LOCIART: A UNIFIED PARADIGM FOR FREE-STYLE LOCAL VIDEO STYLIZATION

Anonymous authors
Paper under double-blind review

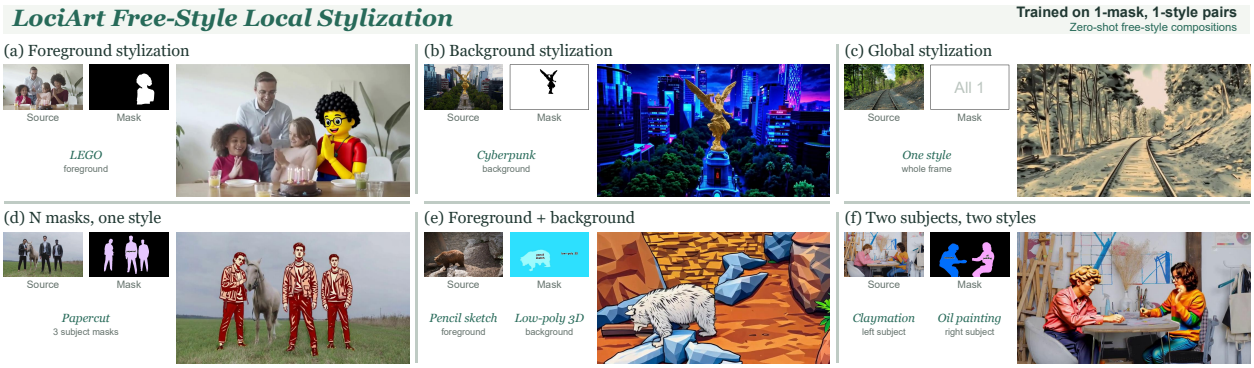


Figure 1: LociArt free-style local video stylization samples. Training uses one paired dataset with a single edited region and target style per sample. At inference, spatial control supports local or global editing, one style across multiple regions, and compositions of learned styles without compositional training pairs. Small images show sources and masks. Binary masks use white (1) for editing and black (0) for preservation. Colors identify styles.

ABSTRACT

Video stylization typically transforms entire scenes, while local artistic editing requires precise region–style association, spatial control, and temporal consistency. We introduce **LociArt**, a unified paradigm for free-style local video stylization that explicitly specifies *where to stylize and what style to apply* while preserving non-target content. At its core, **GRNStyle** builds on Generative Refinement Networks (GRN) and introduces a polar style-aware mask that embeds style identities as equal-norm directions within selected regions. Complementary representation–geometry conditioning combines encoded mask features with raw-mask spatial support, integrating spatiotemporal context with geometric control. To provide paired supervision, we develop an automated pipeline combining global stylization, video object removal, matting, and compositing, and construct the **LociArt-100K Dataset** with 100K local stylization pairs. We further establish the **LociArt Benchmark**, comprising 80 source videos and 84 target masks covering 168 foreground/background stylization cases, with a **Style Prototype Similarity (SPS)** evaluation scheme for regional style quality and complementary measures of non-target preservation and temporal consistency. Experiments show that GRNStyle outperforms existing methods in regional style quality and spatial control while maintaining strong temporal consistency. Qualitative examples further illustrate joint stylization of multiple regions with the same or different learned styles, despite training only on single-region, single-style pairs. Together, these components provide a unified foundation for learning and evaluating controllable local video stylization. Anonymous demo page: `https://lociart-page.pages.dev/`.

1 INTRODUCTION

Video stylization transforms recorded footage into artistic renditions for film, advertising, and digital art. Building on image style transfer (Gatys et al., 2016; Huang & Belongie, 2017) and diffusion-based image generation and editing (Rombach et al., 2022; Brooks et al., 2023), video methods use text or reference images to guide appearance while preserving motion (Geyer et al., 2024; Ye et al., 2025; Song et al., 2026). Yet artistic control often requires more than a global transformation: an artist may turn a moving subject into a sketch within an otherwise unchanged scene,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reimagine the surroundings as a watercolor while keeping the subject intact, or bring distinct artistic styles to different objects in the same shot, as shown in Fig. 1. We call this task *free-style local video stylization*: controlling both *where to stylize* and *what style to apply* across space and time, while preserving non-target content.

Video segmentation provides a practical starting point. A few user interactions can define masks that propagate across frames, with streaming inference supporting real-time processing (Ravi et al., 2025; Carion et al., 2026). Such masks offer an accessible spatial prior for local stylization (Song et al., 2025; He et al., 2026), extending control from foreground subjects to the less-explored task of stylizing their surroundings. Yet selecting a region does not specify its artistic transformation: a binary mask indicates editing extent but not style identity. With style supplied separately through text, the model must learn to associate the intended appearance with the selected region.

Reliable local control also requires preserving region boundaries throughout motion. Despite advances in attention supervision (Chen et al., 2026; Long et al., 2026) and mask-conditioned editing (Jiang et al., 2025; Bian et al., 2025; Li & Pan, 2026; Kang et al., 2026), mask representations face a trade-off between spatiotemporal context and spatial fidelity. VAE-based processing aligns masks with learned video features (Yatim et al., 2025), but compression can blur their geometry; direct resampling retains spatial structure without the same learned temporal context (Hu et al., 2026). Thus, a useful conditioning mechanism must both bind styles explicitly to regions and retain the spatial detail needed to localize their effects over time.

These modeling challenges are compounded by gaps in supervision and evaluation. Training pairs must demonstrate local artistic changes, faithful preservation elsewhere, and natural transitions along moving boundaries. Global stylization pairs (Song et al., 2026; Li et al., 2026) do not directly supply this supervision, while instruction-based video-editing datasets (Bai et al., 2026; Long et al., 2026) cover broader editing operations. Likewise, global similarity and video-averaged scores can conceal weak regional style, changes outside the target, and transient failures. Free-style local video stylization therefore calls for coordinated advances in paired data, region–style conditioning, and region-aware evaluation.

We introduce *LociArt*, a unified paradigm addressing these needs through the **LociArt-100K Dataset**, the **GRNStyle** model, and the **LociArt Benchmark**. Our automated data pipeline combines global stylization, video object removal, matting, and compositing to construct 100K curated pairs. Each example contains a source video, an editing mask, a style instruction, and a locally stylized target, supervising transformation and preservation together. Constructing these targets requires handling contour changes introduced by stylization. Object removal supplies clean background content for newly exposed areas, while alpha mattes estimated separately for original and stylized objects guide compositing along their respective contours, avoiding object remnants and enabling natural boundary transitions.

Trained on these pairs, **GRNStyle** builds on Generative Refinement Networks (GRN) (Han et al., 2026) and the source-evidence editing formulation of GRNEdit (Xie et al., 2026). Its core design is a *polar style-aware mask*, which embeds style identities as equal-norm directions within selected regions, unifying *where to stylize* and *what style to apply* without imposing a scalar ordering on styles. Complementary *representation–geometry conditioning* combines VAE-encoded mask features with raw-mask spatial support to integrate spatiotemporal context and geometric control, while source evidence guides content preservation throughout refinement. The same interface accommodates distinct style codes across regions: qualitative examples illustrate joint stylization of multiple non-overlapping regions with the same or different learned styles, using only single-region, single-style supervision and no additional training.

To evaluate local stylization, the **LociArt Benchmark** contains 80 source videos and 84 object–mask units, each evaluated under both foreground and background stylization protocols, yielding 168 stylization cases. It uses *Style Prototype Similarity (SPS)* to measure regional style fidelity, together with complementary measures of non-target preservation and temporal consistency.

Our contributions are threefold:

- **A Unified Paradigm for Local Video Stylization.** We introduce LociArt, a unified computational paradigm that formulates local video stylization as region-conditioned artistic transformation, explicitly specifying where to stylize and what style to apply while preserving non-target content and temporal consistency.
- **GRNStyle with Style-Aware Region Control.** We propose GRNStyle, a GRN-based model that unifies region and style through polar style-aware masks and combines spatiotemporal context with geometric control. Qualitative examples further illustrate multi-region, multi-style compositions learned from single-region, single-style supervision.
- **Scalable Data Construction and a Region-Aware Evaluation.** We develop an automated pipeline to construct the LociArt-100K Dataset and establish the LociArt Benchmark with 80 source videos and 84 object–mask units, integrating SPS with preservation and temporal consistency measures for region-aware evaluation.

2 Related Work

Image style transfer. Neural style transfer evolved from per-image optimization (Gatys et al., 2016) to feed-forward arbitrary transfer with AdaIN (Huang & Belongie, 2017). Building on latent diffusion (Rombach et al., 2022), StyleShot (Gao et al., 2025) separates style encoding from content conditioning, while TeleStyle (Zhang et al., 2026) learns content-preserving transfer from curated triplets and extends it to video. These approaches advance reference-based stylization; localized editing additionally requires control over where the transformation applies.

Localized stylization. At the image level, DiffStyler (Li & Pan, 2026) combines style LoRA with masked feature and attention control. RegionRoute (Chen et al., 2026) learns regional attention from masks and operates without masks at inference, while StyleGallery (He et al., 2026) matches semantic regions for training-free image stylization. For video, Xia et al. (2021) learn localized photorealistic transfer with spatiotemporal feature alignment. UniVST (Song et al., 2025) instead combines mask propagation, AdaIN-guided latent and attention manipulation, and optical-flow smoothing without training. We study local artistic video stylization through paired supervision, using fixed, equal-norm polar codes to bind style identity to editable regions.

Video stylization and generative editing. PickStyle (Mehraban et al., 2025) trains lightweight style adapters using synthetic clips derived from paired images. VISTA (Song et al., 2026) and MoStyle (Wei et al., 2026) uses triplets of style references, source videos, and stylized videos, while DreamStyle (Li et al., 2026) learns text-, image-, and first-frame-guided stylization from paired videos. For broader editing, VACE (Jiang et al., 2025) unifies video conditions, including masks, while VideoPainter (Bian et al., 2025) provides context-conditioned video inpainting. Beyond diffusion, GRN (Han et al., 2026) synthesizes videos through binary-state refinement, and GRNEdit (Xie et al., 2026) treats source content as evidence for editing those states. GRNStyle extends this evidence-based formulation by coupling VAE-encoded mask features with raw-mask geometry, combining temporal semantics with spatial fidelity.

Data and evaluation. Ditto-1M (Bai et al., 2026) scales instruction-based video editing, while CoinVE-200K and CoinVE-Bench (Long et al., 2026) target compositional editing, including regional stylization. StyleShot (Gao et al., 2025) and MegaStyle (Gao et al., 2026) provide image datasets organized around stylistic consistency and diversity; StyleShot also introduces StyleBench. Evaluation ranges from image stylization quality with ArtFID (Wright & Ommer, 2022) and style similarity with CSD (Somepalli et al., 2024) to video quality and temporal consistency with VBench (Huang et al., 2024). RegionRoute (Chen et al., 2026) further evaluates regional style and background preservation. The LociArt-100K Dataset complements these resources with video pairs aligning source content, editing regions, style labels, and local targets. The accompanying LociArt Benchmark evaluates regional style, non-target preservation, temporal consistency, and boundary behavior.

3 LociArt-100K Dataset Construction

To supervise local stylization and non-target preservation, we construct the **LociArt-100K Dataset** of 100K pairs by separating global style generation from spatial composition. Specifically, we globally stylize source and object-removed videos, then recombine their components into complementary foreground and background targets. Each example contains a source video $\mathbf{V}_{\mathrm{src}}$, binary editing mask $\mathbf{M}$, style instruction p, and locally stylized target $\mathbf{V}^{\mathrm{sty}}$. All videos are provided at 720p resolution. As illustrated in Fig. 2, the pipeline comprises the following four stages.

Object identification and removal. We first identify the objects to separate from their surroundings. Source videos are collected from SA-V (Ravi et al., 2025) and Ditto-1M (Bai et al., 2026), Qwen3-VL-32B (Bai et al., 2025) identifies salient objects whose category prompts guide SAM3 (Carion et al., 2026) to produce temporally tracked masks $\mathbf{M}_{\mathrm{obj}}$. After vision-language, temporal-coverage, and area checks, each accepted mask guides SVOR (Hu et al., 2026) to remove the selected object, yielding $\mathbf{V}_{\mathrm{rem}}$. This reconstructed background fills areas exposed by stylization-induced contour changes, avoiding original-object remnants during compositing. Appendix Section A.1 details the mask selection and filtering procedures.

