation between frames; Zhou et al. [WZC\*25] adapt pretrained image-to-video diffusion models for keyframe interpolation using forward and backward temporal losses; Zhang et al. [ZCZ\*25] enhance diffusion-based interpolation under large motion; and Zhang et al. [ZCL\*24] propose TVG as a training-free approach that interpolates in latent space using Gaussian process regression and frequency-aware fusion. These methods achieve strong performance when trained on large curated datasets, but generally assume small temporal gaps and semantic consistency. However, they are not directly applicable to artistic or cross-domain video transitions across diverse clips. In Section 5, we present a comparison with a recent generative inbetweening approach [WZC\*25] as well as TVG.

**Concurrent work.** Recent methods explicitly target transitions across clips with larger visual differences. MatchDiffusion [PPZ\*24] defines a *joint* diffusion process to align coarse structure and motion for two different text prompts, followed by a *disjoint* diffusion phase that generates detail-preserving, match-cut-ready

video pairs. Although the results show broad structural coherence between the produced clips, they still require a manual blend step to bridge visual inconsistencies. VTG [YZY\*25] introduces a versatile diffusion-based framework<sup>†</sup> for transitions, employing bidirectional motion fine-tuning and representation alignment, and evaluates on a dedicated benchmark (TransitBench). While promising, the method lacks fine structural control, which can lead to inconsistent motion or structural collapse in challenging cases. In contrast, our method integrates structural guidance (line structures and motion guidance) with pretrained generative inbetweening, enabling smoother and more semantically consistent transitions without additional training.

### 3. Design Considerations

Artists and video creators often design video transitions manually, guided by heuristics that preserve perceptual continuity while allowing creative freedom. From examining such workflows from classical books [Ond00, Pea16], blog posts [Viv25, Mar25], and popular social media examples, we distilled three principles that inform the design of SAGE.

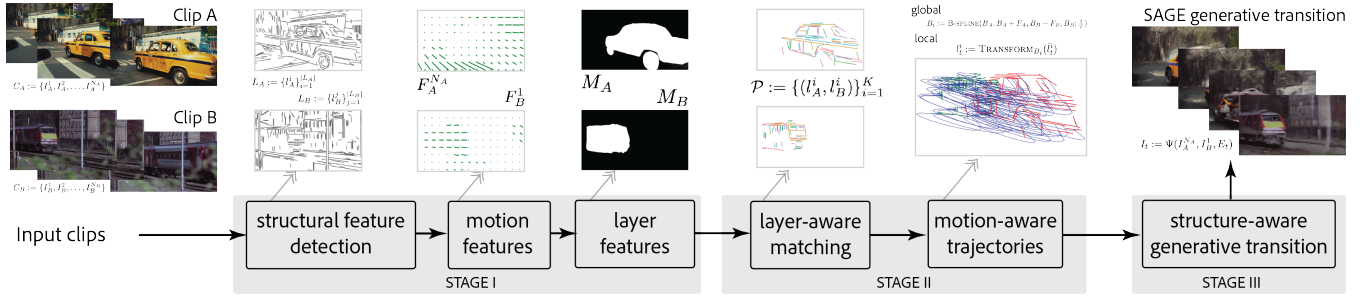
**(i) Structural anchoring.** Transitions are smoother when dominant structural cues, such as edges, silhouettes, or perspective lines, are preserved across clips. Artists often align or morph these structures, even when the content changes significantly, to avoid abrupt scene collapses. This motivates our use of line maps and their motion encoding to provide structural guidance.

**(ii) Motion continuity.** Smooth transitions guide viewers' attention through consistent motion. Artists typically match or extrapolate dominant motion trajectories (e.g., camera pans, object flows, vanishing directions), ensuring that transitions feel fluid rather than chaotic and avoiding unnecessary crossings. This observation motivates our use of optical flow to preserve motion direction and magnitude, and blending using smooth B-spline paths to guide the intermediate transitions.

**(iii) Layered blending.** Manual transitions often separate global and local changes. For example, we observed that artists gradually faded backgrounds while foreground objects are interpolated or morphed. Such layering avoids ghosting and reduces visual clutter. Inspired by this, we design SAGE to combine structural and motion guidance with generative synthesis, while making use of foreground/background information; thus, enabling both smooth global blending and coherent local transformations.

Following these principles, we design SAGE to extract structural lines and motion cues that anchor transitions across clips. By encoding these structures and using them to guide a pretrained generative inbetweening model, our approach operates in a zero-shot setting. This avoids the need for curated training data, which is scarce for artistic transitions across diverse clips.

<sup>†</sup> code was not publicly available at the time of publication.



**Figure 3: Method overview.** Given two clips, we extract structural lines, optical flow, and foreground masks (Stage I). We match and interpolate these structures using motion-aware B-spline trajectories (Stage II), producing intermediate line sets  $\{L_t\}_{t=1}^T$ . These are then used to condition a pretrained generative inbetweening model (Stage III), yielding smooth and motion-aware transitions between diverse clips.

