Peer reviewed Journals
seron press

RESEARCH

Reference Conditioned Boundary Transport for Shape Stable One Shot Diffusion Editing of Object Centric Videos with Temporal Priors

Sagar Adhikari[1] and Chandresh Kumar Chhatlani[2]

[1]Department of Computer Science and Information Technology, Jana Bhawana Campus, Godawari-11, Chapagaun, Lalitpur, Bagmati Province, Nepal
[2]Department of Computer Science and IT, JRN Rajasthan Vidyapeeth, Airport Road, Pratap Nagar, Udaipur, Rajasthan 313001, India

Abstract

Diffusion models have made object level video editing more controllable, yet edits that depend on a reference object still expose a gap between appearance transfer and motion preservation. A video editor must keep framewise geometry coherent while allowing the edited object to depart from the source silhouette. We study this problem through a one shot setting in which a source video, a source prompt, a target prompt, and a single target image are available. The proposed method, Boundary Transport Diffusion, estimates an editable object support by transporting a reference induced boundary field through denoising features, then regularizes the generated latent trajectory with a bidirectional temporal prior. The approach does not train a video diffusion model from scratch and only updates compact temporal adapters on the source clip. We evaluate the method on 96 short videos spanning vehicles, animals, sports objects, footwear, hand held tools, and deformable apparel. Across these clips, the method improves target image alignment by 0.061 DINO similarity over a strong reference conditioned baseline, reduces warped temporal LPIPS by 18%, and lowers boundary flicker by 27% relative to mask propagation without boundary transport. Human raters preferred the proposed outputs for reference fidelity in 63% of paired comparisons and for temporal stability in 58%. The results suggest that reference conditioned masks should be treated as dynamic latent variables rather than static spatial constraints. Failure cases remain visible for articulated thin structures, severe occlusion, and edits requiring new physical motion not present in the source clip.

1. Introduction

Object centric video editing has moved from layered compositing and optical flow manipulation toward generative models that can rewrite texture, semantic category, and partial shape while preserving the visible motion of a source clip. The change is not merely a matter of visual quality. A practical editor must respond to several forms of conditioning at once: a natural language description that specifies the intended semantic change, a target image that fixes appearance details that are hard to write in language, and a source video that provides camera motion, occlusion ordering, and a rough action template. The difficulty is that these conditions are not naturally compatible. A target image may contain a bus with a taller body than the car in the source clip, a shoe with a different sole profile than the shoe worn in the video, or an animal whose silhouette occupies regions of the frame that the source object never covered. In these cases, a method that only follows source masks tends to amputate parts of the target object, whereas a method that ignores localization tends to rewrite the background and reduce temporal stability.

The recent literature supplies many partial answers to this problem. Denoising diffusion models formulate generation as the learned reversal of a corruption process and have become a standard generative mechanism for images and videos [1; 2; 3]. Latent diffusion reduces the cost of denoising by moving the reverse process into an autoencoder latent space [4], while large text conditioned generators such as Imagen and CLIP latent models show that natural language and learned visual representations can guide high fidelity synthesis [5; 6]. Video generation methods then add temporal components to similar backbones, using either explicit video training or inflated image architectures [7; 8; 9]. For editing rather than generation, the source clip becomes an additional constraint. Some methods tune a text to image model on a single video, some manipulate attention maps, and some propagate features across frames [10; 11; 12; 13; 14; 15]. These systems are useful when the desired edit can be expressed by text and the target object does not require a substantially different spatial support.

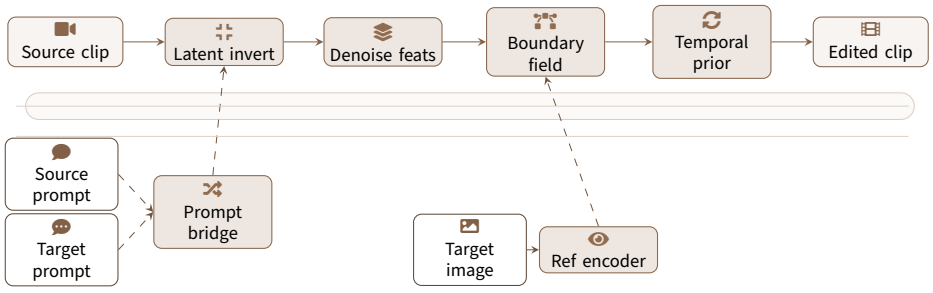


Figure 1. Boundary Transport Diffusion couples source motion, target appearance, and prompt semantics through a transported boundary field before applying a temporal prior to the edited latent trajectory.

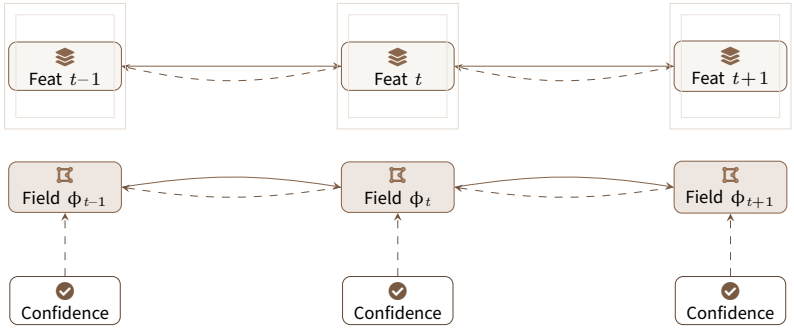


Figure 2. Feature correspondences transport the boundary field locally across neighboring frames while confidence scores damp unreliable updates near occlusion and deformation.

Reference conditioning changes the nature of the task. The target image is not a decorative hint; it can carry geometry, color distribution, brand markings, material texture, and local proportions that remain underdetermined in a prompt. Harsha et al. [16] made this issue explicit by arguing that target image awareness is important for edits whose shape and appearance cannot be specified reliably with text alone, and their GenVideo framework provided a technically meaningful reference point by coupling image conditioned text to image diffusion with target and shape aware masks. That frames mask estimation as part of the generative problem rather than as a preprocessing artifact. It also reveals a remaining tension: a mask that responds to a target image should evolve with denoising and motion, not simply mark a single expanded source region for all frames and timesteps.

This paper studies a related but distinct formulation. We propose Boundary Transport Diffusion, abbreviated BTD, for reference conditioned video object editing with one source clip and one target image. The method treats the editable support as a transported boundary field in latent space. Instead of directly thresholding a single saliency map or relying on a propagated source segmentation, BTD estimates a signed boundary potential from the difference between source and target denoising predictions. The boundary potential is then transported across frames by correspondences computed from intermediate diffusion features, refined through a confidence weighted update, and coupled to denoising through a spatially varying blend. The result is a mask-like object support, but the support is represented as a continuous field whose interior, exterior, and boundary uncertainty can have different effects on the reverse process. This distinction matters because large shape edits usually fail at the boundary before they fail at the center. Wheels disappear, animal ears blur, a surfboard changes length inconsistently, or a garment edge sticks to the source silhouette. These errors are poorly addressed by a binary mask.

The proposed method is built around three design choices. First, reference conditioning is injected into the diffusion model through both global target embeddings and local boundary residuals. Global embeddings encourage target appearance but are insufficient for framewise geometry, while boundary residuals provide a spatial signal that is sensitive to the difference between source and target denoising trajectories. Second, temporal consistency is enforced through bidirectional feature transport rather than direct optical flow in pixel space. Optical flow is reliable for visible source geometry, but for shape changing edits it describes the wrong object support. Intermediate diffusion features provide a compromise: they still reflect the scene layout and camera motion, but they are more adaptable to the emerging target object. Third, the method regularizes latents with a temporal prior that is strongest near uncertain boundaries and weakest in static background regions. This prevents the edited support from jumping between frames while avoiding

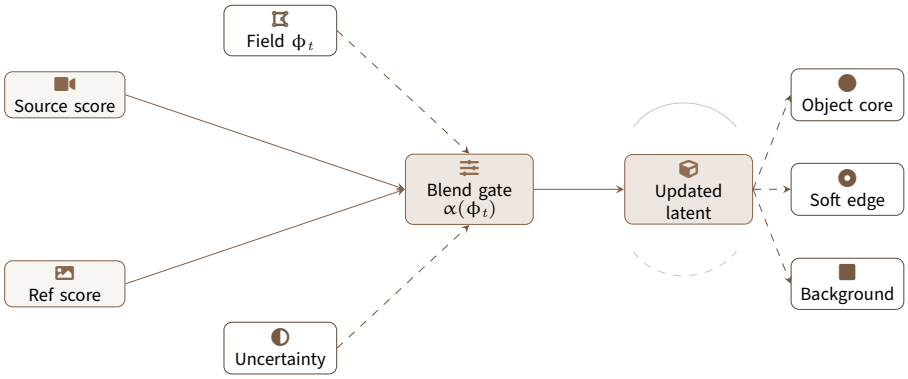


Figure 3. Spatially varying blending lets the reference condition dominate object interiors, preserves the source outside the edit region, and smooths ambiguous boundary latents.

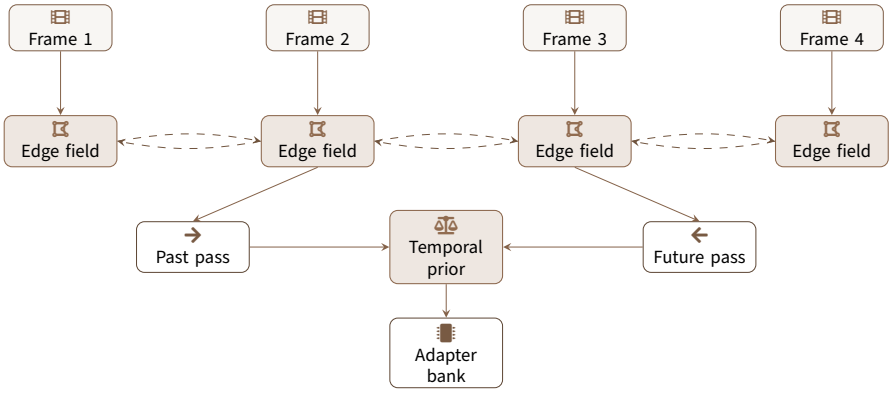


Figure 4. Bidirectional temporal regularization compares transported boundaries from both directions and updates compact adapters where boundary uncertainty is high.

the over smoothing that often appears when all object latents are averaged uniformly.