Global stylization and filtering. With source and reconstructed background videos available, we run both VACE (Jiang et al., 2025) and Ditto (Bai et al., 2026) on each under a stylization instruction p. For either global stylizer G, we get two stylized outputs:

$$\mathbf{V}^{\mathrm{sty}}_{\mathrm{src}} = G(\mathbf{V}_{\mathrm{src}}, p), \qquad \mathbf{V}^{\mathrm{sty}}_{\mathrm{rem}} = G(\mathbf{V}_{\mathrm{rem}}, p). \tag{1}$$

Before using these outputs for local composition, GPT-5.5 is used to assess style quality, content preservation, and temporal quality through two cascaded gates: the source-video pair is evaluated first, followed by the object-removed pair. Only candidates passing both proceed to matting and compositing. Sampling, scoring, and rejection criteria appear in Appendix Section A.2.

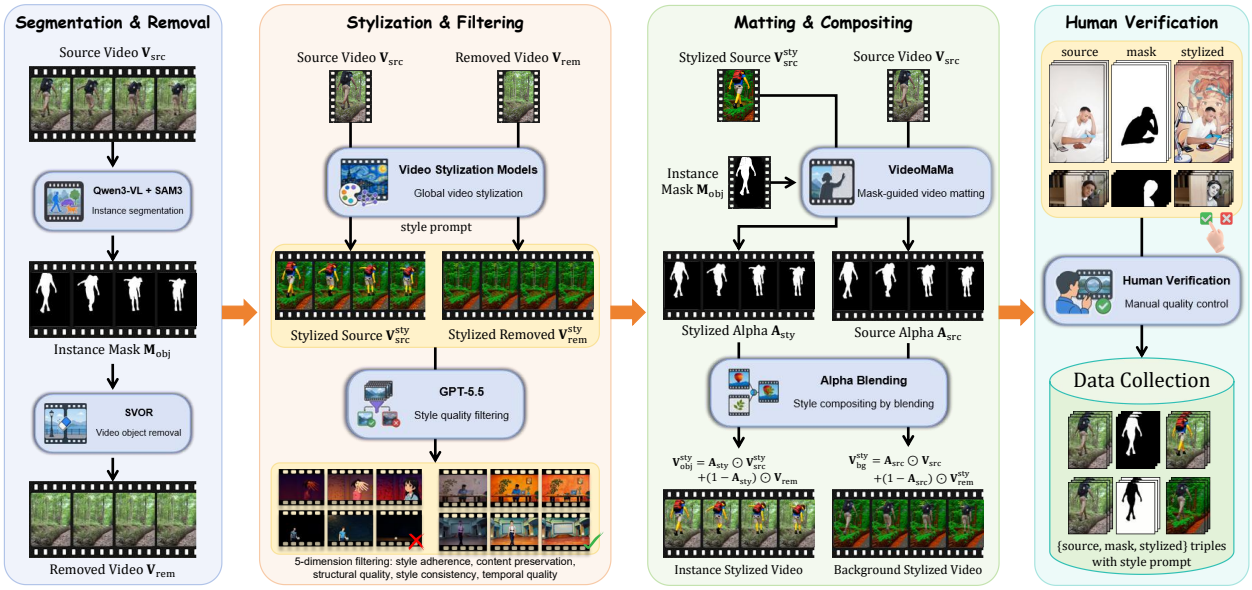


Figure 2: **Overview of the data construction pipeline.** Source and object-removed videos are globally stylized and filtered, then combined through mask-guided matting to produce foreground- and background-stylized pairs for human verification.

Matting and localized compositing. We next localize the accepted stylizations. Since stylization can change object contours, we use VideoMaMa (Lim et al., 2026) to estimate separate alpha mattes $\mathbf{A}_{\text{src}}$ and $\mathbf{A}_{\text{sty}}$ from $\mathbf{V}_{\text{src}}$ and $\mathbf{V}^{\text{sty}}_{\text{src}}$ with the same tracked mask $\mathbf{M}_{\text{obj}}$, respectively. These mattes follow the respective object boundaries through time and are used for compositing instead of binary masks. The resulting composites stylize the foreground while retaining the background, or vice versa:

$$\begin{aligned} \mathbf{V}^{\text{sty}}_{\text{fg}} &= \mathbf{A}_{\text{sty}} \odot \mathbf{V}^{\text{sty}}_{\text{src}} + (1 - \mathbf{A}_{\text{sty}}) \odot \mathbf{V}_{\text{rem}}, \\ \mathbf{V}^{\text{sty}}_{\text{bg}} &= \mathbf{A}_{\text{src}} \odot \mathbf{V}_{\text{src}} + (1 - \mathbf{A}_{\text{src}}) \odot \mathbf{V}^{\text{sty}}_{\text{rem}}. \end{aligned} \tag{2}$$

Here $\odot$ denotes element-wise multiplication, with alpha values broadcast across color channels. Both targets are paired with $\mathbf{V}_{\text{src}}$ and p, using editing masks $\mathbf{M}_{\text{obj}}$ and $1 - \mathbf{M}_{\text{obj}}$, respectively.

Human verification. Finally, because compositing may introduce artifacts not captured by global filtering, annotators inspect each completed example for content preservation, boundary quality, and temporal consistency. Only qualified samples are retained, yielding a final dataset of 100K video pairs covering 18 style categories. Appendix Section A.3 provides more details on the verification criteria and dataset statistics.

4 GRNStyle: Style-Aware Local Video Stylization

As the model within LociArt, **GRNStyle** binds style identities to target regions and combines encoded mask features with explicit geometric support (Fig. 3). Source evidence guides content preservation throughout generation, while the same spatial conditioning interface accommodates multiple regions at inference.

4.1 Problem Formulation and Binary Evidence

Given a source video $\mathbf{X} \in [0,1]^{T\times H\times W\times 3}$, an editing mask $\mathbf{M} \in \{0,1\}^{T\times H\times W\times 1}$, and a text instruction p, we derive the target style $s \in \{1,\dots,K\}$ from p, where K is the number of style categories. We embed s into $\mathbf{M}$ to form a style-aware mask $\mathbf{M}_s$ and learn $\widehat{\mathbf{Y}} = F(\mathbf{X}, \mathbf{M}_s, p; \boldsymbol{\theta})$ from locally stylized targets $\mathbf{Y}$, aiming to transform the selected region while preserving non-target content.

To preserve source content during generation, we build on GRN (Han et al., 2026) and GRNEdit's source-evidence formulation (Xie et al., 2026). An encoder $\mathcal{E}$ and hierarchical binary quantizer $\mathcal{Q}$ map the source to $\mathbf{B}^{\mathbf{X}} = \mathcal{Q}(\mathcal{E}(\mathbf{X})) \in \{0,1\}^{L\times D}$, with L latent positions and D bits per position. GRN uses one-hot bit codes $\phi(b) = (1-b, b)$ and parallel

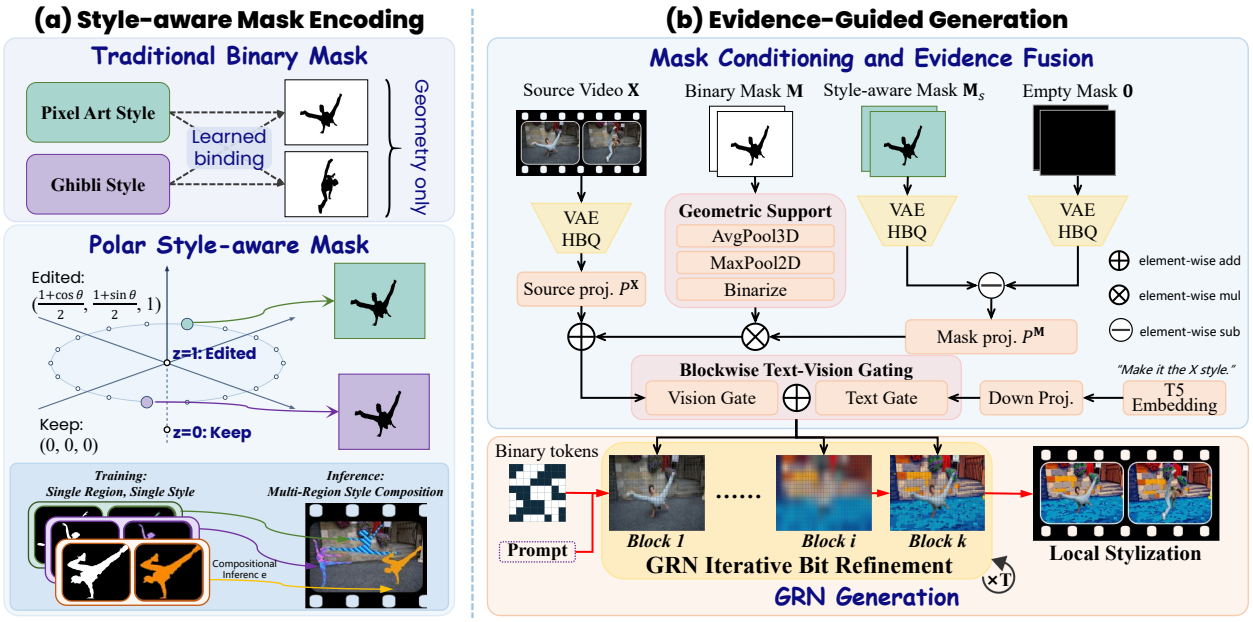


Figure 3: **Overview of GRNStyle.** (a) Polar style-aware masks bind styles to regions and support multi-region composition. (b) Anchored mask evidence and geometric support are combined with source features through block-wise gating to guide local video stylization.

iterative refinement, while GRNEdit supplies source evidence for retaining or changing target bits. GRNStyle adds explicit style–region binding and geometric control to localize the transformation.

4.2 Equal-Norm Style–Space Conditioning

A binary mask specifies where to edit but leaves the associated style to a separate text condition. To bind style directly to space, we assign each style a fixed polar angle θ_s at unit radius and embed its coordinates within the selected region (Fig. 3(a)). The first two channels of the three-channel VAE input store the rescaled Cartesian coordinates, while the third retains binary occupancy:

$$\boldsymbol{a}_s = (\cos\theta_s, \sin\theta_s), \qquad \mathbf{M}_s(q) = \mathbf{M}(q)\left(\frac{1+a_{s,1}}{2}, \frac{1+a_{s,2}}{2}, 1\right). \tag{3}$$

Here q indexes spatiotemporal pixel positions. Since $\|\boldsymbol{a}_s\|_2 = 1$, styles differ by direction rather than magnitude, avoiding an arbitrary scalar ordering. Unedited positions remain zero, and occupancy is directly recoverable from the third channel. This representation adds neither channels nor conditioning tokens relative to a repeated binary mask. Appendix Section B.1 provides encoding details.

Crucially, the resulting field contains the region–style interaction before video encoding, as a linear readout recovers

$$\mathcal{R}(\mathbf{M}_s(q)) = \big(2\mathbf{M}_{s,1}(q) - \mathbf{M}_{s,3}(q),\ 2\mathbf{M}_{s,2}(q) - \mathbf{M}_{s,3}(q)\big) = \mathbf{M}(q)\boldsymbol{a}_s. \tag{4}$$

Thus, the input explicitly associates each edited position with its style code. Appendix Section B.3 derives this readout and compares it with additive readouts of separate region and style conditions.

4.3 Representation–Geometry Mask Conditioning

The mask must be encoded on the backbone's latent grid in video generation. However, Hu et al. (2026) reported that common mask compression strategies can compromise spatial localization. We therefore combine encoded mask features with geometric support from binary occupancy, retaining spatiotemporal context while localizing mask evidence.

Representation and anchoring. To align mask conditioning with source representations, we encode both through the frozen map $\Phi = \phi \circ \mathcal{Q} \circ \mathcal{E}$, including input normalization and spatiotemporal flattening. Since an all-zero mask does not encode to zero, we anchor mask evidence to the encoding of an empty mask of the same shape:

$$\mathbf{Z}^{\mathbf{X}} = \Phi(\mathbf{X}), \qquad \Delta\mathbf{Z}^{\mathbf{M}} = \Phi(\mathbf{M}_s) - \Phi(\mathbf{0}). \tag{5}$$

This subtraction ensures zero mask evidence for an absent edit and represents nonempty masks through changes relative to the shared empty-condition baseline.

