FILM STYLE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

In film post-production, a colorist needs to apply the visual look of a whole film — its *film look* — to target frames. The standard approach is manual colour grading, accurate but labour-intensive. We propose to automate this step by introducing a novel computer vision task, termed *film style transfer* (FST): a set of reference frames sampled from a film represents its film look, which needs to be transferred to target frames. One intuitive solution is to leverage off-the-shelf style transfer methods. However, this suffers several issues: (1) Existing methods often assume *a single style image*, whereas FST is defined over a set of reference frames; (2) applying the film look uniformly over target frames tends to degrade their fine structure, rendering such direct application suboptimal. To address these, we introduce a dedicated FST method, *AnswerPrint*, that starts with film look extraction, followed by a coarse-to-fine transformation: first transferring the global film look, then recovering the fine content details the global transfer destroys, echoing the spirit of the established colorist workflow. Our method integrates with existing image generative models (*e.g.*, Stable Diffusion) without extra training. To enable evaluation, we build *Reel-30*, a new FST benchmark with thirty films as well as metrics in film identity (source-film recognisability) and structural fidelity (local structure preservation). Extensive experiments show that AnswerPrint excels over existing alternatives (*e.g.*, LUT, StyleID). Code and data will be released.

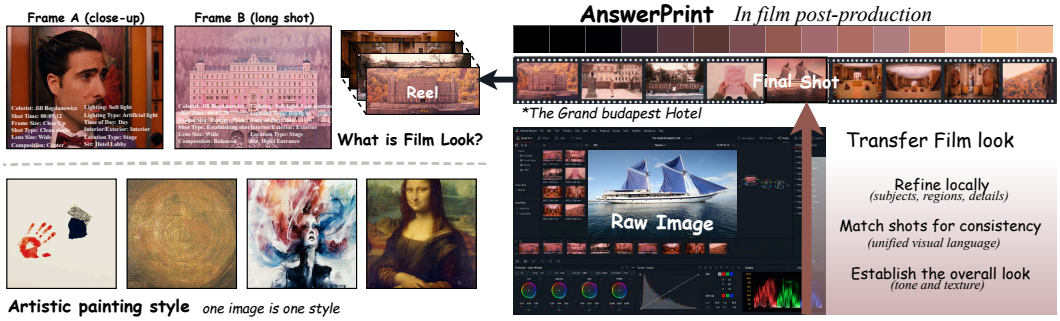

Figure 1: **(Right)** Our AI colorist **AnswerPrint** works like a colorist in film post-production: it first transfers a film's look onto a raw shot holistically, establishing the overall tone and texture (primary) for a unified visual language; Second, it refines locally to preserve fine content structure (secondary). Ultimately, the graded shot sits seamlessly within the film. **(Left)** In artistic painting style transfer, one image is one style. However, film is different: a close-up and a long shot from *The Grand Budapest Hotel* differ in nearly every measurable attribute, from framing and lighting to palette, yet both are unmistakably Wes Anderson. A film look is a property of the *reel* as a whole, not of any single frame, and thus we define the *film-style transfer* (FST) task.

1 INTRODUCTION

In film production, raw footage does not yet carry the *film look* it will eventually have. Producing that look is a separate stage, post-production (Giorgianni & Madden, 2009; Van Hurkman, 2014), carried out by a colorist who color-grades each shot, unifying footage captured under different lighting, cameras, or days into one consistent look. Typically, the colorist first applies *a primary grade*, an initial correction spanning the whole film, before reconciling it with each shot through *a secondary grade*: relighting a face with a power window (a shape-based mask) or recovering a highlight with a