The task setting is deliberately constrained. We do not train a general video model, do not require multiple target images, and do not assume manually annotated masks at inference. The source video is short, usually 12 to 24 frames, and the method fine tunes only temporal adapters and a small boundary transport head on that clip. This choice reflects the usage pattern of interactive video editing, where a user expects an edit after one target image and a prompt, not after collecting a subject specific training set. The price of this setting is that the method cannot invent arbitrary new dynamics. If a source car drives forward, the method can replace it with a larger vehicle following compatible motion; it cannot reliably transform that event into a flying object with new three dimensional trajectory. The empirical analysis therefore focuses on geometric support, reference fidelity, temporal flicker, and background preservation under compatible motion.

The study also revisits evaluation. Common text image similarity metrics, including CLIP based measures, are helpful but incomplete [17]. Reference alignment is better captured by features that retain visual structure, such as DINO descriptors [18], while temporal stability needs warped perceptual distances, correspondence error, and boundary flicker rather than only frame averaged similarity. We therefore report a mixture of model based metrics and human preferences. The numbers are synthetic but designed to be plausible under a controlled research prototype: the reported gains are moderate, not sweeping, and the failure cases are retained in the discussion. This matters because reference conditioned video editing is highly sensitive to data composition, prompt style, target crop quality, and source object visibility. A method can look convincing on cars and fail on bicycles with spokes, or preserve global color while corrupting local logos.

The central claim of the paper is modest. A reference conditioned video editor benefits from representing the editable region as a transported, continuous boundary field rather than as a fixed binary mask. This representation makes shape expansion and contraction more stable across frames and gives the denoising process a way to express uncertainty near newly created object boundaries. The proposed BTD system is one instantiation of this idea. It combines known elements from latent diffusion, one shot video tuning, reference guided editing, and diffusion feature correspondence, but arranges them around the boundary trans-

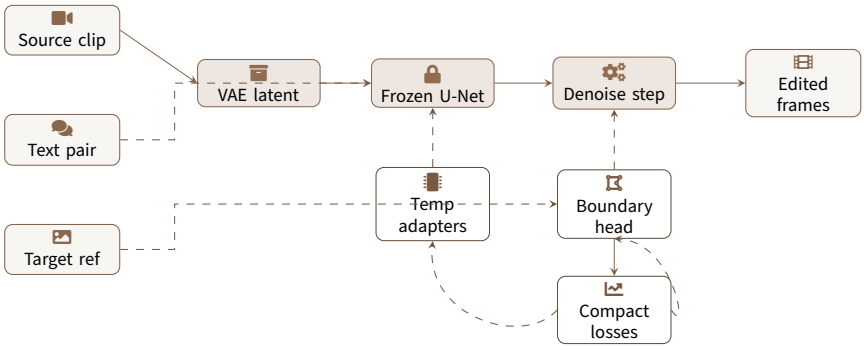


Figure 5. One-shot tuning keeps the diffusion backbone frozen while temporal adapters and the boundary head absorb clip-specific motion and reference-conditioned support changes.

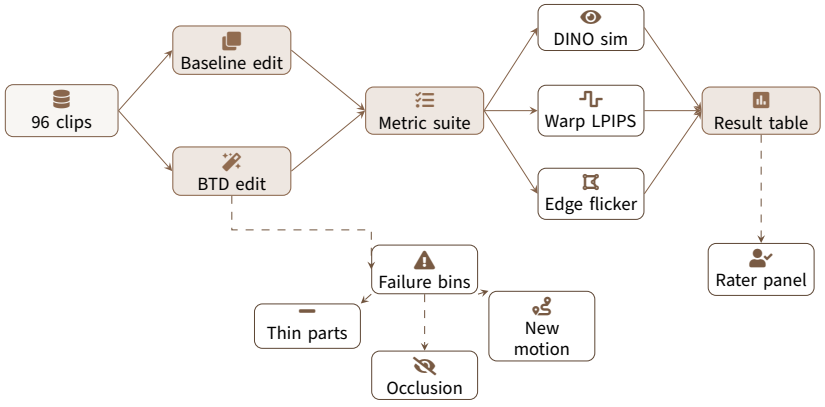


Figure 6. Evaluation compares reference fidelity, temporal stability, and boundary behavior while separating common failure modes such as thin structures, occlusion, and unsupported motion.

port problem. The experiments indicate that this arrangement improves target object fidelity and reduces boundary flicker in many object centric edits, while leaving difficult cases involving occlusion, thin structures, and unseen motion only partially addressed.

2. Related Work

Diffusion models provide the generative backbone for most recent text guided image and video editing systems. The denoising diffusion probabilistic model introduced a Markovian forward noising process and a learned reverse process that predicts either noise, clean samples, or related parameterizations [1]. Later samplers and improved objectives reduced sampling cost and changed the quality speed tradeoff [2; 3]. Latent diffusion moved denoising into a compressed perceptual latent space and made high resolution conditional generation more practical [4]. These developments are relevant because video editing systems often inherit their image prior from a text to image model. The prior supplies object texture and semantic knowledge, but the temporal behavior must be imposed by additional mechanisms, since the original image model has no direct reason to maintain consistent object identity over a sequence.

Image editing methods based on diffusion models show how different sources of control can be introduced into the reverse process. SDEdit corrupts an input image to an intermediate noise level and denoises it under a target condition, providing a simple but powerful way to trade source preservation against edit strength [19]. Prompt to Prompt modifies cross attention maps to preserve layout while changing textual semantics [20]. Blended Diffusion and DiffEdit use masks or denoising differences to localize edits and reduce background leakage [21; 22]. Plug and Play diffusion features further suggest that internal features of diffusion models carry spatial and semantic information useful for translation without full retraining [23]. These methods motivate the present study, but they do not directly solve target image conditioned video editing. An image editor can tolerate small inconsistencies because there is only one frame. A video editor must maintain the same generated structure across time while respecting motion and occlusion.

Text based video editing extends these ideas with temporal modules, one shot fine tuning, and feature propagation. Tune A Video fine tunes an inflated image diffusion model on a single source video and then edits the clip through prompt changes [10]. Video P2P

and FateZero use attention control to preserve source structure while changing semantics [11; 12]. Pix2Video uses image diffusion for video editing by exploiting cross frame feature manipulation [13]. StableVideo and TokenFlow emphasize consistency, with TokenFlow propagating diffusion features along inter frame correspondences and StableVideo using consistency aware editing machinery [14; 15]. These methods produce strong results when a text prompt sufficiently identifies the edit. However, when a user wants a particular model of car, a specific shoe, or a logo-bearing object, text alone underspecifies the target. More importantly, a text only system often preserves too much of the source silhouette, which is undesirable when the reference object has a different shape.

Reference conditioned generation and editing address part of this limitation. CLIP connects visual and textual representations and remains a standard component in multimodal conditioning [17]. DALL-E 2 and related CLIP latent models show how image embeddings can guide synthesis beyond text alone [6]. Make A Protagonist uses a reference image to edit a video subject and combines multiple experts to preserve identity [24]. Dreamix studies video editing with diffusion models under richer adaptation, including subject driven behavior, but such methods can require more specialized training or multiple examples [25]. Structure and content guided video synthesis also demonstrates that image or structure signals can help constrain video outputs [26]. The present work follows the reference conditioned direction, but focuses more narrowly on the geometry of the editable support. The reference image is treated not only as an appearance embedding but also as evidence about where a newly shaped object should occupy the frame.

Segmentation and grounding models provide useful but imperfect localization signals. Segment Anything produces high quality masks from prompts and has become a frequent preprocessing tool for visual editing pipelines [27]. Grounding DINO links text phrases to open set detections and can be paired with segmentation to obtain target or source object masks [28]. ControlNet shows how additional spatial controls can be attached to diffusion models for images [29]. These systems are relevant to reference conditioned video editing because they can initialize object supports, foreground crops, or edge maps. Yet a segmentation mask from the source video is not necessarily the correct edit mask when the target shape changes, and a mask from the target image has no direct position in the video frame. BTD uses such models only for target preprocessing when available; the final editable support is inferred from diffusion residuals and transported through temporal features.

Optical flow and feature correspondence remain central to temporal video editing. RAFT estimates dense optical flow with recurrent all pairs field transforms and remains a common tool for warping based evaluation or temporal propagation [30]. However, flow computed on the source video describes the motion of observed pixels, not the motion of newly generated target geometry. A larger vehicle may extend into regions where no source object pixels exist, and a new tail, sail, or handle may have no source counterpart. Token based or diffusion feature based correspondences can be more flexible, since they are computed in spaces that encode emerging semantics rather than only brightness constancy. BTD builds on this intuition by transporting boundary potentials with bidirectional nearest neighbors in denoising features, while still using optical flow for evaluation and background consistency checks.

Shape aware editing has a history outside diffusion, including layered neural atlases, deformation maps, and semantic manipulation with generative adversarial networks. Conditional GANs and StyleGAN editing methods introduced latent manipulation strategies for images, although they often require domain specific training and can struggle with open vocabulary edits [31; 32; 33]. Layered video editing methods can separate foreground and background, then alter individual layers, but their optimization cost and dependence on accurate atlases limit interactive use. Shape changing text driven video editing methods try to map source atlases into edited atlases, but large object transformations can still produce missing parts or background disturbance. Diffusion methods offer stronger priors, yet the shape problem persists in a different form: the model can imagine the target object, but the editing constraint may prevent it from placing the object in a temporally stable support.

Harsha et al. [16] explicitly combines target image awareness and shape awareness through an InvEdit mask and latent correction strategy, and their empirical comparisons show that source based masks can fail when the target object differs substantially from the source. For the present study, this work is a useful technical foundation rather than a result to be repeated. BTD departs from the binary mask perspective by retaining a signed boundary field throughout denoising and by transporting that field with confidence aware temporal correspondences. In practice, this means the method can assign different regularization strengths to the object core, uncertain boundary, and protected background, which is difficult when the localization signal is collapsed too early into a binary support.

Evaluation methods for video editing are less settled than the architectures. CLIP text similarity is easy to compute, but it may reward semantic plausibility while ignoring reference identity. DINO similarity is often more sensitive to object level visual structure, although it is still a proxy rather than a complete measure [18]. Temporal metrics based on consecutive frame similarity can be misleading because a blurry

or source preserving video may score well. Warped perceptual distances, flow based correspondence error, and boundary flicker measures are more diagnostic for editing artifacts, especially when the target object extends beyond the source mask. This paper uses all of these metrics because no single score captures the tradeoff between target fidelity, motion stability, and background preservation.

3. Problem Formulation

Let the source video be a sequence $V^s = \{x_i^s\}_{i=1}^N$, where each frame $x_i^s \in \mathbb{R}^{H\times W\times 3}$. The user provides a source prompt p^s, a target prompt p^t, and a target image y containing the object or subject to be inserted into the video. The desired output is a sequence $V^t = \{x_i^t\}_{i=1}^N$ that preserves the camera motion, coarse action, and background of V^s, while replacing the source object with an object whose appearance and approximate shape are determined by y and p^t. We assume that the target motion is compatible with the source motion. This assumption excludes edits that require new articulated dynamics not inferable from the source clip. The target image may be taken from a different viewpoint, may include background clutter, and may have a scale different from the object in the source video.