Geometric support. Following the VAE's temporal grouping and spatial strides, we use average-pool $\mathcal{P}$ to binary occupancy onto the latent grid, then apply spatial max pooling $\mathcal{D}$ and threshold:

$$\Omega(\mathbf{M}) = \mathbf{1}[\mathcal{D}(\mathcal{P}(\mathbf{M})) \geq 0.5]\,. \tag{6}$$

Max pooling uses unit stride and size-preserving padding, providing a kernel-controlled boundary margin without further downsampling. Complementing empty-mask anchoring, the flattened support $\Omega(\mathbf{M}) \in \{0,1\}^{L\times 1}$ localizes projected mask injection without hard-masking source features or output pixels.

4.4 Evidence-Guided Generation and Compositional Inference

Evidence fusion. To condition the backbone on source content and local style, we project their evidence into a shared d-dimensional space using $\boldsymbol{P}^{\mathbf{X}}$ and $\boldsymbol{P}^{\mathbf{M}}$. The mask projection is bias-free to preserve the zero anchored response. These projections are shared across injection blocks and form $\mathbf{V} = \boldsymbol{P}^{\mathbf{X}}(\mathbf{Z}^{\mathbf{X}}) + \Omega(\mathbf{M}) \odot \boldsymbol{P}^{\mathbf{M}}(\Delta\mathbf{Z}^{\mathbf{M}}) \in \mathbb{R}^{L\times d}$, combining source context with geometrically localized mask evidence. To adapt this shared signal to each block, we compute the channel gate $\boldsymbol{g}_k = g_{\mathrm{vis},k}(\mathbf{V}) + g_{\mathrm{text},k}(p) \in \mathbb{R}^d$ from visual evidence and text, broadcast it across the L spatiotemporal positions, and inject the gated residual:

$$\mathbf{R}_k = \mathbf{V} \odot \boldsymbol{g}_k, \qquad \mathbf{H}_k = \mathcal{G}_k(\mathbf{H}_{k-1} + \Pi_{\mathrm{vis}}(\mathbf{R}_k))\,. \tag{7}$$

Here $\mathcal{G}_k$ is the k-th DiT block with output $\mathbf{H}_k$, and Π_{vis} inserts residuals at existing visual positions without adding tokens. Blocks without evidence injection use $\mathbf{R}_k = 0$.

Learning and generation. With clean source and mask evidence as conditions, GRNStyle predicts target bits from a partially randomized state. We retain GRN's full-video bitwise cross-entropy to jointly supervise stylization and preservation. Training mixes style-specific text with generic mask-following instructions, encouraging the model to read spatially encoded styles while retaining text guidance. At inference, random target bits are iteratively refined under the same conditioning and decoded into video.

Compositional inference. This mask-following interface also supports multiple regions through the same codebook. For disjoint regions $\{\mathbf{M}^j\}_{j=1}^N$ and learned styles $\{s_j\}_{j=1}^N$, we compose the encodings in Equation 3:

$$\mathbf{M}_s^{\mathrm{multi}}(q) = \sum_{j=1}^{N} \mathbf{M}^j(q) \left(\frac{1+a_{s_j,1}}{2}, \frac{1+a_{s_j,2}}{2}, 1 \right), \qquad \mathbf{M}^{\mathrm{union}} = \sum_{j=1}^{N} \mathbf{M}^j. \tag{8}$$

Disjointness ensures a valid style code or empty condition at each position, with $\mathcal{R}(\mathbf{M}_s^{\mathrm{multi}}(q)) = \sum_j \mathbf{M}^j(q)\boldsymbol{a}_{s_j}$ by Equation 4. We jointly encode the composite field and localize its mask evidence using $\Omega(\mathbf{M}^{\mathrm{union}})$. All regions are generated in a single refinement trajectory under a generic mask-following instruction, without adding channels or conditioning tokens. Qualitative examples illustrate these compositions after single-region, single-style training, without further training on region–style combinations.

5 Experiments

5.1 Experimental Setup

Training. We train the 2B GRNStyle model on LociArt-100K with 8 $\times$ NVIDIA H200 GPUs, using all 18 styles unless otherwise specified. We fully fine-tune the backbone and conditioning branches while keeping the video tokenizer and T5 encoder frozen. More training details are provided in Appendix Section E.

Benchmark. We introduce the LociArt Benchmark for local video stylization task, which contains 80 source videos and 84 corresponding object masks. Each video–mask pair is evaluated under two complementary stylization protocols: foreground stylization edits the target mask M, while background stylization edits its complement $1 - M$. The benchmark therefore contains 168 cases in total. More details are provided in Appendix Section C.1.

Metrics. We use *SPS* (Style Prototype Similarity to measure whether the edited region matches the requested style using a prototype bank in the CSD (Somepalli et al., 2024) feature space. Regional *CLIPT* (Radford et al., 2021) measures alignment between the edited region and the textual style instruction, while *PSNR* (Hore & Ziou, 2010) measures fidelity to the source in the non-target region. *TC* (Temporal Consistency) (Zi et al., 2025) evaluates temporal consistency by measuring feature similarity between adjacent frames. We additionally use Gemini 3.6 Flash to

Table 1: **Comparison on the LociArt Benchmark.** The fg/bg suffixes denote stylization of the target mask M and its non-target complement $1 - M$, respectively. FT denotes local-stylization fine-tuning; $*$ denotes training-free methods. † denotes direct mask-compositing methods; their PSNR is omitted because the non-edited regions are copied from the source. ‡ denotes image-based methods. Best and second-best results are shown in bold and underlined, respectively.

Method	Venue	SPS-fg↑	SPS-bg↑	CLIPT-fg↑	CLIPT-bg↑	PSNR-fg↑	PSNR-bg↑	TC↑	Style↑	Preserve↑
LEAST*†‡	arXiv '24	0.283	0.302	0.158	0.173	-	-	0.953	2.20	4.50
VACE-1.3B	ICCV '25	0.156	0.191	0.114	0.132	**32.35**	**29.17**	0.981	0.05	**5.00**
UniVST*‡	TPAMI '25	0.191	0.286	0.127	0.165	27.69	23.92	0.969	1.83	4.40
RegionRoute‡	CVPR '26	0.317	0.491	0.154	0.183	<u>29.38</u>	13.17	0.955	2.41	2.75
StyleGallery*†‡	CVPR '26	0.230	0.325	0.146	0.183	-	-	0.957	3.39	<u>4.77</u>
CoinVE-Edit	arXiv '26	0.248	0.251	0.153	0.159	20.71	16.76	<u>0.983</u>	3.11	3.93
Seedance 2.5	Jul. 2026	0.347	0.358	**0.173**	<u>0.175</u>	18.50	18.76	**0.984**	3.98	2.40
VACE-1.3B (our FT)	-	<u>0.386</u>	<u>0.676</u>	0.155	<u>0.185</u>	25.43	25.49	0.980	<u>4.52</u>	4.56
GRNStyle-2B (ours)	-	**0.407**	**0.705**	<u>0.160</u>	**0.199**	25.92	<u>26.15</u>	<u>0.983</u>	**4.62**	4.64

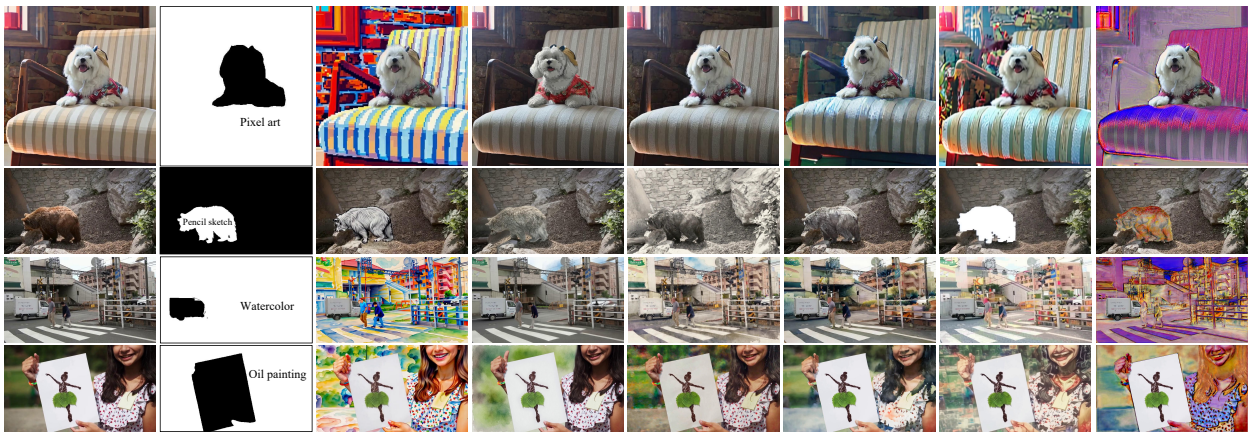


Figure 4: **Qualitative comparison of local video stylization.** Matched examples compare artistic expression inside the editing region, preservation outside it, and consistency across time. Foreground and background cases expose different control failures: spill into neighboring content, distortion of preserved subjects, and style changes along moving boundaries. GRNStyle combines distinct local style with coherent boundaries and stable non-target appearance.

provide separate *Style* and *Preserve* scores for regional stylization quality and non-target content preservation. Detailed definitions and implementation protocols are provided in Appendices Section C.2–Section C.5.

5.2 Comparison with Existing Methods

We compare with methods spanning regional image stylization and video editing: LEAST (Singh et al., 2024), UniVST (Song et al., 2025), RegionRoute (Chen et al., 2026), and StyleGallery (He et al., 2026) are image-based methods, whereas VACE-1.3B (Jiang et al., 2025), CoinVE-Edit (Long et al., 2026), and the commercial system Seedance 2.5 (ByteDance Seed Team, 2026) target video generation or editing. LEAST and StyleGallery directly composite the non-edited region from the source using the mask, while the other methods generate or reconstruct the complete output; accordingly, PSNR is not reported for the former two because source-pixel copying makes it non-comparable. To isolate the contribution of task-specific supervision, we additionally fine-tune VACE-1.3B on LociArt-100K using the same architecture-independent hyperparameters as GRNStyle.

Quantitative and perceptual comparison. As shown in Table 1, GRNStyle achieves the best SPS for both foreground (0.407) and background (0.705) stylization, together with the best background CLIPT (0.199). It also maintains strong preservation, with the second-best background PSNR (26.15 dB) and near-best temporal consistency (TC: 0.983). The VLM scores further indicate a favorable balance between regional stylization and non-target preservation, achieving the highest Style score (4.62) while retaining a high Preserve score (4.64).

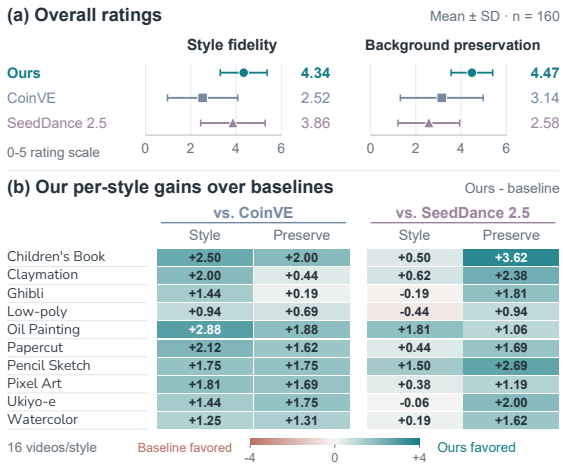


Figure 5: Human evaluation: mean ratings ± SD and per-style gains.