## 4. Algorithm

Given a pair of diverse video clips  $C_A, C_B$ , we extract  $\{I_A^1, I_A^2, \dots, I_A^{N_A}\}$  frames from  $C_A$  and  $\{I_B^1, I_B^2, \dots, I_B^{N_B}\}$  frames from  $C_B$ . We aim to synthesize  $T$  inbetween frames  $\{I_t\}_{t=1}^T$  that are temporally smooth and semantically coherent across *diverse clips*. Guided by the design considerations in Section 3, SAGE proceeds in three stages: (i) *feature extraction* (Section 4.1); (ii) *motion-aware structural interpolation* (Section 4.2); and (iii) *conditional generative synthesis* (Section 4.3) (see Figure 3). Our key contributions are the introduction of *line-based structural extraction* to anchor transitions, and a novel *B-spline-guided propagation* scheme that couples local line interpolation with global motion trajectories.

### 4.1. Feature Extraction

We extract three complementary features from the boundary frames  $I_A^{N_A}$  and  $I_B^1$ :

(i) *Structural features.* We detect two sets of line segments on the last frame  $N_A$  of video clip  $C_A$  and the first frame of video clip  $C_B$ , where

$$L_A := \{I_A^i\}_{i=1}^{|L_A|}, \quad L_B := \{I_B^j\}_{j=1}^{|L_B|},$$

using a pretrained line detector (we use GlueStick [PSY\*23]). Each line  $l = \{x_1, y_1, x_2, y_2\}$  is encoded by its endpoints, collectively representing silhouettes and dominant contours.

(ii) *Motion features.* We estimate optical flow fields  $F_A^{N_A}$  and  $F_B^1$  using SEA-RAFT [WLD24]:

$$F_A^{N_A} := \phi(I_A^{N_A-k}, I_A^{N_A}), \quad F_B^1 := \phi(I_B^1, I_B^k),$$

with a small temporal span  $k$ , the number of offset frames used for optical flow computation. These capture local motion cues aligned with the structural features. Empirically, we set  $k = 3$  in our tests.

(iii) *Layer features.* Foreground masks  $M_A, M_B$  are predicted with SAM [KMR\*23] with a user-specified click or coarse bounding box, isolating salient regions for line selection and preserving perceptual continuity.

### 4.2. Interpolation via Structural Guidance

To obtain a smooth yet content-aware transition between  $C_A$  and  $C_B$ , we interpolate a sequence of intermediate structural primitives from the features in Section 4.1 by making use of the available motion cues. Specifically, we adopt a line-based interpolation scheme that produces intermediate line sets  $\{L_t\}_{t=1}^T$  between  $L_A$  and  $L_B$ , where  $T$  is the number of inbetween frames. These line sets serve as geometric anchors that guide synthesis toward temporally coherent, semantically aligned transitions. A core design choice of SAGE is to propagate structural cues not by naïve linear blending, but by coupling (i) *layer-aware line matching* and (ii) *motion-aware B-spline trajectories*, thereby avoiding trajectory crossings and ensuring that interpolated structures respect both local geometry and global motion.

#### 4.2.1. Layer-aware Line Matching

We enforce structural consistency by performing *layerwise* matching of lines using  $L_A, L_B$  and the corresponding segmentation masks  $M_A, M_B$ . Directly matching all lines in  $L_A$  to those in  $L_B$  is brittle: background structures can dominate the objective and degrade the transition, particularly for *diverse clips*. We therefore proceed in three steps, each reflecting a deliberate design decision:

(i) *Foreground selection.* We restrict attention to lines that lie in the foreground regions, selecting

$$L_A^{\text{fg}} := \{l \in L_A \mid l \cap M_A \neq \emptyset\}, \quad L_B^{\text{fg}} := \{l \in L_B \mid l \cap M_B \neq \emptyset\}.$$

This restricts subsequent marching on semantically salient objects while suppressing background clutter.

(ii) *Canonical normalization.* For each foreground region, we compute a tight bounding box  $B$  and normalize line endpoints into this canonical frame to obtain  $L_A^{\text{fg}}, L_B^{\text{fg}}$ . This makes matching robust to absolute position and scale, and ensures that correspondences are determined by relative geometric positioning rather than raw pixel coordinates.

(iii) *Hungarian matching.* We define a cost matrix  $C \in \mathbb{R}^{|L_A^{\text{fg}}| \times |L_B^{\text{fg}}|}$  for all candidate line pairs,

$$C_{ij} := \|c(I_A^i) - c(I_B^j)\|^2$$

where  $c(\cdot)$  is the canonical line center. Note that the cost matrix can be modified to include a weighted combination of other geometric attributes, such as line orientation and length. We prioritized predictable results for artists and opted for simpler position-only matching, as adding other features can cause line paths to contradict with each other later in the method. We use Hungarian matching to produce a set of one-to-one correspondences,

$$\mathcal{P} := \{(\hat{l}_A^i, \hat{l}_B^i)\}_{i=1}^K,$$

with unmatched lines discarded. These matched pairs  $\hat{l}_A^i \leftrightarrow \hat{l}_B^i$  form the foundation for motion-guided interpolation, as described next.