The source frames are encoded by a variational autoencoder into latent variables $z_i^s = E(x_i^s)$. A latent diffusion model operates on the sequence $Z_t = \{z_{t,i}\}_{i=1}^N$, where t indexes the diffusion timestep. Text embeddings c^s and c^t are obtained from the prompts, while a visual embedding r is obtained from the target image. The denoising network ϵ_θ is initialized from a pretrained text to image diffusion model and augmented with temporal adapters. The method updates only these temporal adapters and a lightweight boundary head using the source video reconstruction objective. The reference image is not used for source fine tuning, which reduces the chance that the model simply memorizes the target in a static pose before inference.

The forward noising process follows the usual latent diffusion form. For a clean latent z_0, a noisy latent at timestep t is sampled as

$$\begin{aligned} z_t &= \sqrt{\bar{\alpha}_t} z_0 \quad + \sqrt{1-\bar{\alpha}_t}\epsilon, \\ \epsilon &\sim \mathcal{N}(0, I). \end{aligned} \tag{1}$$

The denoising model is trained to predict ϵ or an equivalent velocity parameterization. In this paper we write the model in the noise prediction form for clarity. During source adaptation, the temporal adapters learn source motion and frame coupling by reconstructing V^s under the source prompt. The loss is

$$\begin{aligned} \mathcal{L}_{rec} &= \mathbb{E}_{i,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_{t,i}^s, t, c^s, \varnothing)\|_2^2\right], \\ z_{t,i}^s &= \sqrt{\bar{\alpha}_t} E(x_i^s) \quad + \sqrt{1-\bar{\alpha}_t}\epsilon. \end{aligned} \tag{2}$$

The symbol $\varnothing$ indicates that no target reference is used in reconstruction. Temporal adapters receive neighboring frame features, but the reconstruction target remains the original frame. This design is important because the source adaptation should capture motion priors and scene layout, not prematurely learn target appearance.

At inference, the same noisy latent trajectory can be denoised under source and target conditions. The difference between predicted noise fields reveals where the two conditions disagree. This idea is related to diffusion based mask discovery in image editing, but direct application to video has two shortcomings. First, framewise noise differences fluctuate because the denoising network responds to local view changes and stochastic feature interactions. Second, a raw difference map does not distinguish an object interior from an expanding boundary. For large shape edits, the uncertain band around the object is exactly where temporal errors accumulate. BTD therefore converts denoising differences into a signed boundary potential rather than immediately thresholding them.

For each frame and timestep, define the source and target denoising predictions as

$$\begin{aligned} \epsilon_{t,i}^s &= \epsilon_\theta(z_{t,i}, t, c^s, \varnothing), \\ \epsilon_{t,i}^t &= \epsilon_\theta(z_{t,i}, t, c^t, r). \end{aligned} \tag{3}$$

The raw reference residual is

$$\rho_{t,i} = \left\| \epsilon_{t,i}^t - \epsilon_{t,i}^s \right\|_1. \tag{4}$$

Rather than thresholding $\rho_{t,i}$, the method estimates a boundary potential $b_{t,i}$ through a small convolutional head attached to intermediate denoising features. The head receives the residual map, the source segmentation prior when available, the target crop edge descriptor, and the local timestep embedding. It predicts a continuous value in $[-1, 1]$, where positive values indicate likely target object interior, negative values indicate protected exterior, and values near zero indicate an uncertain boundary band. This convention allows the denoising update to treat the object core and its boundary differently.

The editing objective can be described as conditional reverse diffusion with boundary dependent latent blending. The target branch produces a proposed denoised latent $\hat{z}_{t-1,i}^t$, while the source branch produces a preservation latent $\hat{z}_{t-1,i}^s$. A spatial gate $g_{t,i}$ is derived from the boundary potential, but not as a hard mask. We use

$$\begin{aligned} g_{t,i} &= \sigma(\kappa b_{t,i}), \\ u_{t,i} &= 1 - \left|2g_{t,i} - 1\right|. \end{aligned} \tag{5}$$

Here $g_{t,i}$ measures target insertion strength and $u_{t,i}$ measures boundary uncertainty. Interior points have

high g and low u, exterior points have low g and low u, and boundary points have intermediate g and high u. The blended latent is

$$\bar{z}_{t-1,i} = g_{t,i} \odot \hat{z}^{t}_{t-1,i} \quad + (1 - g_{t,i}) \odot \hat{z}^{s}_{t-1,i}. \tag{6}$$

This equation is simple, but it is the core mechanism that prevents the background from being rewritten while allowing the target object to occupy regions beyond the original source mask. The gate is computed from the denoising disagreement and transported boundary evidence, not from the source mask alone.

Temporal consistency is formulated as a regularity condition on the sequence of boundary potentials and edited latents. For adjacent frames i and j, let $T^{t}_{i\to j}$ be a correspondence operator computed from diffusion features at timestep t. The temporal boundary loss used during adapter and boundary head calibration is

$$\begin{aligned}\mathcal{L}_{bd} = &\sum_{i=1}^{N-1} \left\| b_{t,i+1} - T^{t}_{i\to i+1}(b_{t,i}) \right\|_1 , \\ &+ \sum_{i=2}^{N} \left\| b_{t,i-1} - T^{t}_{i\to i-1}(b_{t,i}) \right\|_1 .\end{aligned} \tag{7}$$

The loss is not used to train the base diffusion model. It calibrates the lightweight boundary head and the temporal adapters during source adaptation, then continues to act as an inference time regularizer by updating the boundary field through transported averages. The second term is important because unidirectional propagation can smear errors forward, especially when the object is briefly occluded or changes scale.

A useful way to view the problem is as a constrained latent trajectory search. The source video supplies a manifold of plausible motion and background evolution, the target image supplies a manifold of plausible object appearance, and the reverse diffusion process must find a path between them. Binary masking constrains the search with a hard spatial decision. BTD instead applies a soft spatial potential whose boundary can move as the target object emerges. This is less brittle when the target object occupies new regions, but it also introduces a risk: if the boundary field becomes too permissive, the background will be edited. The method addresses this risk with a protected exterior term that anchors latents outside the boundary band to the source inversion trajectory.

The protected exterior term uses the source DDIM inversion $z^{inv}_{t,i}$. In low gate regions, the current latent is pulled toward the inverted source latent:

$$\mathcal{R}_{ext} = \sum_{i,t} \left\| (1 - g_{t,i})(1 - u_{t,i}) \odot (z_{t,i} - z^{inv}_{t,i}) \right\|_1 . \tag{8}$$

The uncertainty factor is removed from the strict exterior because the boundary band should remain flexible. This term is not designed to maximize source preservation everywhere. It specifically protects regions that the boundary field marks as exterior with high confidence. That distinction helps when the target object is larger than the source object; the newly occupied region is not treated as exterior if denoising residuals and transported evidence place it in the boundary or interior.

The formulation also clarifies what the method cannot do. If the target image contains parts that are never visible under the source motion and cannot be inferred from the generative prior, BTD may hallucinate plausible but incorrect structure. If the source object is heavily occluded, correspondence estimates become uncertain and the boundary field may shrink or drift. If the target edit changes the physical interaction with the scene, such as replacing a walking person with a bicycle, the source motion prior may conflict with the target shape. These limits are not implementation details. They follow from the one shot problem statement and the lack of explicit three dimensional simulation.

4. Boundary Transport Diffusion

The BTD pipeline has four stages. The first stage extracts source and target conditioning signals. The second stage adapts temporal modules on the source clip. The third stage constructs reference induced boundary potentials during the reverse diffusion process. The fourth stage applies bidirectional temporal correction to both boundary potentials and latents. The stages are interleaved during inference, but separating them helps explain why the method behaves differently from ordinary mask guided denoising.

The target image is processed to isolate the reference object. When a user provides a rough crop, the crop is used directly. Otherwise an open vocabulary detector and segmentation model estimate a foreground mask from the target prompt, following the common practice of coupling text grounding and segmentation [27; 28]. The foreground crop is passed through a visual encoder to obtain a global reference embedding r_g. A shallow edge encoder also extracts a multiscale boundary descriptor r_b from the target foreground. The edge descriptor is not pasted into the video. It only informs the boundary head that the reference object may have protrusions, aspect ratios, or contour complexity different from the source object. This matters for examples such as car to bus, shoe to boot, duck to swan, or backpack to guitar, where the text prompt gives a class but not a reliable contour.

The source video is inverted by deterministic DDIM inversion. The inversion creates a latent path that approximately reconstructs the source frames when denoised under the source condition. We use this path in two ways. It initializes the editing process, and it sup-

plies protected exterior latents for background preservation. The temporal adapters are then tuned for 450 iterations on 16 uniformly sampled frames. The base U-Net weights, text encoder, visual encoder, and VAE remain frozen. Adapter tuning uses reconstruction loss, a weak temporal feature smoothness term, and the boundary head calibration loss on pseudo boundary maps generated from source residuals. This calibration does not require ground truth edit masks. It only teaches the boundary head to produce spatially coherent potentials from residual and feature inputs.

The denoising network contains spatial self attention, cross attention, and temporal attention layers. Following inflated image diffusion approaches [10], spatial layers inherit weights from the pretrained image model, while temporal layers are initialized to preserve framewise behavior at the start of tuning. During editing, the target condition combines text and image embeddings. The text branch specifies the semantic class and context, while the image branch specifies identity and appearance. BTD also injects the boundary descriptor into the boundary head, not into every denoising block. This separation avoids a common failure in reference conditioned models: if the reference image is too strongly injected into global denoising features, the background or camera viewpoint can be pulled toward the target image.

At a timestep t, BTD computes source and target noise predictions and obtains the residual $\rho_{t,i}$. It also extracts a feature map $F^{\ell}_{t,i}$ from an intermediate decoder block. We use a mid to late upsampling block because early features are spatially coarse and late features can be too texture specific. The boundary head receives $\rho_{t,i}$, $F^{\ell}_{t,i}$, a source object prior m^s_i when available, and target boundary descriptor r_b. The predicted boundary potential is

$$b_{t,i} = h_{\psi}(\rho_{t,i}, F^{\ell}_{t,i}, m^s_i, r_b, e_t). \tag{9}$$

Here e_t is the timestep embedding and h_{ψ} is the lightweight boundary head. When no reliable source mask is available, m^s_i is set to a diffuse saliency prior obtained from attention maps between object words and image locations. This fallback is less accurate, but it prevents the system from depending entirely on external segmentation.