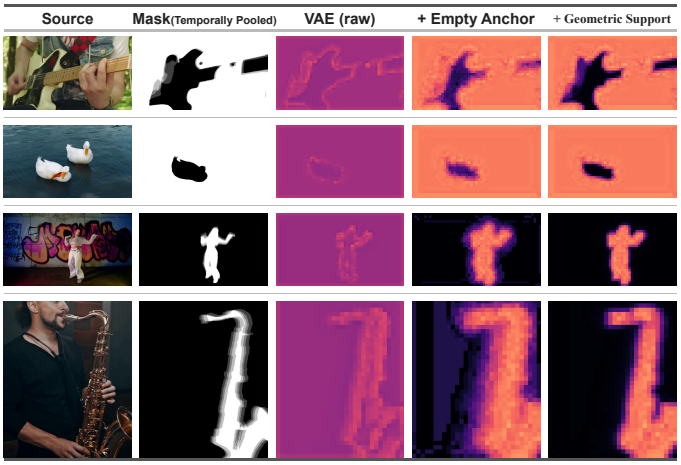


Figure 6: Mask-conditioning ablation: representation–geometry encoding improves spatial localization.

Table 2: **Component ablation.** RG denotes representation–geometry conditioning. Metrics are the unweighted mean over foreground- and background-stylization cases.

Configuration	SPS↑	CLIPT↑	PSNR↑
Base	0.541	0.180	20.09
Base + RG	0.544	0.179	25.07
Base + Polar	0.553	0.181	21.34
GRNStyle (Base + RG + Polar)	0.554	0.181	25.78

Qualitative comparison. Fig. 4 provides representative regional edits. GRNStyle renders the armchair and room in pixel art while preserving the dog, and adds pencil strokes to the bear without desaturating the surrounding rocks. CoinVE-Edit produces weaker transformations; Seedance 2.5 alters content beyond the bear mask, while UniVST reduces the bear to a white silhouette. GRNStyle introduces painterly texture while preserving the held illustration. These cases highlight the distinction between localized stylization and weak editing, global recoloring, or loss of subject detail. Additional examples are provided in Appendix Section D.

5.3 User Study

We conducted a user study with 20 participants on 168 cases. Each participant was assigned one result per case. Since the study evaluates generated videos, we compare GRNStyle only with the two video-based methods, CoinVE-Edit and Seedance 2.5. As shown in Fig. 5(a), GRNStyle achieves the highest mean ratings for style fidelity (4.34) and background preservation (4.47), indicating stronger artistic expression together with better scene preservation. Fig. 5(b) further shows consistent benefits across styles: GRNStyle outperforms CoinVE-Edit on both criteria for all ten styles, improves preservation over Seedance 2.5 for every style, and improves style fidelity for seven styles.

5.4 Ablation of Mask Conditioning

We evaluate two complementary mask-conditioning strategies: RG (representation–geometry conditioning), which combines empty-mask anchoring with raw-mask geometric support, and Polar conditioning, which encodes an explicit regional style identity through a polar representation.

Table 2 shows that the baseline performs worst overall. RG substantially improves PSNR while leaving SPS and CLIP-T nearly unchanged, indicating improved preservation without sacrificing stylization. As illustrated in Fig. 6, empty-mask anchoring removes the reference response to an empty input, while geometric support suppresses spatial spillover. Polar encoding supplies an explicit regional style identity, yielding a clear SPS improvement and a modest PSNR gain. Combining the two components achieves the strongest overall trade-off between style conditioning and spatial control.

Table 3: **Replay-free style expansion from 10 to 15 classes.** Expanded models train only on the five new classes. Retention and acquisition are evaluated on 1,680 original-style and 840 new-style cases, respectively.

	10 classes (old)			5 classes (new)		
Configuration	SPS↑	CLIPT↑	PSNR↑	SPS↑	CLIPT↑	PSNR↑
Binary, 10 classes	0.547	0.177	25.42	—	—	—
Binary, 15 classes	0.445 $-\Delta 0.102$	0.158 $-\Delta 0.019$	25.43 $+\Delta 0.01$	0.588	0.197	25.39
Polar, 10 classes	0.553	0.180	25.75	—	—	—
Polar, 15 classes	0.531 $-\Delta 0.022$	0.172 $-\Delta 0.008$	25.77 $+\Delta 0.02$	0.590	0.199	25.74

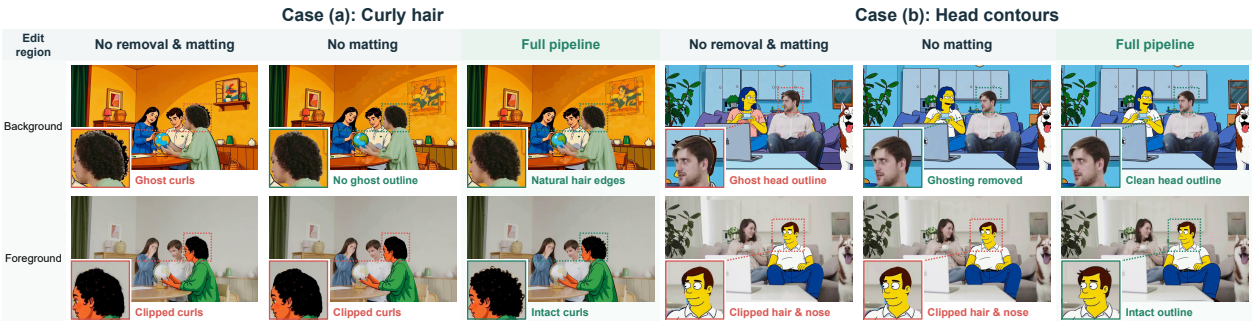


Figure 7: **Data-construction ablation.** For curly hair (a) and head contours (b), columns compare direct compositing, removal without matting, and the full pipeline. Rows show constructed background (top) and foreground (bottom) stylization targets. Insets highlight residual silhouettes, clipped contours, and the full pipeline's cleaner boundaries.

5.5 Learning New Styles without Replay

To examine whether a predefined polar codebook limits expansion to new styles, we extend a separately trained 10-style model to 15 styles using only the five new classes and no replay of the original data. The polar and binary variants follow the same protocol, and retention is measured against their respective 10-class checkpoints.

Table 3 shows that polar conditioning supports this 50% vocabulary expansion with substantially less forgetting: old-class SPS drops by 0.022, compared with 0.102 for binary masks, while the corresponding CLIPT drops are 0.008 and 0.019; PSNR remains stable in both cases. The new-class scores are also slightly higher across all metrics, indicating that improved retention does not hinder adaptation. These results support the extensibility of explicit style identities: new styles can be added while better preserving existing capabilities, without replaying the original data.

5.6 Effect of the Data Construction Pipeline

To assess the contribution of each data-construction step, we compare direct mask compositing, using object removal but no matting, and the complete pipeline for both foreground and background stylization. As shown in Fig. 7, background editing with direct compositing can leave ghost curls or head outlines when stylized subjects extend beyond their source silhouettes; object removal eliminates these artifacts. For foreground editing, source-aligned masks may clip stylized curls, hair tips, and noses, while removal alone cannot resolve the mismatch. Matting follows the actual foreground contour, preserving fine boundaries and providing cleaner supervision for style-induced shape changes.

6 Conclusion

We introduced LociArt, a unified framework comprising automated data construction, the LociArt-100K dataset, GRNStyle, and a region-aware benchmark. Built on GRN, GRNStyle combines representation–geometry conditioning with polar style codes to balance spatial control and temporal context while preserving non-target content, outperforming the evaluated baselines. Category-aware polar style-aware masks further enable zero-shot multi-region compositions from single-region, single-style supervision, reducing the need for exhaustive paired data. Future work will study compositional generalization across training diversity, style vocabulary, and region complexity; further limitations are discussed in Appendix F.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing, to draft portions of the paper, and to generate synthetic datasets used in our data-construction pipeline. All AI-assisted text was reviewed and revised by the authors, and all synthetic data were inspected and validated before use. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide anonymized inference code with the current submission. The appendix documents the training configuration and implementation details, while the anonymous project page presents representative examples from the LociArt-100K dataset. We plan to publicly release the trained model weights and the complete LociArt-100K dataset in the future, subject to the applicable licensing and data-distribution requirements.

REFERENCES

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37971–37981, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

ByteDance Seed Team. One-take creation, flexible referencing: Introducing seedance 2.5, July 2026. URL `https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5`. Product announcement, July 31, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Bowen Chen, Jake Zuena, Alan C Bovik, and Divya Kothandaraman. Regionroute: Regional style transfer with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35736–35746, 2026.

Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Junyao Gao, Sibo Liu, Jiaxing Li, Yanan Sun, Yuanpeng Tu, Fei Shen, Weidong Zhang, Cairong Zhao, and Jun Zhang. Megastyle: Constructing diverse and scalable style dataset via consistent text-to-image style mapping. *arXiv preprint arXiv:2604.08364*, 2026.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Michal Geyer, Omer Bar Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, volume 2024, pp. 1608–1620, 2024.

Jian Han, Jinlai Liu, Jiahuan Wang, Bingyue Peng, and Zehuan Yuan. Generative refinement networks for visual synthesis. In *European Conference on Computer Vision*, pp. 566–585. Springer, 2026.

Boyu He, Yunfan Ye, Chang Liu, Weishang Wu, Fang Liu, and Zhiping Cai. Stylegallery: Training-free and semantic-aware personalized style transfer from arbitrary image references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29092–29102, 2026.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Jiagao Hu, Yuxuan Chen, Fuhao Li, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. From ideal to real: Stable video object removal under imperfect conditions. *arXiv preprint arXiv:2603.09283*, 2026.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1510–1519. IEEE, 2017.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Dongkyung Kang, Jaeyeon Hwang, Junseo Park, Minji Kang, Yeryeong Lee, Beomseok Ko, Hanyoung Roh, Jeongmin Shin, and Hyeryung Jang. Mast: Mask-guided attention control for training-free regional-multi style transfer. *arXiv preprint arXiv:2604.12281*, 2026.

Mengtian Li, Jinshu Chen, Songtao Zhao, Wanquan Feng, Pengqi Tu, and Qian He. Dreamstyle: A unified framework for video stylization. *arXiv preprint arXiv:2601.02785*, 2026.

Shaoxu Li and Ye Pan. An efficient and harmonized framework for balanced cross-domain feature integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6369–6377, 2026.

Sangbeom Lim, Seoung Wug Oh, Jiahui Huang, Heeji Yoon, Seungryong Kim, and Joon-Young Lee. Videomama: Mask-guided video matting via generative prior. *arXiv preprint arXiv:2601.14255*, 2026.

Fuchen Long, Cong Wang, Zitao Gao, Wenhao Zhong, Yu Cheng, Xiaolu Hou, Yan Li, Xiao Cao, Xinlong Sun, Xi Chen, et al. Coinve-200k: A large-scale high-quality dataset for compositional instruction-guided video editing. *arXiv preprint arXiv:2608.17566*, 2026.

Soroush Mehraban, Vida Adeli, Jacob Rommann, Babak Taati, and Kyryl Truskovskyi. Pickstyle: Video-to-video style transfer with context-style adapters. *arXiv preprint arXiv:2510.07546*, 2025.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Silky Singh, Surgan Jandial, Simra Shahid, and Abhinav Java. Least:" local" text-conditioned image style transfer. *arXiv preprint arXiv:2405.16330*, 2024.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

Quanjian Song, Mingbao Lin, Wengyi Zhan, Shuicheng Yan, Liujuan Cao, and Rongrong Ji. Univst: A unified framework for training-free localized video style transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yiren Song, Wangzi Yao, Haofan Wang, and Mike Zheng Shou. Vista: Triplet-supervised video style transfer with diffusion transformers. *arXiv preprint arXiv:2605.17312*, 2026.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

Xin Wei, Yijie Fang, Yanjia Li, Liangyi Wu, Qin Yang, Mingrui Zhu, Nannan Wang, and Xinbo Gao. Stylized video generation via decoupled data synthesis and gated style token injection. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2026.

Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pp. 560–576. Springer, 2022.

Xide Xia, Tianfan Xue, Wei-sheng Lai, Zheng Sun, Abby Chang, Brian Kulis, and Jiawen Chen. Real-time localized photorealistic video style transfer. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1088–1097. IEEE, 2021.

Feng Xie, Jiagao Hu, Fuhao Li, Zepeng Wang, Yuxuan Chen, Dahua Gao, Fei Wang, and Daiguo Zhou. Grnedit: Efficient general video editing from a new binary-evidence perspective in generative refinement networks. *arXiv preprint arXiv:2608.16328*, 2026.

Danah Yatim, Rafail Fridman, Omer Bar-Tal, and Tali Dekel. Dynvfx: Augmenting real videos with dynamic content. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2630–2640. IEEE, 2025.