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

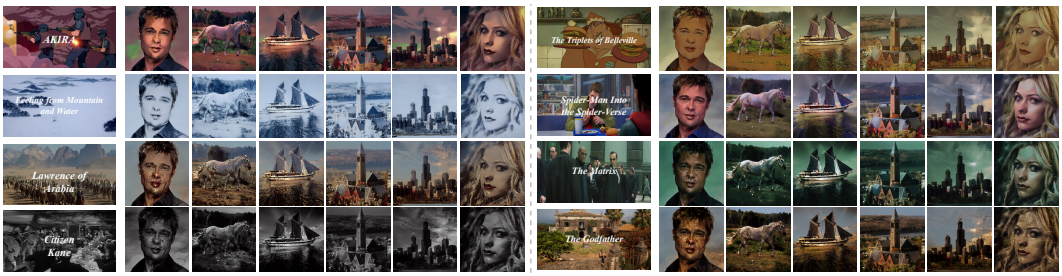

Figure 2: Film look transfer results of APrint. Zoom in for better visualization.

qualifier (a color- or luminance-based mask). Once established, this film look must then be applied to any new footage, *e.g.*, reshoots, inserted VFX plates, or other frames, that needs to match it. This process is labour-intensive: a feature film runs to hundreds of graded shots, and a single shot can take the colorist tens of minutes to bring from an ungraded state to an accepted match.

We propose to automate this step by introducing a novel task, termed *film style transfer* (FST): a set of reference frames sampled from a film represents its film look, which needs to be transferred to target frames. In particular, film style is highly specialized with significance on static tone and texture as focused here. For example, as shown in figure 1, a close-up and an establishing shot from Wes Anderson's *The Grand Budapest Hotel* differ in framing, lighting, and palette, yet share a film look any viewer recognizes at a glance.

A natural solution is to adopt off-the-shelf style transfer methods (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022), but two issues make this less effective. First, these methods consider style from a single artistic image, whereas film style is defined by a set of reference frames, thus presenting an input mismatch. Second, they typically perform a global transfer, at the risk of distorting target's fine structure.

To address these issues, we propose *AnswerPrint*[1], a training-free FST method with two passes as colorist's workflow: (1) The primary grade is a coarse, global transfer of the film look, extracted from the pooled statistics of the reference frames collectively; (2) The secondary grade leverages a sparse graph over the content's original structure to anchor the restoration of fine detail while preserving the transferred film look. Besides, we remove a near-constant bias at the grading process, much as one clears a stray hair from the film gate to lift a fixed contamination from every frame. To enable evaluation, we construct a new benchmark, *Reel-30*[1], including thirty films across eight categories, from animation to live action, monochrome to saturated, as well as metrics in film identity and structural fidelity.

Our *contributions* are: **(1)** a new vision task FST, practically useful in film post-production; **(2)** a novel training-free FST algorithm in the spirit of professional workflow; **(3)** a new benchmark carefully designed for the FST research; **(4)** extensive experiments showing that AnswerPrint not only outperforms existing alternatives but also presents competitive results in comparison with the professionals with significant time saving.

2 Related Work

Automating the look has a long lineage in colour science, from global colour transfer between an image pair (Reinhard et al., 2001) and the closed-form Monge–Kantorovich map underlying LUT-based grading (Pitié & Kokaram, 2007) to learned colour-histogram control (Afifi et al., 2021); these are pointwise methods that lack learning capabilities and their performance depends heavily on human input. With pre-trained DMs, the style-transfer field has largely abandoned per-pair optimization in favour of intervening in a frozen generator. Embedding-based methods project the reference into the text encoder's space, as a learned token (Zhang et al., 2023) or an inverted prompt embedding (Wang et al., 2023), so that it reaches the sampler only through language. Others steer the sampling trajectory with an external objective (Kwon & Ye, 2023), route content and style through

[1]Both names borrow from film production: *answer print* is the first fully colour-graded copy against which release prints are matched (Encyclopædia Britannica, 2026); we repurpose *reel*, traditionally a spool of film, to mean a curated set of representative frames.