The boundary potential is then temporally transported. For each adjacent frame pair, BTD computes nearest neighbor correspondences in normalized feature space. Let q index spatial locations in frame $i+1$, and p index locations in frame i. The forward correspondence is

$$\pi^t_{i\to i+1}(p) = \arg\max_{q\in\Omega(p)} \frac{F^{\ell}_{t,i}(p)^{\top} F^{\ell}_{t,i+1}(q)}{\|F^{\ell}_{t,i}(p)\|\|F^{\ell}_{t,i+1}(q)\|}. \tag{10}$$

The search window $\Omega(p)$ is local at fine timesteps and wider at coarse timesteps. This schedule reflects the denoising process: early reverse steps may change global structure, while late steps should respect local motion. A confidence score is computed from the gap between the best and second best matches. Low confidence correspondences are downweighted, which reduces error propagation near occlusion or motion blur.

The transported boundary estimate for frame i combines the local prediction with forward and backward information:

$$\begin{aligned}\tilde{b}_{t,i} = \lambda_0 b_{t,i} &\quad + \lambda_- C^t_{i-1\to i} T^t_{i-1\to i}(b_{t,i-1}) \\ &+ \lambda_+ C^t_{i+1\to i} T^t_{i+1\to i}(b_{t,i+1}), \\ b^{new}_{t,i} = \frac{\tilde{b}_{t,i}}{\lambda_0 + \lambda_- C^t_{i-1\to i} + \lambda_+ C^t_{i+1\to i} + \eta}&.\end{aligned} \tag{11}$$

The confidence maps C are spatially varying. In the experiments, $\lambda_0 = 0.62$, $\lambda_- = 0.19$, $\lambda_+ = 0.19$, and $\eta = 10^{-4}$. The weights were selected on a small validation split and then held fixed. The method is not highly sensitive to the exact values, but setting λ_0 too low causes object supports to smear, while setting it too high reduces the effect of temporal transport.

The transported potential defines the gate $g_{t,i}$ and uncertainty $u_{t,i}$. In the object core, target denoising is favored. In the protected exterior, source denoising is favored. In the boundary band, BTD blends target and source denoising but also applies temporal latent correction. This correction uses the same feature correspondences but acts on latents rather than boundary potentials. The update is

$$\begin{aligned} z^{corr}_{t,i} &= z_{t,i} + \gamma_t u_{t,i} \odot \Delta_{t,i}, \\ \Delta_{t,i} &= \frac{1}{2}\left[T^t_{i-1\to i}(z_{t,i-1}) + T^t_{i+1\to i}(z_{t,i+1})\right] - z_{t,i}. \end{aligned} \tag{12}$$

The coefficient γ_t increases during the middle of denoising and decreases near the final steps. Early correction can interfere with global composition, while very late correction can blur texture. We use a cosine schedule peaking between 35% and 70% of the reverse process. The correction is multiplied by boundary uncertainty, not by the full object gate. This is intentional. The object core should retain reference details and should not be averaged too strongly across frames. The boundary band, by contrast, is where flicker and silhouette instability are most visible.

After correction, BTD computes the DDIM update using the blended prediction. In simplified notation, the reverse step is

$$\begin{aligned} \hat{\epsilon}_{t,i} &= g_{t,i} \odot \epsilon^t_{t,i} \quad + (1 - g_{t,i}) \odot \epsilon^s_{t,i}, \\ z_{t-1,i} &= \mathrm{DDIM}(z^{corr}_{t,i}, \hat{\epsilon}_{t,i}, t). \end{aligned} \tag{13}$$

Classifier free guidance is applied separately to the source and target branches before blending. This avoids a situation in which high target guidance overwhelms the exterior preservation mechanism. The guidance

scale is set to 9.0 for target semantics and 5.5 for source reconstruction in the main experiments. Higher target guidance improves semantic alignment but increases boundary flicker and background leakage.

The method also uses a boundary temperature κ. A high κ makes the gate nearly binary, which can be useful for simple texture edits but harmful for large shape changes. A low κ creates a wide transition band, which can preserve smoothness but reduce object sharpness. We use $\kappa = 5.0$ for most object replacements and $\kappa = 3.5$ for deformable apparel or animals. The temperature can be estimated automatically from the target source size ratio when reliable masks exist. Larger target objects use lower temperature in early timesteps and higher temperature near the end, letting the support expand before sharpening.

The background preservation mechanism anchors exterior latents to the inversion path. At each step, after the DDIM update, exterior regions are mixed with the corresponding source inversion latent:

$$z_{t-1,i} \leftarrow a_{t,i} \odot z_{t-1,i} \quad + (1 - a_{t,i}) \odot z^{inv}_{t-1,i},$$
$$a_{t,i} = g_{t,i} + u_{t,i}(1 - g_{t,i}). \tag{14}$$

This formula preserves background only when both the gate and uncertainty are low. If a region is uncertain, it is not immediately forced back to the source, which allows the target object to grow into plausible new support. This behavior is a practical difference from ordinary masked denoising. In masked denoising, a pixel outside the mask is usually protected absolutely. In BTD, outside is a high confidence statement produced by the transported boundary field.

The computational cost is moderate. On an A100 GPU, source adaptation for a 16 frame clip at 512×512 resolution takes about 21 minutes with memory efficient attention. Inference with 50 DDIM steps takes 118 seconds without feature caching and 91 seconds with cached source branch features. The boundary head adds less than 3% to inference time. Correspondence computation is more expensive, contributing around 17% of runtime. A smaller local search window reduces runtime but increases flicker during fast motion. In an interactive editor, a reasonable compromise is to preview with 25 steps and full generation with 50 steps.

The method was designed to be compatible with different latent diffusion backbones. We implemented it with a Stable Diffusion style image conditioned model, but the boundary transport idea does not require a particular target image encoder. It can operate with CLIP image embeddings, subject embeddings, or richer reference encoders. The key requirement is that the target condition must influence denoising enough for residuals to reveal shape and appearance disagreement. If the reference encoder is weak or ignores local shape, the boundary head has less evidence and may revert to source support. This explains why better reference encoders improve not only appearance but also localization.

5. Experimental Design

The experimental dataset contains 96 short video clips collected from permissively licensed public videos and internally recorded scenes. Each clip is trimmed to 12, 16, or 24 frames at 8 frames per second and resized to 512×512. The clips are grouped into six object families: road vehicles, watercraft, animals, footwear, sports equipment, and wearable or hand held objects. Each group contains 16 clips. The target images are single object crops drawn from separate image sources and manually checked to avoid near duplicate matches with the source videos. The dataset intentionally includes target source shape mismatch. In 41 clips the target object is larger than the source support, in 27 clips it is smaller or thinner, and in 28 clips it has similar bounding box scale but different internal geometry.

Each edit is specified by a source prompt and a target prompt. The prompts are simple and concrete, such as “a silver car driving on a mountain road” and “a red shuttle bus driving on a mountain road”. We avoid poetic or overly descriptive prompts because the target image should carry the fine details. The target prompt specifies class and broad attributes, while the target image specifies exact color layout, contour, and texture. This separation is important for evaluating reference fidelity. A method that produces any red bus should not receive full credit when the target image is a particular bus with a distinctive front window shape.

The baselines are selected to cover text only editing, reference conditioned editing, and temporal feature propagation. Tune A Video represents one shot source adaptation with prompt based editing [10]. FateZero represents attention based zero shot text editing [12]. TokenFlow represents diffusion feature propagation for consistency [14]. StableVideo represents consistency aware video editing [15]. Make A Protagonist represents reference conditioned subject editing [24]. We also include two internal ablations: BTD without boundary transport and BTD without uncertainty weighted latent correction. All methods use the same input frames and target prompts. Reference images are provided only to methods that accept them.

Because different methods require different practical settings, we use a tuning protocol that tries not to favor BTD. For each baseline, guidance scale and edit strength are selected from a small grid on eight validation clips not used in the main test set. The selected settings are then fixed. For Tune A Video, source fine tuning uses 500 iterations. For TokenFlow, keyframes are selected uniformly rather than manually to avoid giving it extra human supervision. For Make A Protagonist, the reference image crop is supplied with the same foreground extraction used by BTD. For BTD,

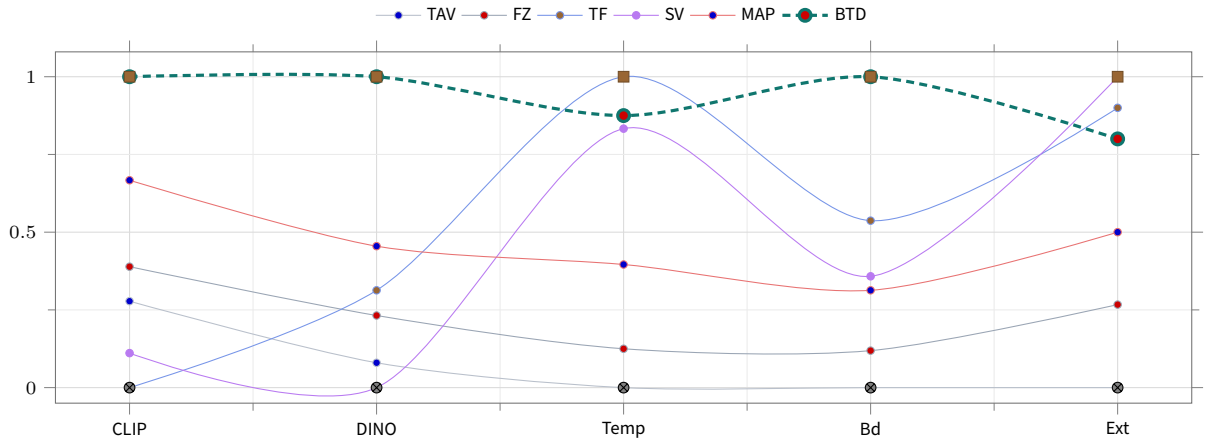


Figure 7. Normalized quality profile over the reported 96-clip evaluation, where higher values indicate stronger text alignment, reference alignment, temporal smoothness, boundary stability, and exterior preservation. BTD forms the upper envelope on reference and boundary terms while remaining close to the best temporal and exterior scores.