#### 4.2.2. Motion-aware B-spline Trajectories

While one could interpolate each matched pair  $(\hat{l}_A^i, \hat{l}_B^i) \in \mathcal{P}$  by blending direction, length, and center, such per-line interpolation ignores global scene dynamics. This often yields physically implausible trajectories (e.g., line crossings or structural collapse); see Figure 4 and Section 5. To address this, we introduce a two-scale interpolation scheme: first, a global foreground trajectory via B-splines; and then, local line blending within this evolving frame.

*Global trajectory (B-spline guidance).* Foreground bounding boxes  $B_A$  and  $B_B$  are computed from  $M_A$  and  $M_B$ . To capture dominant motion, we compute average flow vectors  $F_A, F_B$  around matched lines and displace the bounding boxes accordingly. We then define control points  $\{B_A, B_A + F_A, B_B - F_B, B_B\}$  and fit a cubic B-spline trajectory as,

$$B_t := \text{B-SPLINE}(B_A, B_A + F_A, B_B - F_B, B_B; \frac{t}{T}), \quad t = 1, \dots, T.$$

This simple design ensures that interpolated structures follow a globally smooth and motion-aware path, effectively providing a local frame, avoiding abrupt jumps or unnatural crossings.

*Local line interpolation.* Given each pair,  $\hat{l}_A^i, \hat{l}_B^i$ , which are normalized in the canonical space of  $B_A$  and  $B_B$ , intermediate lines are then blended linearly in the canonical space as,

$$\hat{l}_t^i := (1 - \frac{t}{T})\hat{l}_A^i + \frac{t}{T}\hat{l}_B^i,$$

and mapped back into image space using the transformation defined by  $B_t$ :

$$l_t^i := \text{TRANSFORM}_{B_t}(\hat{l}_t^i).$$

We thus arrive at the resulting line sets,

$$L_t := \{l_t^i \mid (l_A^i, l_B^i) \in \mathcal{P}\}$$

that encodes both local structure and global dynamics.

*Design rationale.* We chose this hierarchical strategy by combining B-spline foreground trajectories with local line interpolation to enforce smoothness of motion and semantic consistency simultaneously. Global trajectories capture camera/object motion, while local blending preserves fine structure. Together, they mitigate failures such as distracting line crossing, ghosting, or structural collapse that arise with naïve linear interpolation.

#### 4.3. Conditional Frame Generation

Finally, we condition a pretrained diffusion-based inbetweening model [ZRW\*25] with the interpolated line maps. The inbetween-

ing model takes  $(I_A^{N_A}, I_B^1, \{E_1 \dots E_T\})$  as input to produce intermediate frames  $\{I_t\}_{t=1}^T$ ,

$$I_t := \Psi(I_A^{N_A}, I_B^1, E_t),$$

where  $\Psi$  denotes the generative diffusion sampler and  $\{E_1 \dots E_T\}$  are edge maps that are passed as frame-wise conditions to the video generation model. The interpolated line maps are rasterized into frame-wise conditions by plotting each line set  $\{L_t\}$  into an edge map  $E_t$ . These edge maps are then injected with  $(I_A^{N_A}, I_B^1)$  via ControlNet-style conditioning. This enables zero-shot synthesis without fine-tuning, guided by structural and motion priors.

### 5. Evaluation

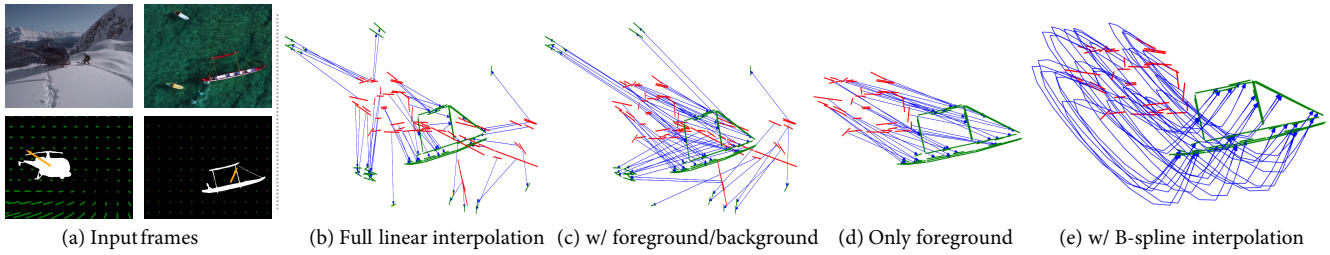
**Datasets.** We evaluate SAGE and competing methods on a diverse set of video clip pairs. Our test set is drawn from three sources: (a) artist-designed transitions, which provide high-quality reference examples; (b) image pairs adapted from related work, where we generate short video clips using an image-to-video workflow; and (c) diverse clips collected from public sources to span a wide range of motion, style, and complexity. The full dataset, along with reference transitions, will be released to enable reproducibility and facilitate future benchmarking.

**Comparisons.** We compare our approach against both classical and generative baselines. For content-aware morphing, we use Diff-Morpher [ZZX\*24]. Among generative transition methods, we test against FILM [RKT\*22], TVG [ZCL\*24], and Generative Inbetweening [WZC\*25] (GI). Finally, we also evaluate against VACE [JHM\*25], a universal video generation model capable of editing and synthesis (see project page). On the project page, we also qualitatively compare against SEINE [CWZ\*23].