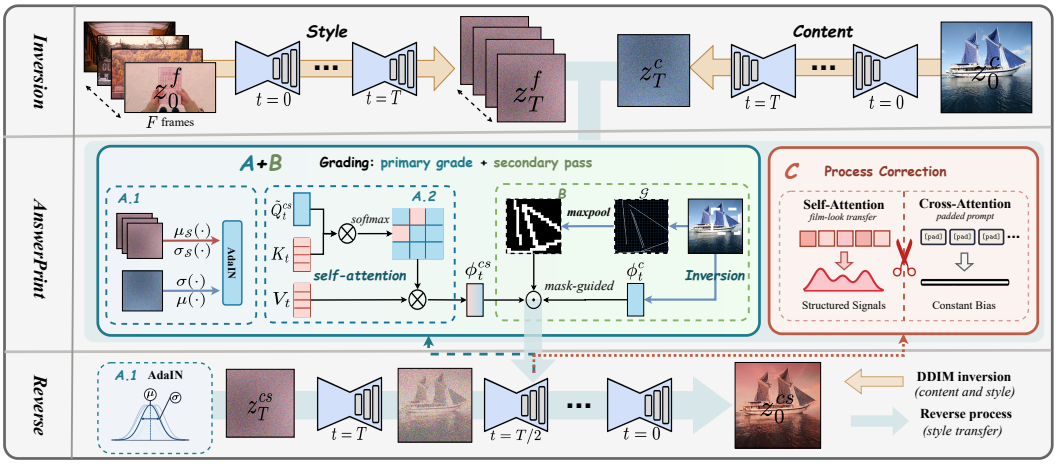

Figure 3: **Overview of AnswerPrint.** Given the inverted set, there are two passes in sequence: A. a primary grade that carries the film's global look from the set (section 4.1): AdaIN on the starting latent (A.1) and attention-bank injection (A.2); B. a secondary pass that rasterizes a sparse structure graph $\mathcal{G}$ into a mask and re-instates the content's attention ϕ_t^{cs} at those locations (section 4.2). Besides, a process correction cuts the CA's constant, content-independent bias while leaving the film-look SA intact (section 4.3). These passes are applied at every step of the DDIM reverse process, denoising z_T^{cs} into the stylized output with the backbone kept frozen.

separate latent pathways (Jeong et al., 2024), or add a trained image-conditioning branch (Ye et al., 2023; Wang et al., 2024). The current mainstream keeps the generator intact and edits only the attention layers: StyleID (Chung et al., 2024) substitutes the reference's self-attention keys and values while preserving the content's queries, StyleAligned (Hertz et al., 2024) shares attention across a batch, and RB-Modulation (Rout et al., 2025) and StyleSSP (Xu et al., 2025) refine feature aggregation and the sampling start point respectively. All share a single-image reference and we instead address the practically important setting where the reference is a reel.

Pre-trained DMs have enabled many inversion-based methods. Plug-and-Play (Tumanyan et al., 2023) observes that residual features and self-attention maps fix the layout of a synthesis, and injects them from an inverted source image so that a new text condition changes appearance without moving structure. MasaCtrl (Cao et al., 2023) retains the source's self-attention keys and values to preserve appearance under text prompts. Style transfer inherits these mechanisms: StyleID uses DDIM inversion (Song et al., 2021) and SDEdit (Meng et al., 2022) to supply the content anchor, and AdaIN (Huang & Belongie, 2017) to initialise the start point from the style reference. However, in style transfer the self-attention that carries the reference's appearance also rewrites the content's structure, leading to a loss of detail. To address this, our secondary pass leverages a sparse set of anchors to re-instate it.

3 Preliminaries

Latent Diffusion Model. In our work, we utilize Stable Diffusion (SD, Rombach et al. (2022)), which is a latent diffusion model trained on the low-dimensional latent space. With the pretrained encoder, its network predicts the noise added to the latent and subtracts it at each step. SD is based on a U-Net architecture in which a block for each resolution comprises a residual block, self-attention block (SA), and cross-attention block (CA), sequentially. Among these modules, we focus on SA layer to transfer styles. Given a feature ϕ, the self-attention layer performs as follows:

$$\phi_{\text{out}} = \text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \qquad Q = W_Q(\phi),\ K = W_K(\phi),\ V = W_V(\phi), \tag{1}$$

where d denotes the dimension, and W is a projection layer. Notably, we use no text prompt throughout, so the conditioning input is always empty.