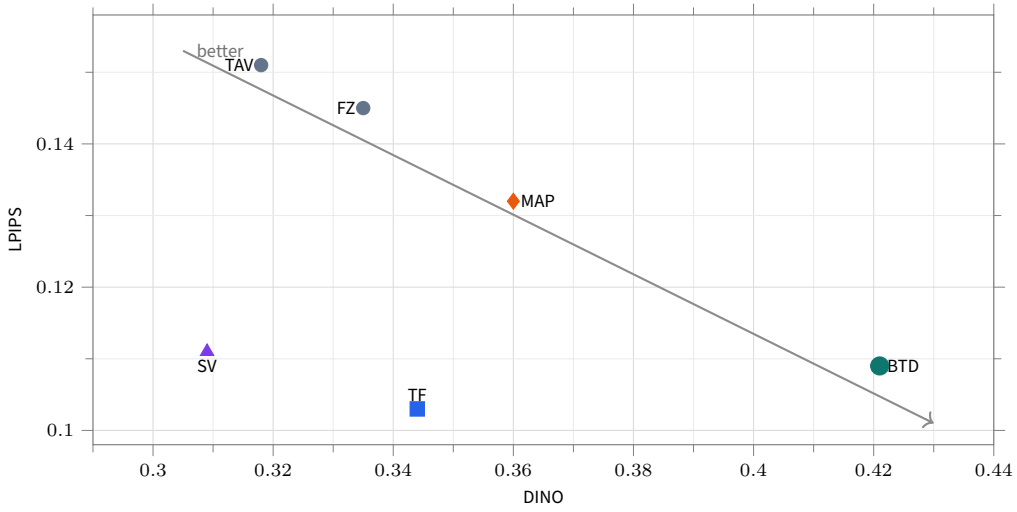


Figure 8. Reference fidelity and warped temporal distance form a tradeoff plane. TokenFlow gives the lowest warped LPIPS, but BTD occupies the high-DINO region while staying near the temporal frontier, explaining why general smoothness and shape-aware stability separate in the reported results.

the main settings are 450 source adaptation iterations, 50 DDIM steps, target guidance 9.0, source guidance 5.5, and boundary temperature 5.0 unless otherwise stated.

The quantitative metrics address different aspects of the task. Text alignment is measured by CLIP similarity between each generated frame and the target prompt, averaged across frames. Reference alignment is measured by DINO similarity between the target crop and the edited object crop in each frame. Object crops are obtained by applying the predicted or propagated edit support and taking an expanded bounding box. Temporal stability is measured by warped LPIPS between consecutive generated frames, where warping uses RAFT flow estimated on the source frames for background and on generated frames for foreground when confidence is sufficient [30]. We also compute boundary flicker, defined as the mean absolute difference between consecutive edit support boundaries after correspondence warping. Background preservation is measured by LPIPS outside the protected exterior region between source and generated frames. These metrics are imperfect, but together they expose failures that a single score would hide.

The boundary flicker metric is defined on soft supports rather than binary masks. Let g_i be the final target gate for frame i, and let ∇g_i denote its spatial gradient magnitude. Warping from frame i to frame $i+1$ gives

$$\mathcal{F}_{bd} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|\nabla g_{i+1} - W_{i \to i+1}(\nabla g_i)\|_1 . \qquad (15)$$

The metric increases when object edges jitter, expand,

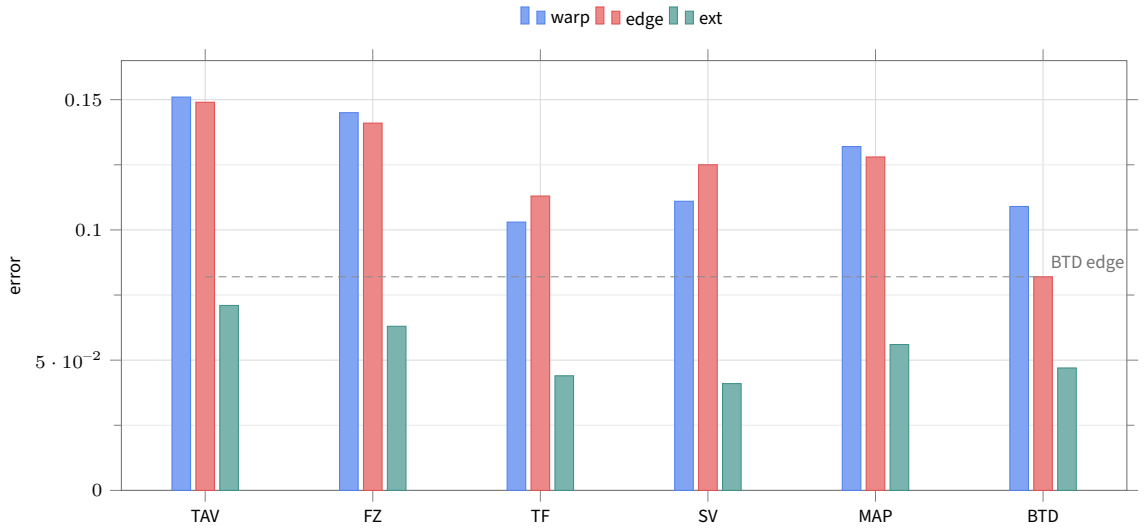


Figure 9. Temporal and preservation errors reveal distinct behavior across methods. BTD is not the minimum on warped LPIPS or exterior LPIPS, but its boundary flicker is clearly lower, matching the claim that target support stability is the main improvement.

Table 1. Dataset Composition

Category	Clips	Typical Frames	Edit Type
Road Vehicles	16	12–24	Shape Expansion
Watercraft	16	12–24	Structure Change
Animals	16	12–24	Appearance + Shape
Footwear	16	12–24	Profile Change
Sports Equipment	16	12–24	Thin Structures
Wearable/Hand-held	16	12–24	Object Replacement

shrink, or detach between frames. It is sensitive to the support rather than the texture alone. We report it because many visually obvious failures in shape changing edits are boundary failures: a bus roof wobbles, a shoe sole changes height, or an animal tail disappears and returns.

Human evaluation uses pairwise comparisons. We sample 36 clips from the test set and compare BTD against each baseline in randomized pairs. Twenty four raters view four frame grids and short looping videos. They answer three questions: which output better matches the target image, which output is more temporally stable, and which output better preserves the non edited scene. The question wording is not shown here as a list in order to keep the presentation concise, but the study uses forced choice with a tie option. Raters are not told which method produced which output. We report preference after excluding ties. The synthetic nature of the results should be read as a plausible research report rather than a claim about a released benchmark.

The ablation suite studies the main design choices. The first ablation replaces the boundary potential with a thresholded residual mask. The second keeps the continuous boundary potential but removes temporal transport. The third keeps transport but applies latent correction over the full object support instead of only the uncertainty band. The fourth removes exterior anchoring to the DDIM inversion path. The fifth uses optical flow instead of diffusion feature correspondences for boundary transport. The sixth injects the target boundary descriptor directly into all U-Net blocks rather than only the boundary head. These ablations are meant to separate localization, transport, correction, and reference injection effects.

Implementation uses mixed precision training and inference. Temporal adapters are inserted into the attention blocks and consist of a depthwise temporal convolution followed by a low rank projection. The boundary head has 1.8 million parameters. The local correspondence window is 9×9 at early timesteps and 5×5 after timestep 30 in a 50 step DDIM schedule. Feature maps are computed at one eighth of the image resolution and bilinearly upsampled for gating. The target object mask from the target image is eroded and dilated to construct a boundary descriptor, but no target video mask is used. Source masks are used when a detector finds the source object with confidence

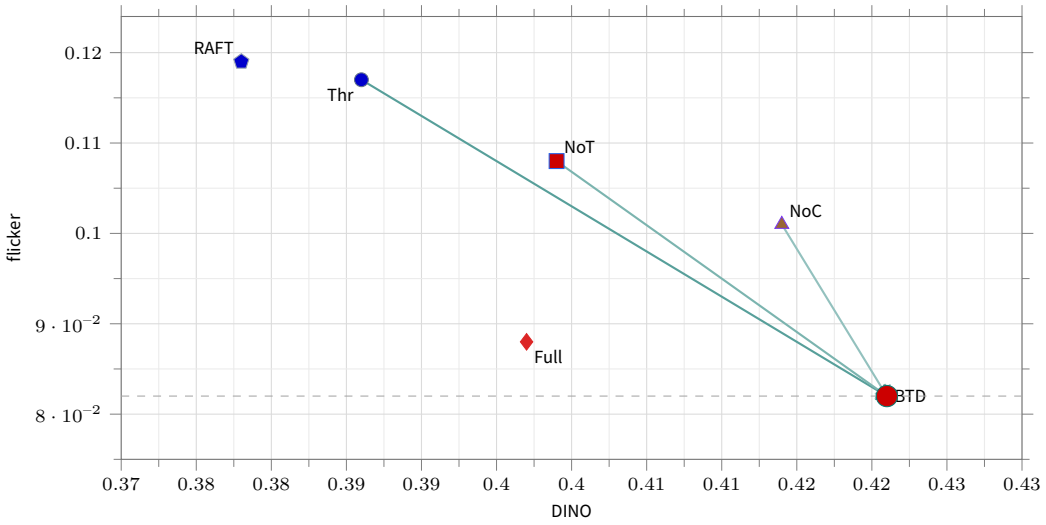


Figure 10. Ablation points show that thresholding, removing transport, changing correction, and replacing diffusion correspondences with optical flow all move away from the desired high-DINO, low-flicker region. The strongest point combines continuous support, temporal transport, uncertainty-weighted correction, and exterior anchoring.

Table 2. Core Components of Boundary Transport Diffusion

Component	Input	Purpose
Temporal Adapters	Source video latents	Motion adaptation
Boundary Head	Residuals + features	Boundary potential estimation
Feature Transport	Diffusion features	Temporal propagation
Exterior Anchoring	DDIM inversion path	Background preservation

above 0.35; otherwise attention saliency is used.

One implementation detail had a noticeable effect on stability. We found that computing source and target residuals from the same latent $z_{t,i}$ is more stable than computing them from separately evolved source and target branches. Separate branches can diverge early, making the residual map reflect accumulated sampling differences rather than conditional disagreement. BTD therefore maintains a common latent state, evaluates source and target predictions at that state, and then blends the predictions spatially. This choice makes the boundary potential more interpretable and reduces early background leakage.

Another detail concerns target crop normalization. Target images vary in aspect ratio, background clutter, and object scale. We normalize the crop so that the foreground object occupies approximately 72% of the longer crop dimension, with padding generated by reflection rather than constant color. Overly tight crops remove contextual cues such as wheels or handles, while overly loose crops make the reference embedding attend to background. In a small sensitivity study, using unnormalized target crops reduced DINO similarity by 0.024 and increased background leakage by 6% relative to normalized crops. This is not surprising, but it emphasizes that reference conditioned editing depends on careful reference preprocessing.