**Metrics.** For the artist examples, we use the handcrafted transition as reference. We compute (i) motion similarity, defined as the cosine similarity between optical flows extracted from the reference and generated transitions, and summed over the pixels; and (ii) image and video similarity temporal smoothness metrics (FID, FVD). These allow us to capture both the motion smoothness and perceptual quality of transitions.

**Implementation details.** We use SEA-RAFT [WLD24] for computing optical flow, and normalize the results to unit vectors for similarity calculations. We take the transition videos generated by artists as ground truth videos when computing FID and FVD for the generated transition. Unless stated otherwise, we use default hyperparameters across baselines. All method outputs are normalized to the same resolution and temporal length (13 in-between frames), for fair comparison.

**Runtime.** The structural preprocessing is lightweight: on a single RTX 3090, feature and mask extraction take (1s) and line matching plus interpolation (2s); i.e., (~3s) total per transition. The runtime is then dominated by transition generation with the underlying diffusion model (5mins for 13 frames).



**Figure 4: Trajectory ablations for structural guidance.** Trajectories (blue) shown from  $L_1$  (red) to  $L_T$  (green). (a) Input frames with computed optical flow and segmentation; (b) Linear interpolation of all detected lines, matched regardless of foreground/background, resulting in trajectory crossings and line mismatches due to neglect of semantic structure; (c) Linear interpolation of detected lines, matched by foreground/background, retaining undesired crossings and mismatches; (d) Linear interpolation restricted to foreground lines (selected by  $M_A$ ,  $M_B$ ), yielding clearer trajectories for salient structures but still exhibiting crossover and motion inconsistency; (e) Motion-aware guidance combining the global bounding-box trajectory  $\{B_t\}_{t=1}^T$  with local line trajectories  $\{L_t\}_{t=1}^T$ , aligning structural evolution with scene/camera motion, as indicated in (a), while reducing trajectory crossovers.

**Qualitative evaluation.** Figure 5 shows representative examples across domains, including artistic edits, object-centric scenes, and natural footage. Compared to interpolation methods, SAGE produces more coherent structure and consistent, smooth motion, even diverse motion direction (CAB-TRAIN and SURFER-LSPEEDBOAT), diverse scale changes (CANDY-CLOUD), or diverse object category (HORSE-DOG). However, generative baselines occasionally collapse or introduce spurious content. For instance, FILM often results in a cross-fade-like transition, which can disrupt the fundamental structure and semantics of the figure (as illustrated in SKATEPARK-BIKER in Figure 6). Similarly, GI typically produces small, localized changes near the boundary frames, yet culminates in a sudden and abrupt mid-transition, indicating a failure to achieve a smooth temporal flow. Although generative baselines are capable of producing transitions that are more semantically and structurally meaningful, their inherent lack of explicit structural prior yields undesirable artifacts. Examples of this include the introduction of an unrelated human figure in the SKATEPARK-BIKER result for DiffMorpher, and the consistent, distracting left-to-right wiping artifact observed in the outputs of TVG for both the SKATEPARK-BIKER and HANDKERCHIEF-CRUISE examples, which disregard the actual motion direction. Additional examples are provided in the project page.

**Table 1: Quantitative comparisons.** We report metrics on image quality (FID), video quality (FVD), and motion adherence (flow similarity). While some baselines achieve strong image/video scores (e.g., GI on FID, TVG on FVD), they fall short in preserving motion consistency. Our method attains the best flow similarity while remaining competitive in FID/FVD, demonstrating that a good transition must balance both visual fidelity and motion adherence rather than optimizing one at the expense of the other.

Method	DiffMorp	GI	TVG	FILM	ours
FID ↓	151	<b>147</b>	157	157	153
FVD ↓	2641	2696	<b>2093</b>	2404	<b>2185</b>
Flow similarity ↑	<b>0.61</b>	0.55	0.57	0.56	<b>0.69</b>