DDIM Inversion. DDIM (Song et al., 2021) improves sampling efficiency and quality through a non-Markovian diffusion process. Within SD, deterministic DDIM sampling denoises a latent

through the network ϵ_θ along a fixed trajectory, so that the output is determined by the start point z_T alone. By expressing this sampling as an ODE and inverting the reverse process, a real image ($t = 0$) can be mapped back to gaussian noise ($t = T$) that regenerates it, yielding the inversion trajectory $[z_t^*]_{t=0}^T$.

4 METHOD

Unlike artistic style transfer with only a single style image I_s, a film $\mathcal{S}$ has a set of F style frames $\mathcal{I}_s = \{I_s^1, \ldots, I_s^F\}$. The task is to modify the style of a content image I_c to that film's look. Then, the stylized image I_{cs} would maintain the content of I_c while its style is transferred from $\mathcal{I}_s$. Accordingly, in the following sections we regard the content, style, and stylized images as their encoded counterparts z_0^c, z_0^f, and z_0^{cs}, and their DDIM-inverted latents as z_T^c and z_T^f.

4.1 THE PRIMARY GRADE

The primary grade carries a consistent look across the reel, the diffusion analogue of a colorist's global grade. We obtain the latent for content and style images with the DDIM Inversion, and then collect the SA features of all frames over the inversion process.

First, we record the query features of content Q_t^c and key/value features of all film frames K_t^f, V_t^f at every timestep $t \in \{0, \ldots, T\}$. After that, we initialize stylized latent noise z_T^{cs} by copying content latent noise z_T^c. Then, at each timestep t we assemble the collected keys and values from F_b frames,

$$K_t = \left[K_t^1; \cdots; K_t^{F_b}\right], \qquad V_t = \left[V_t^1; \cdots; V_t^{F_b}\right], \tag{2}$$

where $[\,\cdot\,]$ denotes concatenation and $K_t, V_t \in \mathbb{R}^{(F_b N) \times d}$ stack N spatial tokens of each frame. During sampling, we transfer the target style to the stylized latent by injecting key K_t and value V_t into the SA layers, replacing the original key K_t^{cs} and value V_t^{cs}. Following (Chung et al., 2024), we blend query of stylized latent Q_t^{cs} and that of content Q_t^c into $\tilde{Q}_t^{cs}$ for the entire reverse process. Then, these style injection at time step t can be expressed as follows:

$$\phi_t^{cs} \;=\; \mathrm{Attn}\Big(\tilde{Q}_t^{cs}, K_t, V_t\Big)\,. \tag{3}$$

We also use the initial style latent z_T^f for the style transfer process, since the tone information is closely related to the channel statistics of the initial latent (Gatys et al., 2016). Concretely, we initialize the starting latent by AdaIN (Huang & Belongie, 2017) against statistics pooled over the set of F_a inverted style latent,

$$z_T^{cs} \;=\; \left(\frac{z_T^c - \mu(z_T^c)}{\sigma(z_T^c)}\right)\sigma_{\mathcal{S}} \;+\; \mu_{\mathcal{S}}, \qquad \mu_{\mathcal{S}},\, \sigma_{\mathcal{S}} = \mu, \sigma\big(\{z_T^f\}_{f=1}^{F_a}\big), \tag{4}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the channel-wise mean and standard deviation, and the set statistics $\mu_{\mathcal{S}}, \sigma_{\mathcal{S}}$ are pooled over frames and spatial positions jointly, where the F_a latents are treated as one population before the per-channel statistics are taken. Based on this, the initial latent z_T^{cs} aligns the channel-wise statistics with style latent from the frames. More details are shown in section 6.3.

4.2 THE SECONDARY PASS

A colorist's secondary correction is governed by a matte that delimits where the correction takes effect. As above, the primary grade applies indiscriminately, including at the very structural detail a viewer relies upon to read the scene, which is what breaks fine objects. Following the colorist's practice, the secondary pass is conditioned on the content's own structure.