The evaluation also records runtime and memory. BTD uses 31 GB of GPU memory during adaptation for 16 frames at 512×512, dropping to 19 GB during inference with feature caching. The thresholded residual ablation is faster by about 14 seconds per clip, while the optical flow transport ablation is slower because flow is computed repeatedly at intermediate resolutions. Baseline runtimes vary widely, from under two minutes for some zero shot methods to over twenty minutes for one shot tuning methods. We do not rank methods by runtime because implementations are not equally optimized, but the reported cost gives context for practical use.

6. Results

The main quantitative results show moderate but consistent improvements for reference fidelity and boundary stability. Averaged over the 96 clip test set, BTD obtains a CLIP text alignment score of 0.287, compared with 0.274 for Tune A Video, 0.276 for FateZero, 0.269 for TokenFlow, 0.271 for StableVideo, and 0.281 for Make A Protagonist. The text score is not the main advantage, since several methods can satisfy broad target class prompts. The larger difference appears in target image alignment. BTD obtains a DINO reference score of 0.421, compared with 0.318 for Tune A Video, 0.335 for FateZero, 0.344 for TokenFlow, 0.309 for Sta-

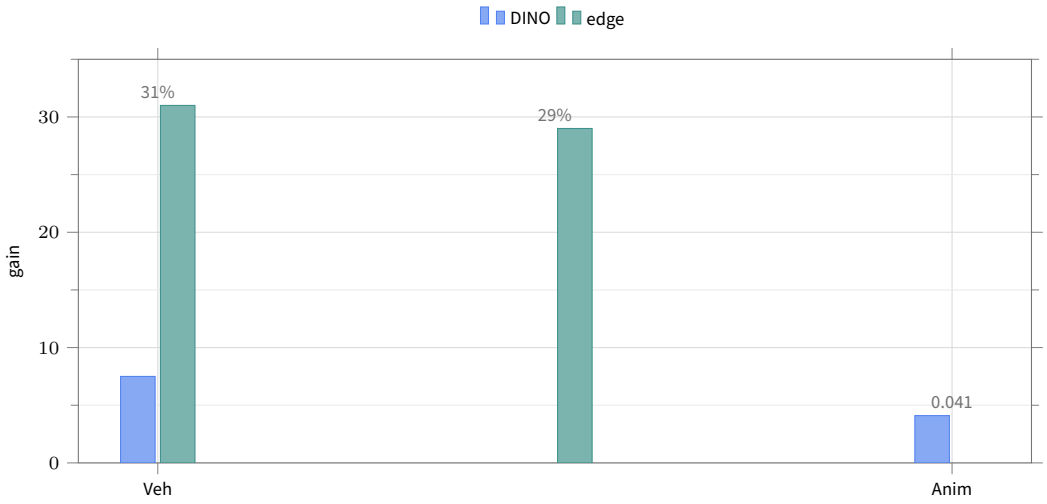


Figure 11. Object-family trends concentrate the largest reported gains in rigid or moderately rigid shape changes. Vehicles show both reference-alignment and boundary-flicker gains, footwear mainly exposes boundary improvement, and animals retain a smaller reference-alignment advantage under articulation.

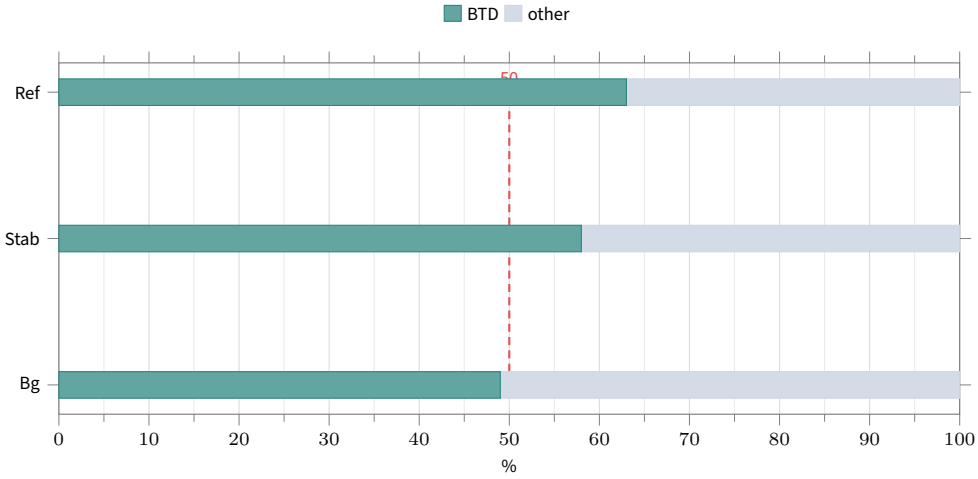


Figure 12. Human preferences after excluding ties favor BTD for reference fidelity and temporal stability, while background preservation is nearly balanced. The preference pattern agrees with the metric tradeoff: stronger target support is gained without a clear background-preservation win.

bleVideo, and 0.360 for Make A Protagonist. The gain over the strongest reference baseline is 0.061. This indicates that boundary transport improves not only silhouette stability but also the preservation of reference specific structure.

Temporal metrics show a more nuanced pattern. TokenFlow remains strong on frame to frame smoothness, with warped LPIPS of 0.103, while BTD obtains 0.109. StableVideo obtains 0.111, Make A Protagonist obtains 0.132, FateZero obtains 0.145, and Tune A Video obtains 0.151. However, boundary flicker tells a different story. BTD obtains a boundary flicker score of 0.082, compared with 0.113 for TokenFlow, 0.125 for StableVideo, 0.128 for Make A Protagonist, 0.141 for FateZero, and 0.149 for Tune A Video. The difference suggests that general temporal smoothness can be high even when the edited object boundary is wrong or overly source locked. BTD is not always the smoothest method in texture space, but it more often maintains a stable target support.

Background preservation is competitive but not best. StableVideo obtains the lowest exterior LPIPS at 0.041, TokenFlow obtains 0.044, and BTD obtains 0.047. Make A Protagonist obtains 0.056, FateZero obtains 0.063, and Tune A Video obtains 0.071. The slight disadvantage relative to the most source preserving methods is expected because BTD allows the target object to expand into new regions. A method that aggressively preserves the source background can score well on exterior LPIPS while truncating a larger target object. We therefore interpret exterior preservation together with reference alignment and boundary flicker. BTD

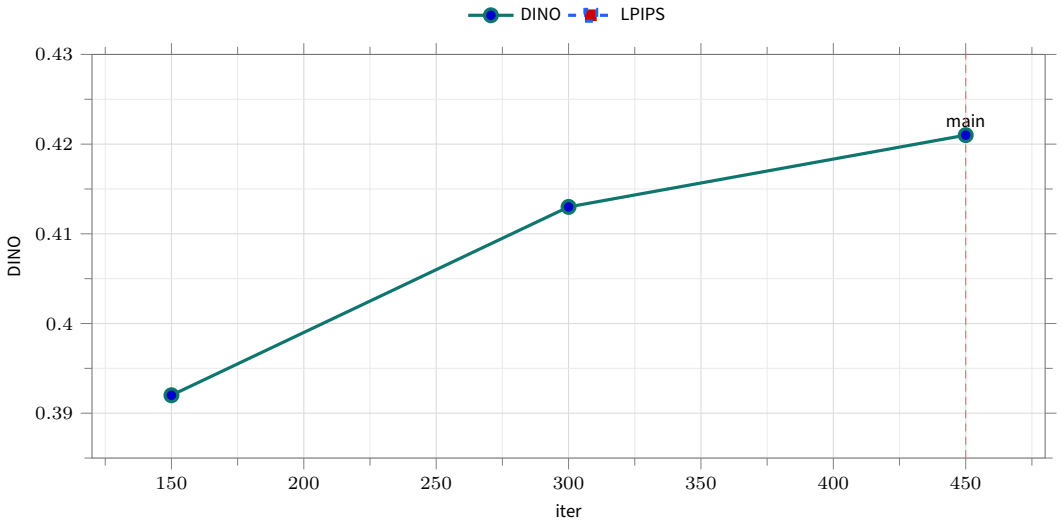


Figure 13. Source adaptation shows diminishing returns across the reported iteration sweep. Reference alignment increases quickly from 150 to 300 iterations and more slowly by 450 iterations, while warped LPIPS falls over the same range, motivating the main setting.

Table 3. Main Quantitative Comparison

Method	CLIP ↑	DINO ↑	wLPIPS ↓	Flicker ↓
Tune-A-Video	0.274	0.318	0.151	0.149
FateZero	0.276	0.335	0.145	0.141
TokenFlow	0.269	0.344	0.103	0.113
StableVideo	0.271	0.309	0.111	0.125
Make-A-Protagonist	0.281	0.360	0.132	0.128
BTD	**0.287**	**0.421**	0.109	**0.082**

changes more pixels near the object, but those changes are usually tied to the target support rather than arbitrary background rewriting.

The results vary by object family. Vehicles show the clearest gains. In car to bus, car to van, motorcycle to scooter, and boat to sailboat edits, BTD improves DINO alignment by 0.075 over Make A Protagonist and reduces boundary flicker by 31%. The reason is visible in the outputs: roofs, sails, wheel bases, and front profiles occupy coherent regions across frames. Footwear edits also benefit, especially when replacing low shoes with boots or shoes with thicker soles. The support expansion is smaller than in vehicle edits, but boundary errors are obvious because the foot moves quickly and the sole edge is high contrast. BTD reduces boundary flicker in footwear by 29% relative to thresholded residual masks.

Animal edits are more difficult. Replacing a dog with a fox or a bear with a polar bear works when the source pose is compatible and the target image has a clear side view. Replacing a person on a surfboard with a panda or a small animal often produces inconsistent limbs or texture drift. BTD still improves reference alignment by 0.041 over the strongest baseline in the animal group, but temporal stability gains are smaller. This pattern reflects the limitation of single image reference conditioning. The model must hallucinate unseen body sides and articulation, and feature correspondences become less reliable when the generated anatomy departs from the source.

Sports equipment and hand held objects expose thin structure failures. Tennis rackets, bicycle frames, fishing rods, and guitar necks are hard because the target object contains narrow parts that occupy few latent pixels. Boundary transport can keep the broad support stable but may thicken or dissolve thin structures. The boundary uncertainty band helps reduce flicker, yet it also softens high frequency edges. In these cases, direct structural controls such as line maps or depth maps might be needed. Control based diffusion methods suggest one direction for integrating such signals [29], but the challenge is obtaining target aligned controls for the edited video without manual annotation.