Shiwen Zhang, Xiaoyan Yang, Bojia Zi, Haibin Huang, Chi Zhang, and Xuelong Li. Telestyle: Content-preserving style transfer in images and videos. *arXiv preprint arXiv:2601.20175*, 2026.

Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. Minimax-remover: Taming bad noise helps video object removal. In *Advances in Neural Information Processing Systems*, 2025.

APPENDIX

Appendix contents. The supplementary material is organized as follows:

- Detailed Data Pipeline (p. 13): data construction, filtering, and human verification.
- Style-Code Design and Local Readout Analysis (p. 15): style-code design and local readout analysis.
- Details of the Evaluation Protocol (p. 17): benchmark construction, SPS, regional CLIP-T, preservation, temporal consistency, and VLM evaluation.
- Additional Qualitative Results (p. 21): additional visual comparisons and examples.
- Model Reproduction Details (p. 25): training hyperparameters, reproduction details, and training dynamics.
- Limitations and Future Work (p. 26): fast-motion consistency and shadow-preservation limitations.

A DETAILED DATA PIPELINE

All videos are standardized to 81 frames at 720p resolution. Frame correspondence is maintained throughout object removal, global stylization, matting, and compositing, so source, removed, stylized, and alpha videos remain aligned. Both VACE (Jiang et al., 2025) and Ditto (Bai et al., 2026) are run on the source and object-removed videos; each stylizer produces candidate outputs for the subsequent quality checks.

A.1 OBJECT INVENTORY AND MASK FILTERING DETAILS

This stage identifies candidate foreground objects and obtains reliable masks for object removal and subsequent local composition. We first build an object inventory, track the selected instances, and then filter their masks for semantic correctness and temporal coverage.

Object inventory. Qwen3-VL-32B (Bai et al., 2025) analyzes each source video to identify prominent foreground objects, excluding broad background regions such as sky, roads, and walls. Each inventory entry contains a concise category label and an appearance- and location-based description. The category label provides the prompt for segmentation.

Instance handling. Using these category prompts, SAM3 (Carion et al., 2026) produces temporally tracked instance masks. When several objects share a category, their masks remain separate if their identities can be reliably tracked, preserving object-specific control for the later removal and compositing steps.

Mask filtering. We then use Qwen3-VL-32B (Bai et al., 2025) to reject missing, flickering, or incorrectly assigned masks. Additional checks require nonempty masks in more than 40% of frames and reject masks whose maximum per-frame area exceeds 85% of the frame. Temporal coverage counts all nonempty frames, not the length of a continuous track. These production thresholds select masks with sufficient visibility and a bounded foreground extent before object removal.

A.2 GLOBAL STYLIZATION QUALITY FILTERING DETAILS

The source and object-removed videos provide complementary content for local composition, so both stylizations must preserve their inputs while expressing the requested style. We assess them through two cascaded gates with a shared sampling and scoring protocol, retaining only candidates that satisfy the acceptance criteria below.

Cascaded evaluation. The source-video pair is evaluated first; failure at this gate stops processing for the candidate. Passing candidates proceed to evaluation of the object-removed pair, and both pairs must pass before matting and compositing. To avoid repeated evaluation, source-gate results are reused for candidates sharing the same source, style instruction, and actual stylized output. Removal-gate results are cached separately for each object-removed video, style instruction, and corresponding stylized output. Distinct removal inputs or generated outputs require separate evaluations, even when the source video and style are shared.

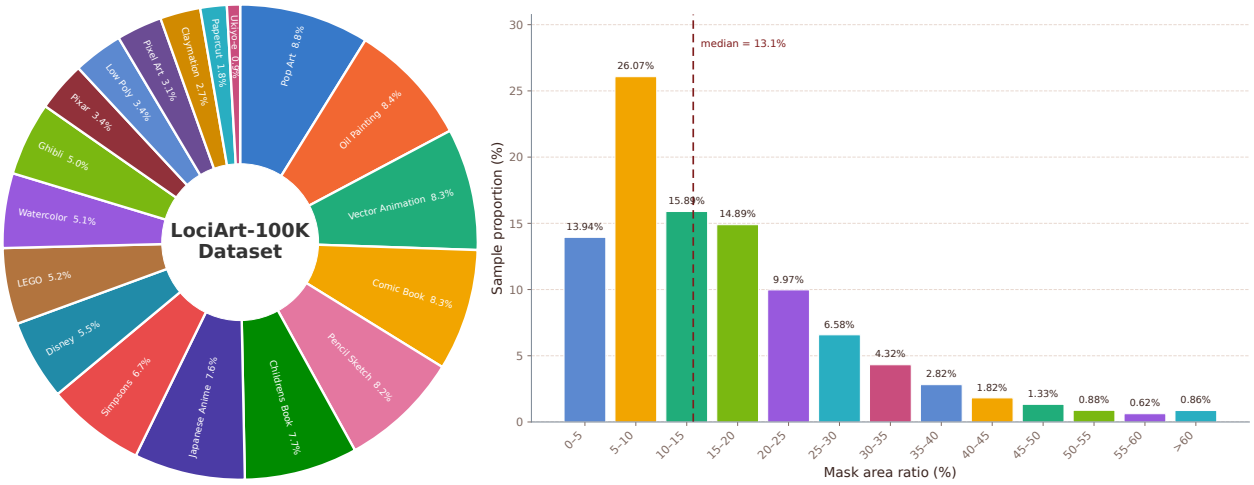
Frame sampling and scoring. At either gate, we sample frame indices 0, 20, 40, 60, and 80 from the aligned 81-frame videos. GPT-5.5 receives ten images in chronological paired order, with an original and a stylized frame at each time point. It scores style adherence, content preservation, structural quality, style consistency, and temporal quality on a 1–5 scale.

Acceptance criteria. The resulting five scores must reach 4, 4, 3, 4, and 4, respectively. In addition, a pair is rejected if it introduces salient objects, removes or duplicates objects, changes semantics or identity, inserts text or logos, or exhibits major layout changes, severe deformation, or ghosting. The total score is reported for reference only and imposes no additional passing threshold.

Evaluation prompt. The following prompt (Listing 1) implements the same sampling, scoring, and acceptance protocol for both gates, with the corresponding input video and stylized output supplied in each case.

Listing 1: VLM prompt for global stylization quality filtering.

```
You are a strict content-preservation evaluator for a full-video stylization result.

Target production style: {production_prompt}
Canonical style category: {style_category}

The input videos contain 81 aligned frames (indices 0 through 80). You receive 10 images from five uniformly sampled time points
in chronological paired order:
- Images 1 and 2: ORIGINAL frame 0 and STYLIZED frame 0
- Images 3 and 4: ORIGINAL frame 20 and STYLIZED frame 20
- Images 5 and 6: ORIGINAL frame 40 and STYLIZED frame 40
- Images 7 and 8: ORIGINAL frame 60 and STYLIZED frame 60
- Images 9 and 10: ORIGINAL frame 80 and STYLIZED frame 80

Each original/stylized pair is the same time point. Compare every pair and use all five time points. Do not decide from the first
pair alone.

Evaluate the requested style while preserving semantic content strictly. Content preservation includes the same people, animals,
objects, object count, identities, roles, actions, interactions, scene layout, camera view, and major background structures.

Do not penalize normal style transformations such as changed colors, simplified details, stylized facial features that preserve
identity, painterly/cartoon/anime/clay/pixel/vector rendering, or moderate texture simplification.

However, be conservative: if an object identity, person identity, object count, action, or scene layout cannot be confidently
matched between an original/stylized pair, score content_preservation below 4 and reject. Do not assume an ambiguous or redesigned
 object is preserved.

Reject a new salient person/object, disappearance or duplication, an object changing semantic category, identity/role/gender
change, action transfer, major layout change, inserted text/logo/graphic, severe deformation, or ghosting.

Judge style adherence against the explicitly requested target style, not against generic stylization or photorealism. Examples:
- Low Poly requires visible facets/polygonal shading.
- Pixel Art requires block-like pixels and a limited palette.
- Claymation requires plasticine material and rounded clay forms.
- Oil Painting requires painterly brushwork.
- Watercolor requires translucent washes.
- Anime or Ghibli requires recognizable hand-drawn animation treatment.
- Other styles must show their characteristic visual language.

Score exactly these five dimensions from 1 to 5:
- style_adherence
- content_preservation
- structural_quality
- style_consistency
- temporal_quality

Passing criteria:
- style_adherence >= 4
- content_preservation >= 4
- structural_quality >= 3
- style_consistency >= 4
- temporal_quality >= 4
- hard_reject_reasons must be empty

Report total as the sum of the five scores for reference only; it is not an additional passing criterion.

Return ONLY one valid JSON object. Do not output analysis, Markdown, or any text outside the JSON object:
{
  "pass": false,
  "style_adherence": 1,
  "content_preservation": 1,
  "structural_quality": 1,
  "style_consistency": 1,
  "temporal_quality": 1,
  "total": 5,
  "hard_reject_reasons": [],
  "reason": "Brief evidence-based explanation"
}
```



(a) Distribution of style categories

(b) Foreground object mask area ratios

Figure 8: Statistics of the LociArt-100K Dataset. The 100K qualified pairs cover 18 style categories. Area ratios describe foreground object masks $\mathbf{M}_{\text{obj}}$, not the complementary background editing masks.

A.3 Human Verification Details

After matting and compositing, human review checks the completed local edits for failures that global stylization filtering may miss. Reviewers assess each source–mask–instruction–target example against the criteria below, then retain qualified pairs to form the final dataset.

Verification requirements. A qualified sample must satisfy four requirements: (i) the requested style is expressed while semantic content, object identities and counts, actions, and scene layout remain consistent with the source; (ii) non-target content is preserved without unintended style spillover; (iii) masks and compositing boundaries remain coherent, without incomplete regions or visible unstylized remnants; and (iv) the video is temporally consistent, without frame breakage or flicker. These checks distinguish unintended content changes from valid artistic transformations: style-consistent changes in color, texture, detail, and object contours are permitted.

Acceptance and archiving. Candidates meeting all four requirements are labeled qualified and retained; those failing the checks are rejected and discarded. Typical rejection cases include weak or incorrect style, semantic or identity changes, inserted or missing objects, unintended changes to preserved content, incomplete masks, and temporal artifacts. Representative examples rejected during human verification are shown in Fig. 9 and on the anonymous demo page: `https://lociart-page.pages.dev/#robustness`.

Production statistics. The pass rate is 52.4% for global stylization VLM filtering and 65% for human verification of the resulting local composites. These rates are measured at their respective filtering stages. The final dataset contains 100K qualified video pairs covering 18 style categories. Figure 8 summarizes the style distribution and foreground object mask areas. The area histogram characterizes $\mathbf{M}_{\text{obj}}$, not its complementary background editing mask, and therefore does not represent the distribution of all editing regions.

B Style-Code Design and Local Readout Analysis

B.1 Codebook Assignment and Input Normalization

We use a fixed 100-slot circular codebook. For $K \leq 100$, let $\nu : \{1, \ldots, K\} \to \{0, \ldots, 99\}$ assign each style to a distinct slot, with $\theta_s = 2\pi\nu(s)/100$. Well-spaced slots are assigned deterministically, and the mapping remains fixed across training and inference. The resulting directions satisfy $\|\boldsymbol{a}_s\|_2 = 1$, distinguishing styles by angle rather than magnitude, analogous to the equal-norm one-hot bit codes in GRN (Han et al., 2026).



Figure 9: Representative rejection cases from the data pipeline.

Following the main-text definition in Equation 3, the style-aware mask is

$$\mathbf{M}_s(q) = \mathbf{M}(q)\left(\frac{1+a_{s,1}}{2}, \frac{1+a_{s,2}}{2}, 1\right), \qquad \boldsymbol{a}_s = (\cos\theta_s, \sin\theta_s). \tag{9}$$

The VAE input normalization $x \mapsto 2x-1$ therefore yields $c_s = (a_{s,1}, a_{s,2}, 1)$ at edited positions and $e = (-1,-1,-1)$ at unedited positions. Thus, the normalized third channel retains the same binary occupancy cue for every style, while the first two channels encode its direction.

B.2 Why Circular Rather Than Spherical Codes?