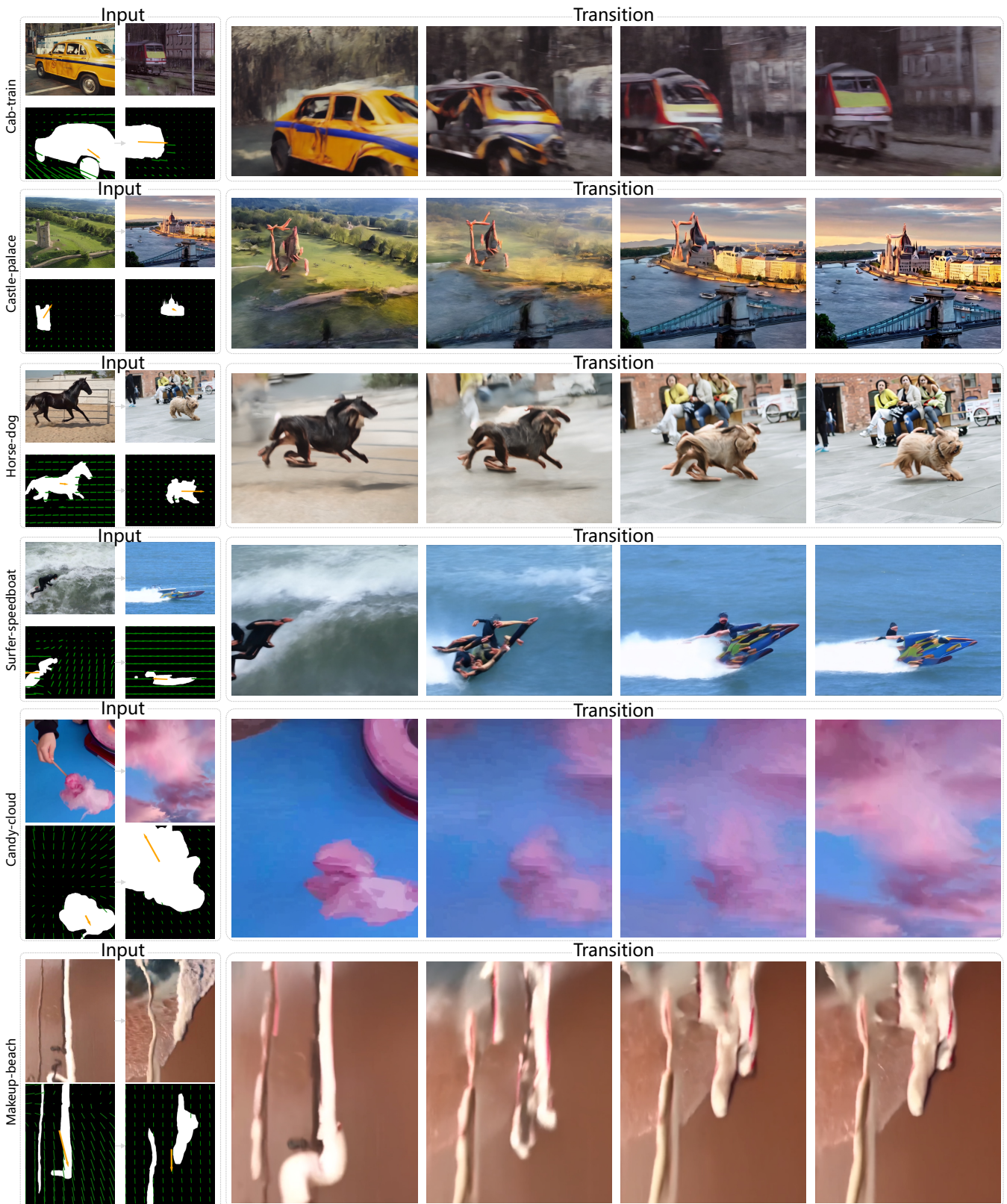
**Quantitative evaluation.** Table 1 presents a quantitative comparison against baseline methods. Our approach, SAGE, achieves the highest flow similarity to the ground truth (GT), which validates the use of motion consistency to constrain the transition. SAGE also secures the second-best results for FID and FVD. Note that, although GI obtains a better FID score, it simply duplicates boundary frames, creating an abrupt transition; as shown in Figure 6 and the project page. Similarly, TVG produces a constant left-to-right transition that fails to adapt to the source and target motion.

**User study.** We conducted a user study to evaluate perceptual quality and preference. We recruited 26 participants with mixed backgrounds in video editing and graphics. Each participant was shown 24 pairs of transitions, comparing our method to one of the four baselines: DiffMorpher, Generative Inbetweening (GI), TVG, and FILM, randomly sampled from these sources. The order of our versus baseline methods was also randomized. Each transition was concatenated with its corresponding input clips as  $(C_A, T, C_B)$  and was looped over unlimited time; the two concatenated clips were aligned side-by-side and played synchronously over time. Figure 7 shows a snapshot of the user study setup.