We realise this matte as a sparse graph $\mathcal{G}$ of the content's salient structure, which localises where detail must be re-instated (Alg. 1). Line segments are first detected with LSD (Von Gioi et al., 2008) and merged into continuous polylines by grouping on orientation, collinearity, and separation, thereby reconstituting a complete structure. We retain only corners (Shi et al., 1994) that fall within a snap distance of a retained segment, ensuring that every node rests on an edge rather than floating in texture. Finally, we discard connected components whose extent falls below a threshold, since

after merging a short blob is a residue rather than a structure. Notably, $\mathcal{G}$ is constructed once before sampling from the content alone and reused throughout.

After that, we apply this re-instatement only in the self-attention layers of the decoder's later, higher-resolution stages, which govern local texture and fine structure, whereas the earlier stages set the global look and leave it untouched. In each layer, we rasterize $\mathcal{G}$ and reduce it to each injected layer's resolution,

$$M_l[u] \;=\; \max_{p \in \mathcal{C}_l(u)} R[p], \tag{5}$$

where R denotes the binary raster of $\mathcal{G}$ at the content's resolution, $\mathcal{C}_l(u)$ is the set of pixels subtended by cell u of the layer l's grid, and $M_l \in \{0,1\}$ is the induced mask. We adopt a maximum since a one-pixel line occupies only a marginal fraction of a cell. Then, we re-instate the content's own recorded attention output,

$$\begin{aligned}\phi_t^{cs} \;\leftarrow\; & \phi_t^{cs} \odot (1 - w_a M_\ell) \\ & + \phi_t^{c} \odot (w_a M_\ell),\end{aligned} \tag{6}$$

where ϕ_t^{cs} is the injected output of equation 3; ϕ_t^c is the content's own SA output, recorded during the content inversion before any style is injected, which carries the content's true geometry uncorrupted by the transfer; $\odot$ is element-wise multiplication along the channel dimension, and w_a is the degree of blending on the range of $[0,1]$, with a higher w_a drawing the anchored structure closer to the content. More details are shown in App. C.2.

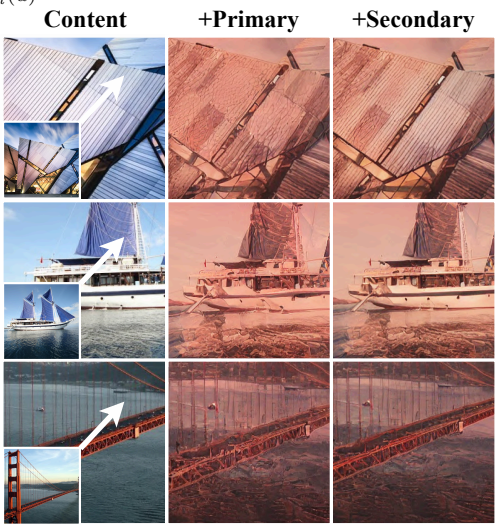

Figure 4: Qualitative comparison of zooming details without vs. with the secondary pass.

4.3 A PROCESS CORRECTION

Alongside the self-attention modified in equation 3, each decoder block contains a cross-attention module. In the unconditional regime we adopt, the prompt is empty, so its key/value are formed from a fixed set of padding embeddings shared across all inputs and timesteps, and the only input-dependent term is the query. Measuring the resulting CA output across contents (Figure 5, left), we find it dominated by a shared, content-independent direction: the mean cosine between per-content outputs stays above 0.95 at every layer and step, and approaches 1.0 in the shallower layers where our operator acts. The output thus decomposes as $\phi_i^x = b + m_{\varepsilon_i}$, a shared near-constant vector b plus a small content-dependent residual m_{ε_i}, so b faithfully summarises the output. This term contributes nothing to the transfer and merely imposes the same offset on every content, much as a fixed contamination is lodged like a hair in the film gate. We therefore delete it with a single-line correction on the CA output,

$$\phi^x \leftarrow (1 - w_x)\, \phi^x, \tag{7}$$

where $w_x \in [0,1]$ is the deletion strength. At $w_x{=}1$ the entire CA output is removed, subtracting essentially the same offset b from every content while leaving the SA that carries the film's appearance untouched. Since b is content-independent, the deletion is expected to act uniformly, which we examine in section 6.3.