A three-coordinate spherical code offers greater class-packing capacity, but uses every channel for identity. Local stylization also needs a stable distinction between edit and keep. We therefore reserve the third channel for occupancy and use the first two for style directions. The shared occupancy channel avoids requiring edit/keep status to be inferred from class-dependent coordinates.

A larger empty-condition margin. Reserving this occupancy axis keeps style codes away from the empty condition. For the circular and origin-centered unit-sphere alternatives,

$$\begin{aligned} \min_{\theta\in[0,2\pi)} \|(\cos\theta, \sin\theta, 1) - e\|_2 &= \sqrt{7-2\sqrt{2}} \approx 2.04, \\ \min_{v\in\mathbb{R}^3:\,\|v\|_2=1} \|v - e\|_2 &= \sqrt{3}-1 \approx 0.73. \end{aligned} \tag{10}$$

The circular representation therefore keeps every style farther from the empty condition than the spherical alternative. For the 100-slot codebook, the smallest distance between adjacent normalized style directions is $2\sin(\pi/100) \approx 0.063$; the empty-condition margin is the larger value $\sqrt{7-2\sqrt{2}} \approx 2.04$.

A shared cue for the frozen encoder. These coordinates denote category identity, not visual color. The pretrained VAE nevertheless processes them through learned RGB features, so greater spherical spacing alone does not resolve the encoding problem. A binary occupancy channel supplies a common boundary cue across styles, while class-dependent variation remains confined to the other two channels. Our 100-slot VAE/HBQ validation found no exact token collisions. We therefore favor explicit occupancy and empty-condition separation over additional spherical capacity.

B.3 Interaction Error and Exact Local Readout

This section isolates the style–region interaction in Equation 4. We compare an additive readout of separate conditions with a linear readout of the polar field, using the same region and style information.

Setup. At one pixel position, write $m = M(q) \in \{0, 1\}$ and let $a_s \in \mathbb{R}^2$ be the unit style direction in Equation 3. The desired local instruction is ma_s. To separate representation from dataset-specific correlations, consider independent $m \sim \text{Bernoulli}(\alpha)$ and $s \sim \pi$, where $0 < \alpha < 1$, and define $\mu = \sum_s \pi_s a_s$. For additive readouts $f(m, s) = f_M(m) + f_S(s)$, the component functions may be arbitrary. We measure squared instruction error $\mathcal{J}(f) = \mathbb{E}\|ma_s - f(m, s)\|_2^2$.

Optimal additive error. The best additive readout and its error are

$$f^\star(m, s) = m\mu + \alpha(a_s - \mu), \tag{11}$$

$$\min_{f=f_M+f_S} \mathcal{J}(f) = \alpha(1-\alpha)\big(1 - \|\mu\|_2^2\big). \tag{12}$$

To see this, decompose $ma_s = f^\star(m, s) + (m - \alpha)(a_s - \mu)$. The residual has zero conditional mean given either m or s, so it is orthogonal to every additive readout. Consequently, $\mathcal{J}(f) = \mathcal{J}(f^\star) + \mathbb{E}\|f - f^\star\|_2^2$. Independence and $\|a_s\|_2 = 1$ then give Equation 12. The factor $1 - \|\mu\|_2^2$ is the variance of the style directions. The minimum is positive when at least two distinct styles have positive probability. It is at most $\alpha(1 - \alpha) \le 1/4$, with equality at $1/4$ when $\alpha = 1/2$ and $\mu = 0$.

Exact polar readout. For the same inputs, the encoded pixel $\mathbf{M}_s(q)$ defined in Equation 9 satisfies

$$\begin{pmatrix} 2 & 0 & -1 \\ 0 & 2 & -1 \end{pmatrix} \mathbf{M}_s(q) = ma_s, \qquad \mathbb{E}\|\mathcal{R}(\mathbf{M}_s(q)) - ma_s\|_2^2 = 0. \tag{13}$$

Polar encoding therefore supplies the interaction explicitly, without adding a new condition. This is a distinction between readout classes, not an error bound on a nonlinear generator: a nonlinear network can learn the interaction from separate inputs, whereas our representation computes it before encoding.

C Details of the Evaluation Protocol

The LociArt Benchmark evaluates target-region style, non-target preservation, and temporal consistency separately. We combine Style Prototype Similarity (SPS), regional CLIP-T, preservation PSNR, and full-frame temporal consistency (TC), together with two-dimensional VLM assessment.

C.1 Benchmark Construction and Evaluation Units

The benchmark contains 80 source videos and 84 object-mask units, each with 81 aligned frames. Videos with two selected subjects contribute separate object-mask units. Of the 80 videos, 13 come from the rejected Ditto (Bai et al., 2026) pool and were excluded from our training data, 46 come from the DAVIS 2017 training split (Pont-Tuset et al., 2017), and 21 come from Pexels[1]. The foreground area ratios range from 0.02 to 0.58, and 29 units contain human subjects.

Styles and comparison coverage. In the quantitative and perceptual comparison, all methods receive the same input videos for the same target style, and frame-alignment protocol for shared cases. Table 4 lists the 18-style vocabulary and the ten-style comparison subset (`bench10`). Each object-mask unit supports foreground and background editing, so the complete ten-style grid contains $84 \times 10 \times 2 = 1{,}680$ results (168 cases, 10 styles).

C.2 Style Prototype Similarity (SPS)

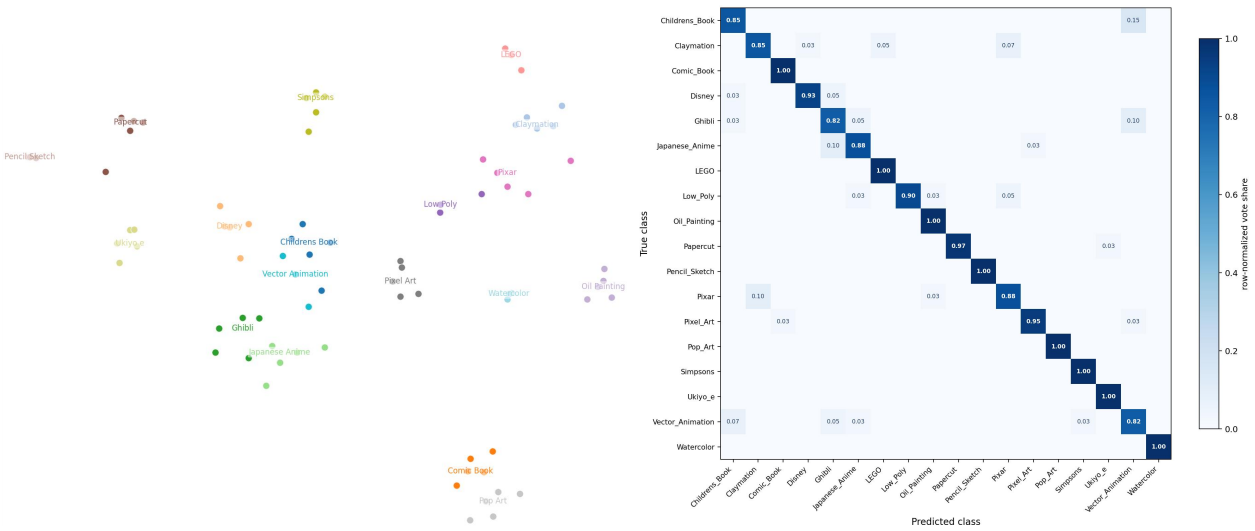
SPS evaluates whether an edited region resembles characteristic visual examples of the requested style. Unlike text–image similarity, it compares outputs with a class-specific bank of visual prototypes. SPS is the prototype-based scoring scheme built on the Contrastive Style Descriptors (CSD) (Somepalli et al., 2024).

Prototype construction. We collect globally stylized source and object-removed videos that pass the data pipeline's quality gates. Every frame is encoded into a 768-dimensional CSD descriptor. Within each style, k-means clustering produces multiple centroids, with $k \in \{3, 5\}$ selected by the silhouette score. The resulting bank contains 80 centroids across 18 styles. Multiple prototypes represent within-style variation rather than forcing distinct rendering modes into a single center. Fig. 10a visualizes the prototype bank with t-SNE (Van der Maaten & Hinton, 2008).

[1] `https://www.pexels.com/`

Table 4: Canonical style phrases for regional CLIP-T. A check mark identifies the ten-style comparison subset.

Style category	Canonical phrase	in `bench10`
Children's Book	children's book illustration	✓
Claymation	claymation	✓
Comic Book	comic book	
Disney	Disney animation	
Ghibli	Studio Ghibli anime	✓
Japanese Anime	Japanese anime	
LEGO	LEGO brick	
Low Poly	low-poly 3D	✓
Oil Painting	oil painting	✓
Papercut	papercut	✓
Pencil Sketch	pencil sketch	✓
Pixar	Pixar 3D animation	
Pixel Art	pixel art	✓
Pop Art	pop art	
Simpsons	The Simpsons cartoon	
Ukiyo-e	Ukiyo-e woodblock print	✓
Vector Animation	flat vector animation	
Watercolor	watercolor painting	✓



(a) Style prototype visualization.

(b) Nearest-prototype classification.

Figure 10: **SPS prototypes and discriminative validation.** (a) t-SNE visualization of 80 CSD prototypes across 18 styles; each point is a prototype, colored by style. (b) Row-normalized confusion matrix under leave-one-source-out validation, with a macro-averaged classification accuracy of 93.61%.

Region extraction. Let $\mathbf{Y}$ be an output video and $\mathbf{M}_{\mathrm{obj}}$ its tracked foreground-object mask. VideoMaMa (Lim et al., 2026) mask-guided matting of the output yields a foreground alpha video $\mathbf{A}_{\mathrm{out}}$, allowing region extraction to follow stylization-induced contour changes. For foreground editing, each frame is cropped to the alpha-support bounding box with a 2% margin. For background editing, the full frame is retained while foreground-object tokens are excluded from the region representation, avoiding an artificial filled foreground area. Denote the corresponding region descriptor by $\psi_d(\mathbf{Y}_t, \mathbf{A}_{\mathrm{out},t})$, where $d \in \{\mathrm{fg}, \mathrm{bg}\}$ specifies the editing direction.

Frame and video scores. For target style c and its prototype set $\mathcal{P}_c$, we compute

$$q_{t,c} = \max_{\boldsymbol{p} \in \mathcal{P}_c} \cos(\psi_d(\mathbf{Y}_t, \mathbf{A}_{\mathrm{out},t}), \boldsymbol{p}) \,, \qquad \mathrm{SPS} = \frac{1}{N} \sum_{t=1}^{N} q_{t,c}. \tag{14}$$

Frame scores are first averaged within each evaluated video; dataset scores then average the corresponding video-level scores, with foreground and background editing reported separately.

Discriminative validation. To test whether SPS distinguishes the 18 style categories beyond source-specific appearance, we perform leave-one-source-out validation: all samples derived from the held-out source video are excluded when constructing its reference prototypes. Each frame is assigned the style of its nearest prototype by cosine similarity, and frame predictions are aggregated by video-level majority vote. This yields a macro-averaged classification accuracy of 93.61%. Fig. 10b shows the class-wise confusion matrix, complementing the prototype visualization in Fig. 10a. Together, these analyses characterize the prototype organization and support SPS's ability to distinguish the evaluated style categories.

C.3 Regional CLIPT and Temporal Consistency

Regional CLIPT measures alignment with the canonical style phrase, complementing SPS's visual prototypes. It uses the same foreground crops or background token selection as SPS, but extracts image and text features with CLIP ViT-L/14 (Radford et al., 2021). Let $\chi_d(\mathbf{Y}_t, \mathbf{A}_{\mathrm{out},t})$ denote the region image embedding and $\eta(t_c)$ the embedding of the phrase for style c in Table 4. Then

$$\mathrm{CLIPT}_{\mathrm{reg},t} = \cos(\chi_d(\mathbf{Y}_t, \mathbf{A}_{\mathrm{out},t}), \eta(t_c)) \,. \tag{15}$$

As with SPS, video-level means are averaged before dataset aggregation. The alpha video defines region selection; it is not itself the RGB image passed to the image encoder.