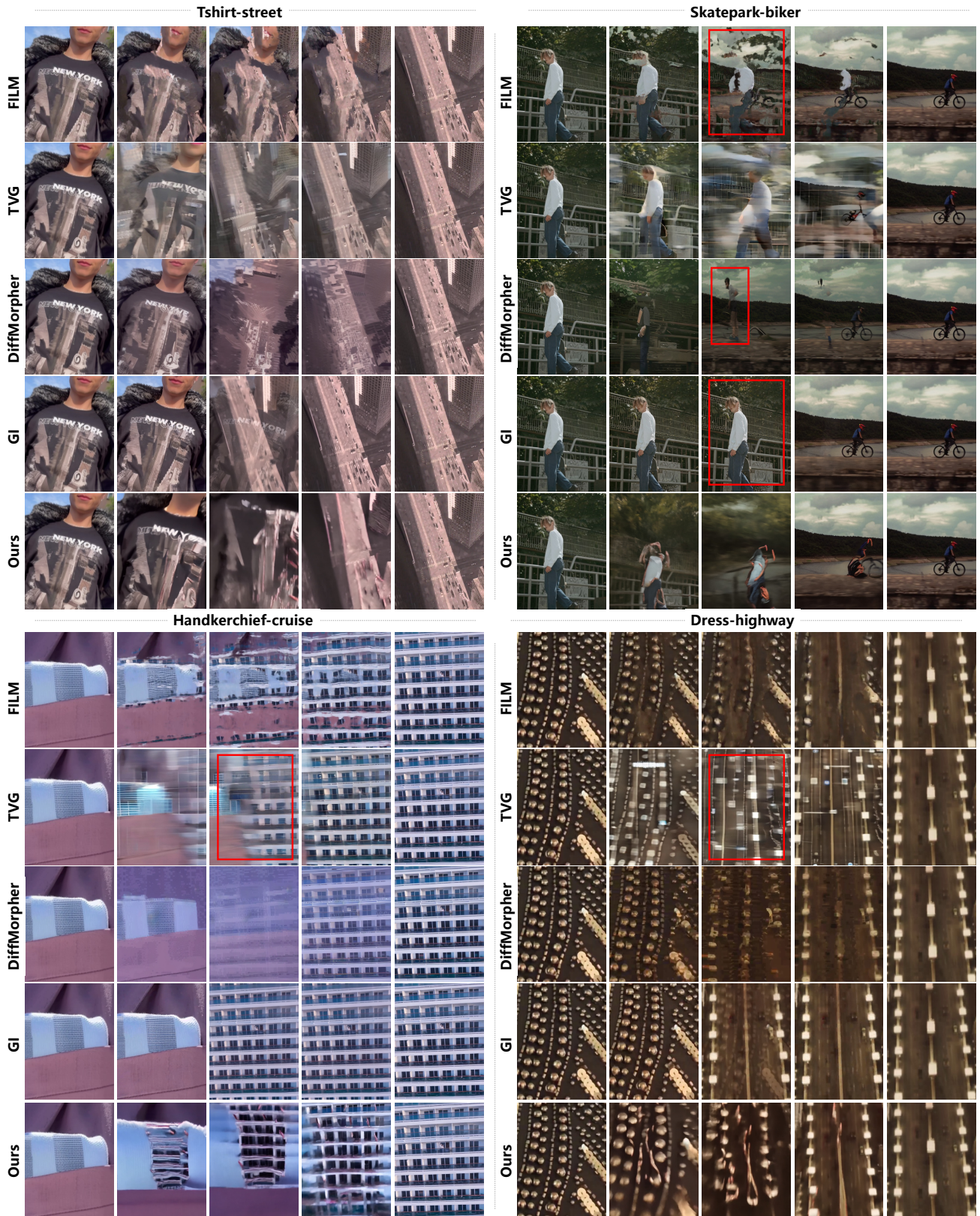
For each pair, participants were given a forced choice task between ours vs a baseline across four different criteria: (i) which transition displayed better temporal consistency over time, (ii) which transition displayed more plausibility, (iii) which

**Table 2: User study results.** SAGE’s video transitions were strongly preferred over all baseline methods. Users consistently rated our transitions higher across all measured aspects: Temporal Consistency, Plausibility, Motion Complexity, and Overall Preference. Note that a score of say  $x\%$  for method- $y$  indicates that SAGE was preferred by  $x\%$  of the participants over method- $y$ .

Method	FILM	TVG	DiffMorp	GI
Temporal Consistency	81.41%	80.12%	91.02%	86.53%
Plausibility	78.84%	76.28%	86.53%	80.12%
Motion Complexity	83.33%	82.69%	90.38%	89.10%
Overall Preference	82.69%	78.20%	89.10%	85.89%
Average	81.57%	79.33%	89.26%	85.42%



**Figure 5: Result gallery.** Qualitative results on diverse video clips, showcasing the model’s performance on complex transitions in scene scale (local-global), object category, and motion direction. Full videos are available on our project page.



**Figure 6: Comparisons.** Qualitative comparison with baseline methods, demonstrating that SAGE generates more plausible video transitions by maintaining consistency in motion, foreground objects, and background scenery. Full videos are available on our project page.

Generative Video Transitions for Diverse Clips

In this study you will be asked to evaluate pairs of video transitions against each other.

A video transition is a set of intermediary frames that transforms one input video into another. In each transition, the 2 input video clips are significantly different – we call them diverse clips.

A transition can be evaluated according to the following definitions:

- **Temporal consistency** – the progression of frames over time is visually consistent
- **Plausibility** – each frame within the transition looks physically realistic, or visually feasible
- **Motion complexity** – indicates smooth, detailed transformations that appear carefully crafted (e.g. objects morphing, structures shifting). Changes are not chaotic or abrupt.



For each pair of transitions, we ask for a comparison across 4 categories:

a) Temporal consistency  
 b) Plausibility  
 c) Motion complexity  
 d) Overall preference

There are 24 comparisons and the form should take 10-15 min in total. Please give each page some time to load the GIFs that display the transitions.

(22/24) Observe the two video transitions, A (left) and B (right).

Please give the transitions a few seconds to load.