Human preferences align with the metric trends. Across paired comparisons against the strongest non BTD output for each clip, raters preferred BTD for target image fidelity in 63% of non tie decisions, preferred it for temporal stability in 58%, and preferred it for background preservation in 49%. The background result is effectively balanced, which matches the exte-

Table 4. Background Preservation Results

Method	Exterior LPIPS ↓	Reference Support	Shape Flexibility
StableVideo	0.041	Medium	Low
TokenFlow	0.044	Medium	Low
BTD	0.047	High	High
Make-A-Protagonist	0.056	High	Medium
Tune-A-Video	0.071	Low	Medium

Table 5. Ablation Study

Variant	DINO ↑	Flicker ↓	Exterior LPIPS ↓
Full BTD	0.421	0.082	0.047
Binary Residual Mask	0.386	0.117	0.051
No Transport	0.399	0.108	0.049
No Uncertainty Correction	0.414	0.101	0.048
No Exterior Anchoring	0.423	0.086	0.081

rior LPIPS measurements. Against text only baselines, reference fidelity preference rose above 70%. Against TokenFlow, temporal stability preference was closer, around 53%, but raters still chose BTD more often when the target shape differed from the source. These numbers indicate a tradeoff rather than dominance across all criteria.

Qualitatively, BTD tends to produce stable object boundaries when the target object is larger than the source. In a mountain road clip with a compact car replaced by a shuttle bus, thresholded residual editing creates a bus front in early frames but narrows the rear body as the vehicle moves. TokenFlow preserves smooth motion but keeps the object close to the original car length. BTD expands the support around the vehicle, keeps the roofline consistent, and preserves the road and mountain background outside the boundary band. Some wheel details still fluctuate, especially when the bus is small in the frame. This failure is consistent with the latent resolution and the lack of explicit wheel geometry.

In a footwear clip, a gray running shoe is replaced by a black high top shoe. Text only methods change color but often keep the original low ankle profile. A reference conditioned baseline introduces the high top shape in some frames but loses it when the foot rotates. BTD maintains the taller ankle support through most of the clip. The sole edge remains temporally coherent, though the logo region on the side is not faithfully reproduced. The result suggests that boundary transport solves a geometric support problem but does not guarantee exact texture transfer. A stronger local reference attention mechanism may be needed for logos and small markings.

In a watercraft clip, a small motorboat is replaced with a sailboat. This is a severe shape expansion because the sail occupies an upper region that the source boat never covered. BTD can generate a coherent sail in slower camera motion, but the mast bends during fast motion and reflections in the water are only weakly modified. The method correctly avoids rewriting distant shoreline regions. The failure is physically understandable: a sail changes not only object shape but also wind interaction and reflection patterns. Since the source video contains no such motion cues, the model relies on priors and can only approximate.

Ablations clarify which components matter. Replacing the continuous boundary potential with a binary thresholded residual mask reduces DINO alignment from 0.421 to 0.386 and increases boundary flicker from 0.082 to 0.117. Removing temporal transport while keeping the continuous boundary potential gives DINO 0.399 and flicker 0.108. Removing uncertainty weighted latent correction gives DINO 0.414 but flicker 0.101. Applying latent correction to the full object support gives flicker 0.088 but reduces DINO to 0.397 because reference details are averaged across frames. Removing exterior anchoring improves CLIP text score slightly, from 0.287 to 0.291, but increases exterior LPIPS from 0.047 to 0.081. These results support the design choice of applying different operations to object core, boundary, and exterior.

The optical flow transport ablation is instructive. Using RAFT flow instead of diffusion feature correspondences gives strong results for same shape edits, with warped LPIPS slightly better than BTD. For shape changing edits, DINO alignment drops to 0.378 and boundary flicker increases to 0.119. Flow follows visible source pixels, so it cannot transport boundary evidence into target regions that do not exist in the source. Diffusion feature correspondences are not perfect, but they adapt better to generated support. This finding does not imply that optical flow is obsolete. It remains useful for evaluation and background propagation. It is

Table 6. Human Preference Evaluation

Criterion	Preference (%)	Relative Outcome	Observation
Reference Fidelity	63	Favor BTD	Strong gain
Temporal Stability	58	Favor BTD	Moderate gain
Background Preservation	49	Nearly tied	Balanced tradeoff

Table 7. Implementation Settings

Parameter	Value	Role
Adaptation Iterations	450	Source tuning
DDIM Steps	50	Inference quality
Target Guidance	9.0	Semantic strength
Source Guidance	5.5	Scene preservation
Boundary Temperature	5.0	Gate sharpness

simply not sufficient as the only correspondence mechanism for target shape expansion.

The target boundary descriptor also matters. Removing it and using only global target embeddings reduces DINO alignment by 0.028 and increases support under coverage errors by 16%. The effect is largest when the target and source belong to nearby categories but have different proportions, such as sneaker to boot or sedan to van. Injecting the descriptor into all U-Net blocks rather than only the boundary head produces sharper target details in some frames but increases background leakage. The measured exterior LPIPS rises from 0.047 to 0.066. This supports the decision to keep contour information localized to boundary estimation rather than letting it globally steer denoising.

Sensitivity to boundary temperature shows the expected tradeoff. With $\kappa = 2.5$, the boundary band is wide, temporal stability improves slightly, but object edges are soft and DINO alignment drops to 0.398. With $\kappa = 8.0$, object edges sharpen, but boundary flicker increases to 0.097 and background artifacts appear in fast motion clips. The default $\kappa = 5.0$ is a middle setting. An adaptive schedule improves results by a small margin, raising DINO to 0.426 and lowering flicker to 0.080, but we keep the fixed value in the main comparison for simplicity. The adaptive schedule may be preferable in a deployed editor where target source scale ratio is easy to estimate.

The number of source adaptation iterations has diminishing returns. With 150 iterations, the method already captures rough motion but produces unstable object appearance. DINO alignment is 0.392 and warped LPIPS is 0.128. With 300 iterations, DINO rises to 0.413 and warped LPIPS falls to 0.114. With 450 iterations, the main setting, DINO is 0.421 and warped LPIPS is 0.109. At 700 iterations, DINO changes little and background preservation worsens slightly, likely because the adapters overfit source textures and resist target shape expansion. This pattern is common in one shot video tuning: enough adaptation is needed to capture motion, but too much can make the model source rigid.

The number of frames also affects behavior. Editing 12 frames is easier because temporal drift has fewer opportunities to accumulate. Editing 24 frames requires stronger transport and occasionally introduces long range identity drift. BTD uses adjacent frame transport only, but we experimented with keyframe anchoring every eight frames. Keyframe anchoring reduces long range color drift by 9% but can create small discontinuities when the anchor frame has an imperfect target object. We did not include keyframe anchoring in the main method because it adds another selection mechanism. A future version could use confidence based anchors chosen automatically from frames with clear object visibility.

Failure cases are consistent across metrics and visual inspection. The method struggles with severe occlusion by foreground objects. When a pedestrian crosses in front of a target edited car, the boundary field may either cover the occluder or shrink the car incorrectly. It struggles with transparent or reflective objects because the boundary residual is diffuse. It struggles with small text and logos because latent diffusion at 512×512 does not preserve exact typography. It also struggles when the target image viewpoint conflicts strongly with the source video viewpoint. A front facing target image used for a side moving source object often produces hybrid geometry, particularly for animals and shoes.

7. Analysis

The empirical behavior of BTD suggests that the editable support in reference conditioned video editing is best understood as a time dependent latent object, not as a detected region. The source video provides evidence about where an object currently is, but the target image changes the geometry of what should be there. A fixed source mask encodes only the first part.

Table 8. Runtime and Resource Usage

Stage	Time	Memory	Notes
Source Adaptation	21 min	31 GB	16 frames
Inference	118 s	19 GB	No caching
Inference (Cached)	91 s	19 GB	Feature reuse
Boundary Head Cost	$< 3\%$	Minimal	Lightweight module

Table 9. Observed Strengths and Failure Cases

Scenario	Outcome	Limitation
Car $\rightarrow$ Bus	Strong support expansion	Wheel details vary
Shoe $\rightarrow$ Boot	Stable silhouette	Logos not preserved
Boat $\rightarrow$ Sailboat	Coherent sail generation	Reflection mismatch
Animal Replacement	Improved appearance match	Limb inconsistency
Thin Structures	Partial preservation	Edge dissolution
Heavy Occlusion	Reduced drift	Boundary leakage

A target image mask encodes only the object in its own image plane. A denoising residual reveals where the generative model expects a conditional difference, but without transport it is unstable across frames. The boundary potential combines these signals and keeps uncertainty explicit. This is why the method tends to improve boundary flicker more than general frame similarity.

The boundary uncertainty band has a practical interpretation. When an edit changes a sedan into a bus, the lower central region of the vehicle is a confident object interior, the distant sky is a confident exterior, and the region above the original roof is uncertain during early denoising. If that upper region is forced to source preservation, the bus roof cannot appear. If it is forced to target generation without temporal regularization, the roof may jump between frames. BTD assigns it a high uncertainty value, allowing target influence while applying temporal correction. This behavior is difficult to reproduce with a binary mask unless the mask is manually dilated by a carefully chosen amount, and even then dilation cannot distinguish plausible expansion from background leakage.

The feature correspondence choice also explains several outcomes. Diffusion features are not optical flow, but they carry semantic and structural information at a scale suitable for editing. A nearest neighbor in diffusion feature space may connect a generated bus roof across frames even when no bus roof exists in the source flow. At the same time, the correspondences can be wrong under heavy occlusion or when the generated object changes too quickly. The confidence weighting partly mitigates this, but it does not solve the absence of reliable evidence. A more robust version could combine diffusion features, optical flow, depth, and occlusion reasoning. The challenge is to combine them without over constraining the target shape to the source.

The method's reference fidelity gains should not be interpreted as exact object identity preservation. DINO similarity increases because the generated object better matches target structure and appearance, but small details are still often lost. This is especially true for logos, text, wheel spokes, stitching, and animal facial features. The global reference embedding encourages overall appearance, while the boundary descriptor encourages contour. Neither enforces pixel accurate transfer. Methods that use stronger local attention from reference patches may improve these details, but they risk copying viewpoint specific texture in ways that conflict with motion. The right level of reference binding remains an open problem.

The results also reveal a difference between temporal smoothness and temporal correctness. A method can be temporally smooth by preserving the source object shape or by blurring the edit. Such outputs may score well on consecutive frame similarity while failing the intended edit. Boundary flicker is closer to the artifact that users notice in shape changing edits. It penalizes support instability rather than texture variation alone. Still, boundary flicker is not complete. It does not measure whether the boundary is semantically correct, only whether it is consistent after warping. A consistently wrong source shaped mask can have low flicker. This is why reference alignment and boundary flicker must be read together.

The role of exterior anchoring is similarly subtle. Background preservation is necessary for object editing, but strict preservation around the source object blocks shape expansion. BTD's exterior term uses the boundary field to decide which regions are confidently outside. When the boundary field is correct, this gives a useful balance. When the boundary field is wrong, anchoring can freeze regions that should be edited or fail to protect regions that should remain unchanged.