To assess temporal stability separately from regional style, TC (Temporal Consistency) uses full-frame image embeddings $\boldsymbol{z}_t$ from a CLIP ViT-H/14 tower, a separate and larger encoder than the ViT-L/14 used for regional CLIPT:

$$\mathrm{TC} = \frac{1}{N-1} \sum_{t=1}^{N-1} \cos(\boldsymbol{z}_t, \boldsymbol{z}_{t+1}), \qquad \boldsymbol{z}_t = \mathrm{CLIP}_{\mathrm{img}}(\mathbf{Y}_t). \tag{16}$$

TC summarizes adjacent-frame feature consistency over the aligned output sequence. It is a global measure and does not replace regional style or preservation scores.

C.4 Non-Target Preservation

Preservation is measured with regional PSNR (Hore & Ziou, 2010) outside the intended edit. For foreground stylization, PSNR is computed on the preserved background region $1 - M$; for background stylization, it is computed on the preserved foreground region M. In both cases, the valid region is derived from the matted alpha on an 8-bit scale—the preserved background core uses $\alpha > 200$ and the preserved foreground uses $\alpha \leq 5$—excluding the anti-aliased transition band in between.

We compare aligned output and source frames after resizing the source to the output resolution and use the common valid temporal range. For each frame, the RGB reconstruction error is averaged over the valid non-edit pixels and channels, and the resulting errors are aggregated over valid frames before conversion to PSNR. Frames with insufficient valid pixels are skipped; when an alpha estimate is empty, the original mask is used as a fallback.

C.5 Details of VLM Evaluation

We use Gemini 3.6 Flash to complement the objective metrics with full-video VLM assessment of two separate quantities: *Style* (`edit_region_style`) and *Preserve* (`non_edit_preservation`). Each is an integer from 0 to 5. The scores are not merged, so strong stylization cannot compensate for damage to non-target content, or vice versa. This evaluation is distinct from the GPT-5.5 data-construction filter, which uses five dimensions on a 1–5 scale.

Inputs and execution. The judge receives the source video first and the edited result second, aligned frame by frame and tagged `SRC` and `RESULT`. A red contour marks the foreground object in the source. For foreground editing, its interior is the target; for background editing, the exterior is the target. The text instruction explicitly identifies the direction, and the contour and display tags are not scored. We use the English system prompt and rubric below with temperature zero. Responses are parsed as JSON; if parsing fails, the same request is retried once.

Scoring anchors. *Style* jointly assesses style correctness, coverage of the edit region, and cross-frame stability. A score of 5 requires all three, 3 indicates visible but imperfect style, incomplete coverage, or occasional flicker, and

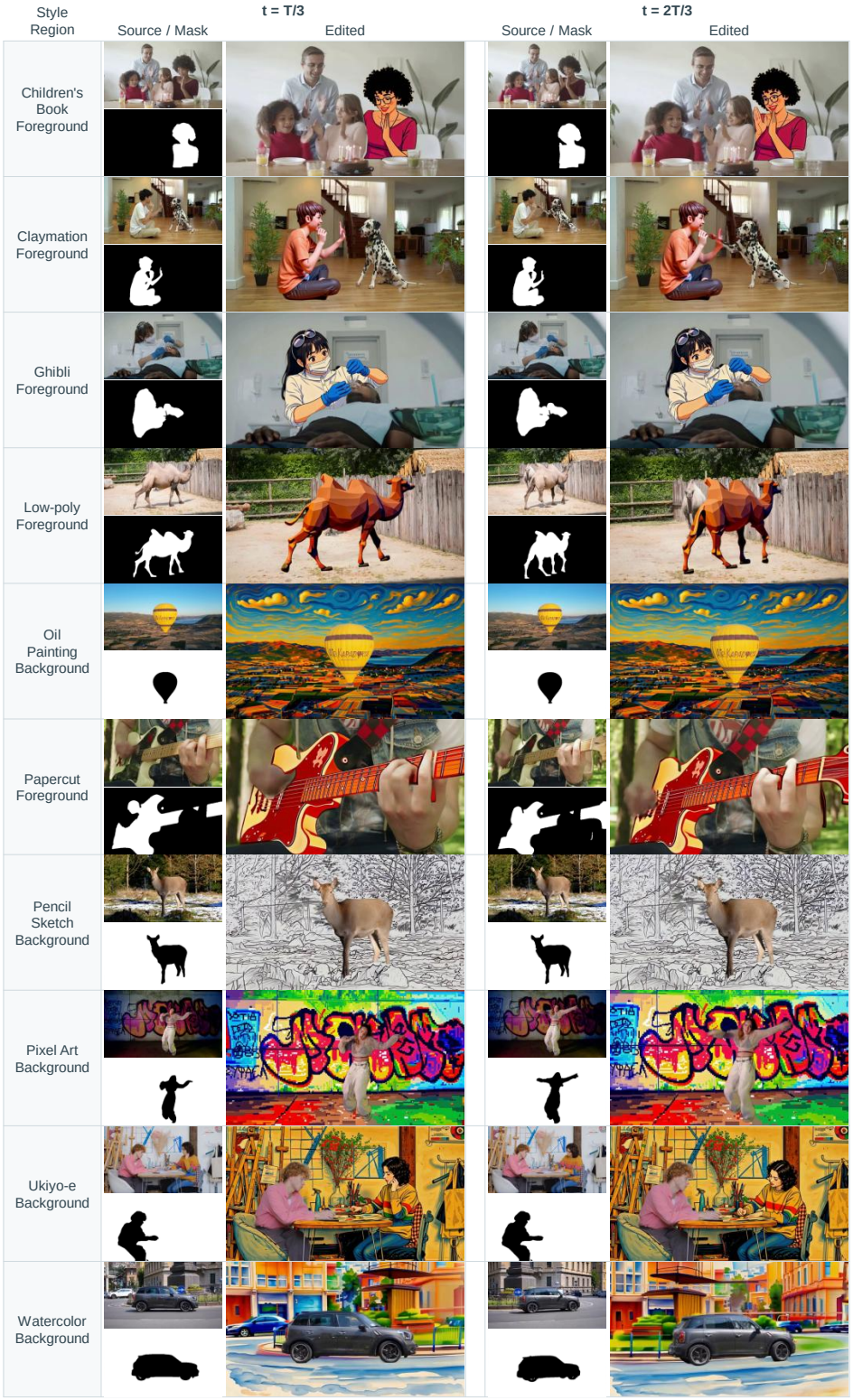
0 indicates absent or incorrect style or severe instability. *Preserve* assesses source fidelity and temporal stability in the non-edit region: 5 indicates a close, stable match, 3 allows minor changes or occasional jitter, and 0 indicates substantial repainting or severe drift. The two video-level scores are aggregated separately across cases.

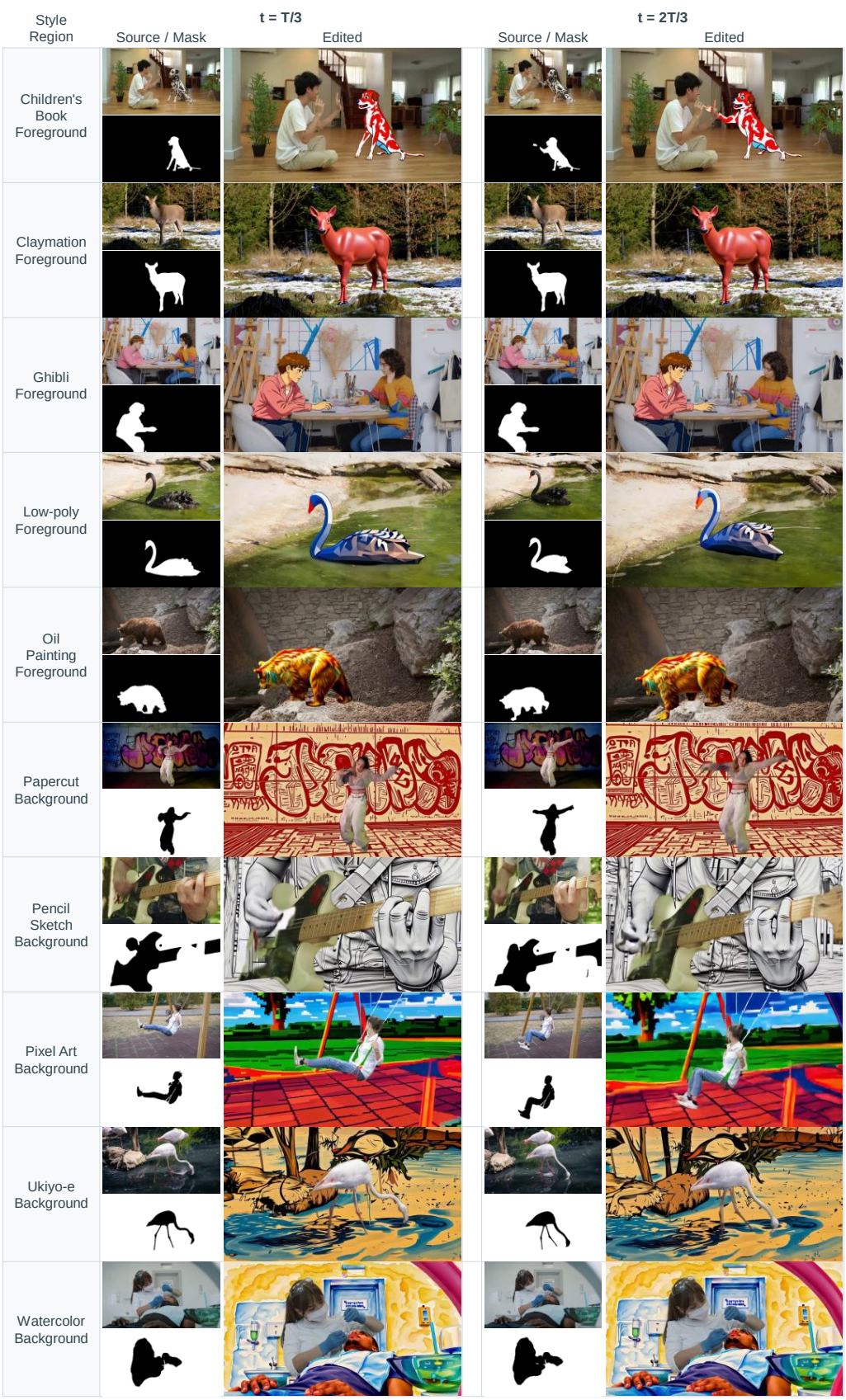
Prompt templates. The following templates retain the English scoring rubric while making the edit direction explicit. For foreground editing, {target_region} is "the foreground object inside the red outline"; for background editing, it is "the background outside the red outline." The {region_description} field names both the edit and non-edit regions accordingly, and {style} is replaced by the requested style.