Rate the better transition based on the 4 categories below:\*

	Transition A (left)	Transition B (right)
Temporal consistency	<input type="radio"/>	<input type="radio"/>
Plausibility	<input type="radio"/>	<input type="radio"/>
Motion complexity	<input type="radio"/>	<input type="radio"/>
Overall Preference	<input type="radio"/>	<input type="radio"/>

**Figure 7: User study design.** We conducted a user study to evaluate our method against multiple baselines. In each question, participants compared our approach with a randomly sampled baseline (FILM, TVG, DiffMorpher, or GI) across four criteria. (Left) Starting instructions shown to participants; (right) Example of a randomly sampled comparison pair.

transition displayed more motion complexity (smooth, detailed changes), and (iv) overall preference. Results (see Table 2) show a clear preference for SAGE across all comparisons in each aspect.

**Ablations.** We ablate the key components of SAGE and show results in the project page. Specifically, we evaluate: (i) without structural guidance – this is Generative Inbetweening [WZC\*25]; (ii) without the layered matching; and (iii) using linear flow interpolation instead of B-spline interpolation for flow guidance. Results show that removing either structural or flow guidance significantly degrades transition quality, confirming the complementarity of both cues. Note that when the motions in the two source clips are already well-aligned, simple linear blending of matched structures performs comparably to our B-spline approach. However, for more diverse clip pairs where motion trajectories differ in direction, scale, or continuity – B-spline interpolation produces smoother global paths and more natural motion inbetweening. Its effect is most visible in cases such as CAB-TRAIN, SKATEPARK-BIKER, and HELICOPTER-BOAT, where clip motions are distinct; linear interpolation leads to sudden or inconsistent structural changes, whereas B-splines produce smoother, more anticipatory motion. It prevents abrupt movements and trajectory crossings that arise with linear blending. Together, these components account for the overall effectiveness of our method.

**Failure cases.** A limitation of our current approach stems from its reliance on a video generative backbone pretrained heavily on human poses. At times, this specificity can cause the model to interpret certain line configurations as limbs and hallucinate human-like structures along structural edges, as illustrated in the CASTLE-PALACE and SURFER-LSPEEDBOAT examples of Figure 5. We believe that employing a more general-purpose video backbone would mitigate these artifacts – note that retraining on transition sequences is not an intuitive option due to the lack of suitable training data.

**Limitations.** While SAGE demonstrates promising results for structure-aware generative transitions on diverse clips, it has several limitations. First, our approach relies on structural guidance from line maps and optical flow; when clips lack salient linear features or when flow estimation fails due to occlusions, textureless regions, or rapid motion, the resulting correspondences may be unreliable. Second, the method assumes that structural correspondences can be meaningfully established between clips; in highly abstract or stylistically divergent content, the extracted matches may be ambiguous or misleading. Finally, our current framework does not explicitly model appearance blending, which may lead to visual discontinuities in texture-rich regions. We believe these limitations can be addressed by extending our structural guidance with semantic cues (e.g., curved lines and/or Dino features), higher-order correspondence costs, and appearance-aware generation, which we leave for future work.

## 6. Conclusion

We have presented SAGE, the first method to produce structure-aware generative transitions between diverse video clips. Inspired by artist workflows, we demonstrated that carefully extracting, matching, and interpolating linear features across source clips provides effective structural guidance for generative video models. This design enables aesthetically pleasing and engaging transitions in a zero-shot setup, making it possible to realize generative workflows even in scenarios where collecting training data is infeasible.

Looking forward, several promising directions remain. First, incorporating semantic cues (e.g., Dino features samples on the lines as in [DMM24]) to score and refine feature matches could improve both robustness and perceptual quality. Second, integrating local smoothness priors and higher-order matching costs may further enhance correspondence estimation. Third, blending structural guidance with appearance information and motion flow could extend our framework into richer generative workflows (e.g., audio-guided [CZL\*14]), potentially combining structural interpolation with adaptive cross-fade strategies for even smoother transitions. We hope that our work provides both a practical tool for video editing and a foundation for future research on structure-aware generative transitions.

**Acknowledgments.** We thank Burak Kizil for his assistance with the VACE comparisons, and Lilian Welschinger and Hao Xu for their valuable comments and suggestions. We are also grateful to the user study participants for their time, and to the anonymous reviewers for their constructive feedback on the original submission. This work was supported in part by generous gifts from Adobe Research and the UCL AI Centre.

## References

- [BLM\*19] BAO W., LAI W.-S., MA C., ZHANG X., GAO Z., YANG M.-H.: Depth-aware video frame interpolation. In *Proc. Computer Vision and Pattern Recognition (CVPR)* (2019). 2, 3
- [BN92] BEIER T., NEELY S.: Feature-based image metamorphosis. *Proc. SIGGRAPH* (1992), 35–42. 2, 3
- [CWZ\*23] CHEN X., WANG Y., ZHANG L., ZHUANG S., MA X., YU J., WANG Y., LIN D., QIAO Y., LIU Z.: SEINE: Short-to-long video