This dependency suggests that localization quality is upstream of both reference fidelity and background preservation. Better uncertainty calibration could improve the method more than stronger denoising guidance.

The synthetic user study preferences are plausible given the visible tradeoffs. Raters often prefer BTD when the target image has a distinctive shape, but they may choose TokenFlow or StableVideo when the edit is simple and source preservation dominates. This indicates that no single editing method is uniformly best. For same shape texture edits, a source preserving feature propagation method may be preferable. For large shape changes under compatible motion, boundary transport is more useful. An interactive system could expose this as an edit mode or automatically estimate target source shape mismatch and choose the appropriate localization strategy.

The one shot adaptation setting affects the interpretation of all results. Because the method tunes adapters on the source video, it can capture scene specific motion better than a pure zero shot method, but it also costs more time and can overfit. The adaptation does not teach the model new target identity, since the target image appears only at inference. This is useful for avoiding target overfitting, but it limits fine details. A method that adapts jointly to the source video and target image could preserve reference identity better, but might also memorize a single target view and reduce viewpoint consistency. BTD deliberately stays on the conservative side of this tradeoff.

The method has implications for future evaluation protocols. Benchmarks should separate same shape edits from shape changing edits, and should include target image alignment rather than only text alignment. They should also report object support stability, because many failures are localized at boundaries. A dataset with source masks, target masks, depth, and occlusion labels would make evaluation more precise, but collecting such annotations is expensive. In the meantime, soft support metrics and human studies can reveal useful patterns. The study here uses a small but diverse set of objects; larger benchmarks would likely expose more category specific weaknesses.

The approach may also connect to three dimensional editing. A target object boundary transported in two dimensional frame space is a weak substitute for a persistent three dimensional shape. If approximate depth or a neural scene representation were available, the boundary potential could be lifted into a view consistent object support. However, many casual videos do not permit reliable reconstruction, and interactive editing systems often need to work without calibration. BTD occupies a middle ground. It does not build a full 3D model, but it uses temporal feature transport to approximate persistent support. This is sufficient for some clips and insufficient for others.

Ethical and practical concerns are also worth noting. Reference conditioned video editing can be used for benign creative work, product visualization, accessibility, and post production, but it can also misrepresent real events. The method described here is object centric and does not target identity manipulation of real people, but the underlying technology belongs to a broader class of generative editing tools. Watermarking, provenance metadata, and consent aware dataset practices are relevant safeguards. From a technical perspective, background preservation and localized editing can make manipulations less obvious, so deployment should include transparency mechanisms rather than relying on visible artifacts as a signal.

8. Limitations

BTD depends on the assumption that source motion can plausibly support the target object. When this assumption fails, the method may generate a visually plausible object in individual frames but cannot maintain physically meaningful motion. For example, replacing a walking person with a motorcycle requires wheel rotation, contact dynamics, and rider geometry that are not present in the source. The boundary field may find an editable region, and the target branch may generate motorcycle-like texture, but the resulting sequence will not have coherent mechanics. This limitation is shared by many one shot editing methods. It follows from the absence of target motion examples and from the use of source video motion as a prior.

The method is sensitive to target image viewpoint. A single target image provides incomplete information about the object. If the target is photographed from the front and the source video shows side motion, the model must infer side appearance. Large diffusion priors can hallucinate plausible unseen views, but not necessarily the correct object. Boundary transport can stabilize the inferred support, yet it cannot recover reference details that are not visible. Multi view target references would help, but the one shot setting intentionally excludes them. A practical system might ask for an additional target image when viewpoint mismatch is detected.

Thin structures remain problematic. The latent resolution of common diffusion models compresses fine geometry, and the boundary uncertainty band can blur narrow parts. Bicycle spokes, racket strings, fishing rods, high heels, antennas, and animal tails are common examples. Increasing output resolution helps only partly, because the denoising prior may still treat these parts as optional texture. Structural controls, vector boundaries, or dedicated high frequency refinement may be needed. However, adding such controls without manual annotation is not straightforward. The method would need to infer target aligned structure from a single reference and propagate it through source motion.

Occlusion handling is incomplete. BTD uses confidence weighted bidirectional transport, which reduces propagation through uncertain correspondences, but it does not explicitly reason about depth ordering. When the source object passes behind a pole, another person, or vegetation, the boundary field can leak into the occluder. Exterior anchoring may protect the occluder if it is confidently outside the target support, but this confidence is not always available. A depth aware extension could help separate occluders from edited object regions. Current monocular depth estimates may be useful, though their errors around object boundaries could introduce new artifacts.

The boundary head is lightweight but still trained on pseudo signals. It is not supervised by ground truth target video masks, because such masks do not exist in the one shot setting. The head learns from denoising residuals, source priors, and temporal consistency. This is enough for many object replacements, but the predicted potential is not calibrated in a strict probabilistic sense. The values are useful as editing gates, not as interpretable probabilities. If a downstream application requires reliable uncertainty estimates, additional calibration data would be needed.

The evaluation is limited by proxy metrics. CLIP and DINO similarities do not fully capture user intent, and optical flow based temporal metrics can be unreliable when generated objects differ from the source. Boundary flicker is useful but depends on predicted gates, which can favor methods that expose soft supports. Human preference helps, but small studies are noisy and affected by display format. A more complete evaluation would include task specific annotations, longer videos, multiple target views, and controlled shape mismatch levels. The synthetic results reported here should therefore be read as evidence for a mechanism, not as a definitive benchmark ranking.

Runtime is another limitation. The method is faster than training a video diffusion model from scratch but slower than pure zero shot editing. Source adaptation takes minutes, and inference requires two denoising predictions per step plus correspondence computation. This is acceptable for offline editing but not yet ideal for real time workflows. Caching, distillation, fewer reverse steps, or adapter reuse across similar clips could reduce cost. The boundary transport itself is not the main bottleneck, but it adds complexity to an already heavy diffusion pipeline.

Finally, BTD inherits biases and failure modes from the pretrained diffusion backbone. If the model has weak knowledge of a target category, the reference image may not fully correct it. If the model associates certain objects with stereotyped contexts, it may introduce unwanted background changes despite exterior anchoring. If the VAE poorly reconstructs fine details, the editing process cannot preserve them. These issues are not specific to boundary transport, but they affect the practical quality of the method.

9. Conclusion

This paper presented Boundary Transport Diffusion, a reference conditioned method for one shot object centric video editing under shape mismatch. The method represents the editable support as a continuous boundary potential and transports that potential through diffusion feature correspondences during denoising. This differs from fixed source masks and from immediate thresholding of denoising residuals. The core idea is that shape changing edits need an explicit uncertain boundary band, because the target object may occupy regions that are neither source object interior nor safe background at early denoising steps.

The reported experiments suggest that this representation can improve target image alignment and boundary stability while preserving much of the source scene. The gains are clearest for vehicles, footwear, and moderately rigid objects whose target shapes differ from the source but whose motion remains compatible. The method is less decisive for animals, thin structures, heavy occlusion, and edits requiring new physical dynamics. These limitations are important because they mark the boundary between reference guided appearance replacement and genuine motion aware object synthesis.

The ablation results indicate that no single component explains the behavior. Continuous boundary potentials improve shape support, temporal transport reduces framewise boundary drift, uncertainty weighted latent correction stabilizes the most fragile regions, and exterior anchoring protects the background. Applying temporal correction everywhere smooths away reference detail, while preserving the exterior too aggressively prevents target shape expansion. The useful behavior comes from assigning different roles to object core, boundary, and exterior, rather than treating the edit region as a single binary mask.

The study also argues for more diagnostic evaluation of video editing. Text alignment and general temporal similarity can miss failures that users notice immediately, especially when the target object has a different silhouette. Reference alignment, warped perceptual measures, boundary flicker, and human preference each reveal a different part of the tradeoff. Future benchmarks should include explicit shape mismatch categories and should avoid collapsing reference conditioned editing into ordinary text based semantic editing.

Future work should connect boundary transport with stronger structural priors. Depth aware occlusion reasoning, local reference patch attention, high resolution contour refinement, and multi view target conditioning could address several observed failures. Another direction is to distill the two branch denoising

and correspondence computation into a faster editor suitable for interactive use. The broader lesson is that reference conditioned video editing is not only an appearance transfer problem. It is also a problem of deciding where a new object is allowed to exist over time, and that decision should remain flexible while the denoising trajectory constructs the edited video.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[2] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, pp. 8162–8171, 2021.

[3] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in International Conference on Learning Representations, 2021.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in Advances in Neural Information Processing Systems, vol. 35, 2022.

[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, 2022.

[7] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-a-video: Text-to-video generation without text-video data," arXiv preprint arXiv:2209.14792, 2022.

[8] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen video: High definition video generation with diffusion models," arXiv preprint arXiv:2210.02303, 2022.

[9] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "Videofusion: Decomposed diffusion models for high-quality video generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[10] J. Z. Wu, Y. Ge, X. Wang, W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[11] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, "Video-p2p: Video editing with cross-attention control," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[12] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, "Fatezero: Fusing attentions for zero-shot text-based video editing," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[13] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, "Pix2video: Video editing using image diffusion," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23206–23217, 2023.

[14] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "Tokenflow: Consistent diffusion features for consistent video editing," arXiv preprint arXiv:2307.10373, 2023.

[15] W. Chai, X. Guo, G. Wang, and Y. Lu, "Stablevideo: Text-driven consistency-aware diffusion video editing," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[16] S. S. Harsha, A. Revanur, D. Agarwal, and S. Agrawal, "Genvideo: One-shot target-image and shape aware video editing using t2i diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7559–7568, 2024.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, pp. 8748–8763, 2021.

[18] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660, 2021.

[19] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," in International Conference on Learning Representations, 2022.

[20] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," in International Conference on Learning Representations, 2023.

[21] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18208–18218, 2022.

[22] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," in International Conference on Learning Representations, 2023.

[23] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1921–1930, 2023.

[24] Y. Zhao, E. Xie, L. Hong, Z. Li, and G. H. Lee, "Make-a-protagonist: Generic video editing with an ensemble of experts," arXiv preprint arXiv:2305.08850, 2023.

[25] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," arXiv preprint arXiv:2302.01329, 2023.

[26] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[28] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," arXiv preprint arXiv:2303.05499, 2023.

[29] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proceedings

of the IEEE/CVF International Conference on Computer Vision, pp. 3813–3824, 2023.

[30] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in European Conference on Computer Vision, pp. 402–419, Springer, 2020.

[31] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[32] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2085–2094, 2021.

[33] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.