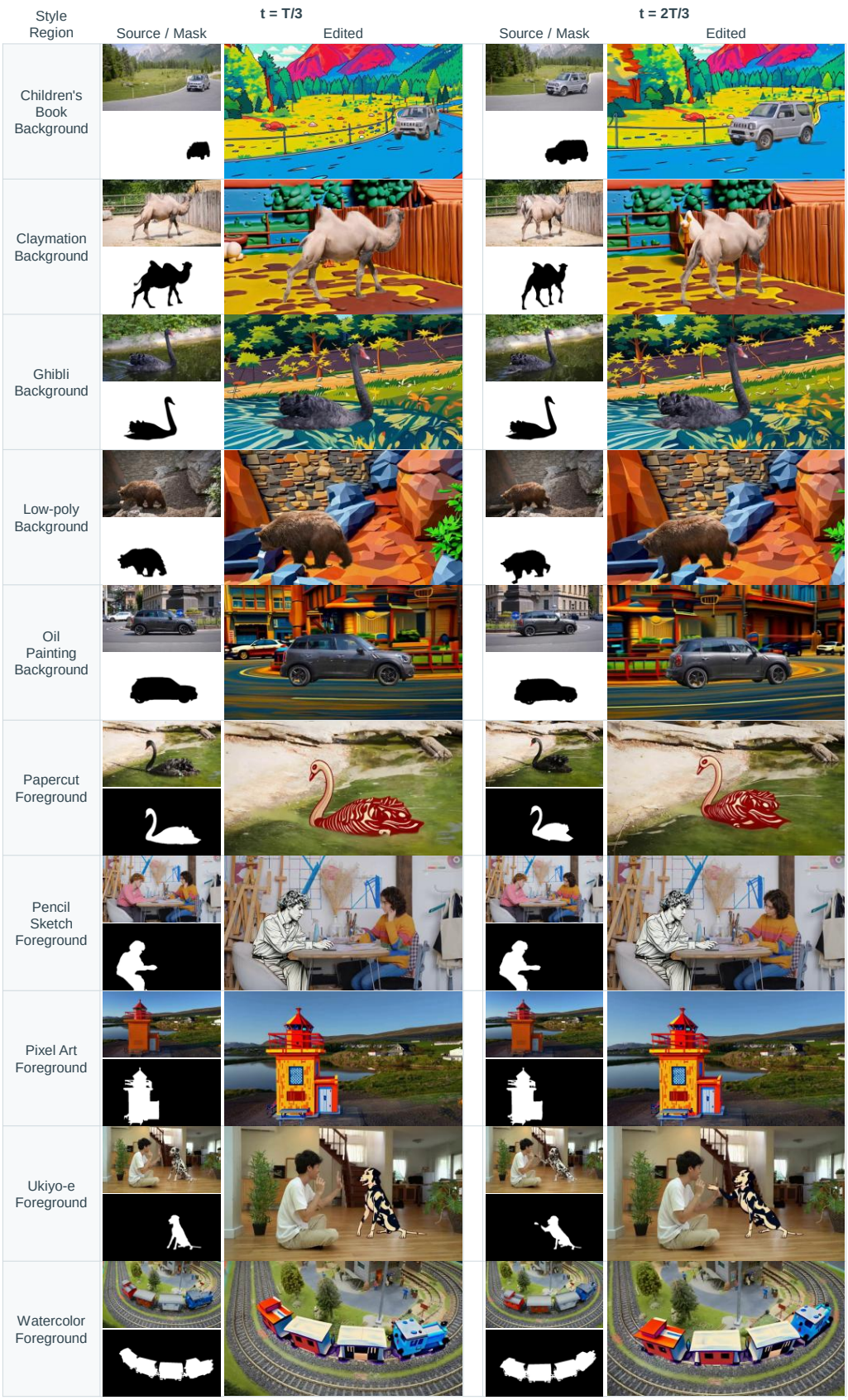
Listing 2: VLM-as-a-judge prompt in our evaluation.

```
You are a professional video local-editing judge. This review contains two videos: the first is the source
video (the foreground object is outlined in red, with an "SRC" tag in the top-left corner), and the second is
the edited result video (tagged "RESULT" in the top-left corner). The red outline and any display differences
caused by it as a marker are not scored; judge only the actual visual changes inside the outlined region and in
 the rest of the frame. Watch both videos in full and compare them segment by segment before scoring. Score
strictly based only on what is visible; do not deduct for video encoding noise, and do not let the order of the
 videos bias you. Output only one line of valid JSON, with no code fences or any other text.


[Task] Local video stylization review
- Edit request: convert only {target_region} in the source video to the {style} style
- Region: {region_description}

[Scoring dimensions] Each dimension is an integer from 0 to 5; score ONLY the following two dimensions:
1. edit_region_style -- quality of the stylization in the edit region (combining style correctness, coverage
completeness, and cross-frame stability):
   (i) Stylization: does it present a pure, typical {style} style
   (ii) Coverage: does the style evenly cover the entire edit region (no gaps, no missed spots)
   (iii) Temporal stability: the appearance of the edit region stays stable and coherent over time, with no
flicker and no frame-by-frame style drift
   5 = a pure, typical {style} style evenly covering the entire edit region, with no noticeable abnormal
flicker or style drift throughout; if any aspect clearly fails, a 5 must not be given
   3 = stylization is present but not faithful to the style, or coverage is incomplete / parts of the region
are missed, or occasional flicker
   0 = almost no stylization, the style is completely wrong, or the edit region flickers or drifts severely
2. non_edit_preservation -- preservation of the non-edit region (combining):
   (i) Consistency with the source: at every corresponding moment the non-edit region keeps the content,
position, texture, and colors of the source video as closely as possible
   (ii) Temporal consistency: the non-edit region stays stable over time, with no frame-by-frame jitter and no
color-shift drift
   5 = the non-edit region matches the source video closely and is temporally stable throughout
   3 = only minor changes, slight color shifts, or local differences, and most of the non-edit region still
matches the corresponding source frames; or only occasional jitter
   0 = the content, texture, colors, or details of the non-edit region are visibly repainted or stylized, or
severe frame-by-frame jitter appears

[Output] Output only the following JSON and nothing else:
{"edit_region_style":n,"non_edit_preservation":n,"comment":"summarize the most prominent issue in at most 20
English words; write none if there is no obvious issue"}
```

D ADDITIONAL QUALITATIVE RESULTS




Style Region
t = T/3
Source / Mask
Edited
t = 2T/3
Source / Mask
Edited
Children's Book Foreground
Claymation Foreground
Ghibli Foreground
Low-poly Foreground
Oil Painting Foreground
Papercut Background
Pencil Sketch Background
Pixel Art Background
Ukiyo-e Background
Watercolor Background


Style Region
t = T/3
Source / Mask
Edited
t = 2T/3
Source / Mask
Edited
Children's Book Background
Claymation Background
Ghibli Background
Low-poly Background
Oil Painting Background
Papercut Foreground
Pencil Sketch Foreground
Pixel Art Foreground
Ukiyo-e Foreground
Watercolor Foreground


Style Region
t = T/3
Source / Mask
Edited
t = 2T/3
Source / Mask
Edited
Children's Book Background
Claymation Background
Ghibli Background
Low-poly Background
Oil Painting Foreground
Papercut Background
Pencil Sketch Foreground
Pixel Art Foreground
Ukiyo-e Foreground
Watercolor Foreground

E Model Reproduction Details

E.1 Hyperparameters

Table 5 summarizes our main GRNStyle 2B training configuration, using 100K paired metadata entries across 18 styles. Section E.2 reports the learning curves of this initial base-initialized run, separately from the later main configuration summarized here.

Table 5: Main training settings for GRNStyle 2B.

Hyperparameter	Setting
Model and representations	
Backbone	GRN 2B
Transformer blocks / hidden width / heads	28 / 2,304 / 18
Transformer blocks per chunk	4
VAE latent channels / HBQ rounds	64 / 4
Temporal compression factor	4
Text feature width / maximum tokens	4,096 / 512
Data and conditioning	
Dataset entries / style categories	100,000 / 18
Video frames / sampling rate	61 / 20 fps
Aspect-ratio buckets	15 buckets, $H/W \in [0.333, 3.0]$
Class-aware polar mask	Enabled
Generic-prompt replacement probability	0.50
Text-condition dropout probability	0
Packed token budget per rank	110,000
Optimization	
Optimizer / momentum coefficients†	AdamW / $(0.9, 0.999)$
Weight decay	0.01
Gradient clipping (global norm)	0.1
Gradient accumulation	1 update per micro-batch
Training objective†	Official all-token bit cross-entropy
Mixed precision / TF32	BF16† / enabled
Learning-rate schedule	
Initial / peak / minimum learning rate	10^{-5} / 4×10^{-5} / 2×10^{-5}
Linear warmup	100 optimizer updates
Post-warmup decay interval / factor	2,000 updates / 0.5
First decay reaching the minimum LR	Update 2,100
Training duration	10 epochs; 18,490 updates
Execution and saving	
Distributed ranks / sharding	8 / FSDP full sharding†
Activation checkpointing	Full transformer blocks
Data-loader workers per rank	2
Checkpoint interval	1,000 optimizer updates
Random seed	1,789,151,675

†Implementation-verified details; remaining entries are reported or directly reconstructed from the log. Batches are token-budgeted rather than fixed by video count. The learning rate remains at its minimum after update 2,100. Generic-prompt replacement retains a text condition and is distinct from text-condition dropout.

E.2 Training and Validation Curves

We initialize from a well-trained GRNEdit general-purpose video editor with strong stylization capabilities. Training thus focuses on mask-guided local control rather than learning stylization from scratch, consistent with the relatively stable global loss and accuracy. Region-wise validation of edited-region stylization and unedited-region preservation reveals clearer improvements in this local control.

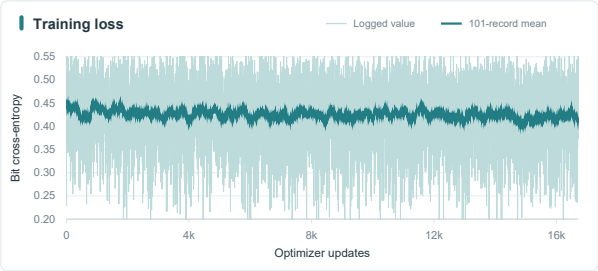


Figure 11: Training bit cross-entropy.

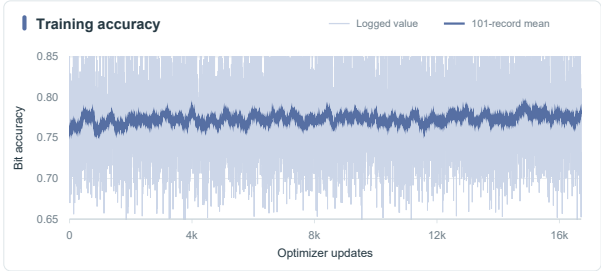


Figure 12: Training bit-prediction accuracy over the same optimizer updates.

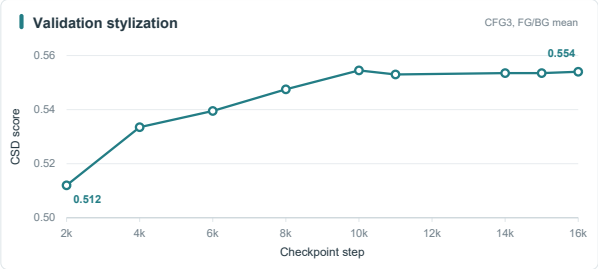


Figure 13: Validation stylization: mean foreground/background SPS.

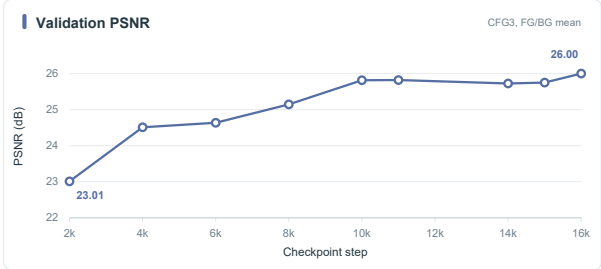


Figure 14: Validation PSNR: mean foreground/background score.

Figures 11 and 12 show 8,365 logged observations across 16,710 optimizer updates. This initial run uses 40,110 metadata entries, a peak learning rate of 5×10^{-5} and an 80,000-token budget per rank; these differ from the later settings in Table 5. Loss is the bit-prediction cross-entropy, and ACC is bit-prediction accuracy, not style-classification accuracy. The plots use the master-rank per-record values rather than the cumulative epoch averages printed in parentheses. Pale traces are clipped to the displayed ranges (ACC is shown as a fraction); solid lines show a centered mean of the unclipped observations over 101 records (approximately 200 updates, with shorter windows at the two endpoints).

Validation performance. Figures 13 and 14 report stylization (SPS) and PSNR on the validation set. At each checkpoint, we take the unweighted arithmetic mean of the reported foreground and background scores. Markers denote the nine evaluated checkpoints; connecting lines are not smoothed. From update 2,000 to 16,000, mean SPS increases from 0.512 to 0.554, and mean PSNR from 23.01 to 26.00 dB. Both metrics show smaller changes from update 10,000 onward.

F Limitations and Future Work

Our current system has two notable limitations. First, temporal consistency can degrade for fast-moving objects, especially near motion boundaries. Rapid motion makes it difficult to maintain stable region alignment and coherent stylization across adjacent frames, which may lead to local flicker or inconsistent boundaries. Improving motion-aware conditioning and boundary tracking is an important direction for future work.

Second, the data-construction pipeline relies on an object-removal model that may also remove the object's cast shadow. As a result, some paired targets lack the original shadow even when it should remain semantically consistent with the scene. Training on such pairs can encourage the stylization model to erase shadows, producing visually unnatural results. Future work should improve shadow-aware data construction and incorporate explicit supervision for preserving scene-consistent shadows.