- diffusion model for generative transition and prediction. In *International Conference on Learning Representations* (2023). 2, 3, 5
- [CZL\*14] CHENG M.-M., ZHENG S., LIN W.-Y., VINEET V., STURGESS P., CROOK N., MITRA N., TORR P.: Imagespirit: Verbal guided image parsing. *ACM Transactions on Graphics* (2014). 9
- [DMM24] DUTT N. S., MURALIKRISHNAN S., MITRA N. J.: Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 4494–4504. 9
- [HZH\*22] HUANG Z., ZHANG T., HENG W., SHI B., LIU S.: RIFE: Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision (ECCV)* (2022). 2, 3
- [JGZ\*24] JAIN A., GHARBI M., ZHANG R., LIU C., FREEMAN W. T., DURAND F., DEKEL T.: Video interpolation with diffusion models. *Proc. Computer Vision and Pattern Recognition (CVPR)* (2024). 2, 3
- [JHM\*25] JIANG Z., HAN Z., MAO C., ZHANG J., PAN Y., LIU Y.: VACE: All-in-one video creation and editing. *Arxiv* (2025). [arXiv: 2503.07598](https://arxiv.org/abs/2503.07598). 1, 2, 3, 5
- [KMR\*23] KIRILLOV A., MINTUN E., RAVI N., MAO H., ROLLAND C., GUSTAFSON L., XIAO T., WHITEHEAD S., BERG A. C., LO W., DOLLAR P., GIRSHICK R. B.: Segment anything. In *International Conference on Computer Vision* (2023), pp. 3992–4003. 4
- [KRK\*25] KYE D., ROH C., KO S., EOM C., OH J.: AceVFI: A comprehensive survey of advances in video frame interpolation. *Arxiv* (2025). [arXiv:2506.01061](https://arxiv.org/abs/2506.01061). 3
- [Mar25] MARY WOODCOCK: Mastering the art of transitions for youtube videos. <https://vimeo.com/blog/post/adding-video-transitions>, 2025. Accessed: 2025-09-20. 3
- [Ond00] ONDAATJE M.: *The Conversations: Walter Murch and the Art of Editing Film*. Knopf, 1900. 3
- [Pea16] PEARLMAN K.: *Cutting Rhythms: Intuitive Film Editing*. Routledge, 2016. 3
- [PPZ\*24] PARDO A., PIZZATI F., ZHANG T., PONDAVEN A., TORR P., PEREZ J. C., GHANEM B.: Matchdiffusion: Training-free generation of match-cuts. *arXiv preprint arXiv:2411.18677* (2024). 3
- [PSY\*23] PAUTRAT R., SUÁREZ I., YU Y., POLLEFEYS M., LARSSON V.: Gluestick: Robust image matching by sticking points and lines together. In *International Conference on Computer Vision* (2023), pp. 9672–9682. 4
- [RKT\*22] REDA F., KONTKANEN J., TABELLION E., SUN D., PANTOFARU C., CURLESS B.: FILM: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)* (2022). 1, 2, 3, 5
- [Viv25] VIVIAN TEJEDA: Video transitions: The ultimate guide in 2025. <https://www.descript.com/blog/article/video-transitions>, 2025. Accessed: 2025-09-20. 3
- [WLD24] WANG Y., LIPSON L., DENG J.: SEA-RAFT: simple, efficient, accurate RAFT for optical flow. In *European Conference on Computer Vision (ECCV)* (2024), Leonardis A., Ricci E., Roth S., Rusakovsky O., Sattler T., Varol G., (Eds.), vol. 15065, pp. 36–54. 4, 5
- [Wol98] WOLBERG G.: Image morphing: a survey. In *The Visual Computer* (1998), vol. 14, pp. 360–372. 2, 3
- [WZC\*25] WANG X., ZHOU B., CURLESS B., KEMELMACHER-SHLIZERMAN I., HOLYNSKI A., SEITZ S. M.: Generative inbetweening: Adapting image-to-video models for keyframe interpolation. *Arxiv* (2025). [arXiv:2408.15239](https://arxiv.org/abs/2408.15239). 3, 5, 9
- [YZY\*25] YANG Z., ZHANG J., YU Y., LU S., BAI S.: Versatile transition generation with image-to-video diffusion. *Arxiv* (2025). 3
- [ZCL\*24] ZHANG R., CHEN Y., LIU Y., WANG W., WEN X., WANG H.: Tvg: A training-free transition video generation method with diffusion models. *Arxiv* (2024). 1, 2, 3, 5
- [ZCZ\*25] ZHANG Z., CHEN H., ZHAO H., LU G., FU Y., XU H., WU Z.: EDEN: Enhanced diffusion for high-quality large-motion video frame interpolation. In *Proc. Computer Vision and Pattern Recognition (CVPR)* (2025). 2, 3
- [ZRW\*25] ZHU T., REN D., WANG Q., WU X., ZUO W.: Generative inbetweening through frame-wise conditions-driven video generation. *Proc. Computer Vision and Pattern Recognition (CVPR)* (2025). 1, 2, 5
- [ZZX\*24] ZHANG K., ZHOU Y., XU X., DAI B., PAN X.: Diffmorpher: Unleashing the capability of diffusion models for image morphing. *Proc. Computer Vision and Pattern Recognition (CVPR)* (2024), 7912–7921. 1, 2, 3, 5