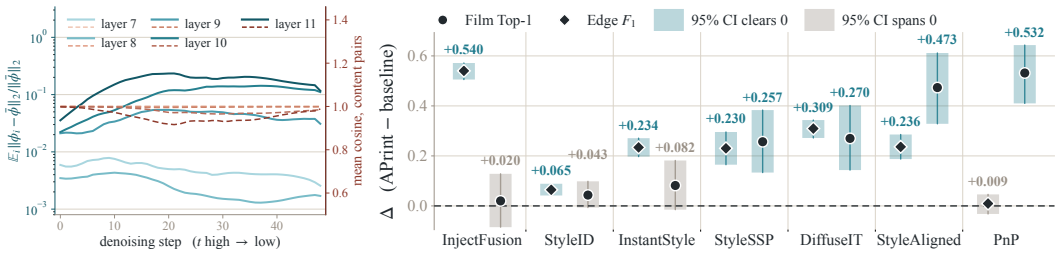

Figure 5: **Left:** the CA output varies little in norm (blue) and stays near-constant in direction (red). **Right:** film-level paired bootstrap of APrint minus each baseline.

Table 1: Quantitative comparison on Reel-30. Best in each column in **bold**, second best underlined. Leads on the ArtFID-family terms reward flawed style transfer, see in section 6.1 and App. B.2.

method	backbone	film top-1 ↑	style sim. ↑	edge F_1 ↑	FID ↓	LPIPS ↓	ArtFID ↓	CFSD ↓
PnP‡	SD1.5	0.075	0.026	0.638	19.12	**0.380**	**27.76**	0.201
StyleAligned	SDXL	0.133	0.048	0.411	**17.92**	0.668	31.55	0.274
InstantStyle	SDXL	0.525	0.211	0.413	20.35	0.594	34.03	0.222
InjectFusion	ADM	0.587	0.218	0.107	19.65	0.888	38.99	0.348
DiffuseIT	ADM	0.337	0.155	0.338	20.66	0.676	36.30	0.278
StyleID	SD1.5	0.563	0.213	0.582	19.38	0.504	30.65	0.183
StyleSSP	SDXL	0.350	0.164	0.417	19.94	0.492	31.25	0.229
APRINT (ours)	SD1.5	**0.607**	**0.222**	**0.647**	20.10	0.458	30.77	**0.169**
APRINT (ours)	SDXL	0.535	0.197	0.552	18.44	0.466	28.50	0.179

‡ Text-conditioned: style reaches it only through the prompt (App. B.1).

5 THE REEL-30 BENCHMARK

Existing style transfer benchmarks define a style as a single artistic image and score appearance with whole-image statistics. Neither fits our task grounded in film post-production. A colorist never grades toward one still: the reference is a reel, a coherent stretch of the film whose look is a property of the set of frames, not of any one of them. And the damage a global grade does is local in area, so a whole-image mean cannot see it, exactly the failure a secondary correction exists to fix. Reel-30 removes both mismatches.

Reel-30 consists of thirty films reported in eight categories (animation to live action, monochrome to saturated), each contributing a pool of curated frames from which the protocol draws F frames as the style set. It employs content images from MSCOCO dataset (Lin et al., 2014) as raw frames in post-grading, which also aligns with the standard single-image protocol (Chung et al., 2024). We hold out the reference samples so that fidelity is never scored against those the method was shown. The full roster, per-film statistics, and category grouping are given in App. A.

Film selection. From a larger pool of candidates, we selected a subset that maximizes the minimum pairwise distance between movies in the embedding space. This criterion guarantees that no two movies in the subset are near-duplicates, and it removes any reliance on subjective impressions during selection. We detail the selection procedure and the measures in App. A.1.

Evaluation metrics. We consider *(i) standard ones*: ArtFID (Wright & Ommer, 2022), FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018), and CFSD (Chung et al., 2024) (see details in App. A.2). *(ii) Film identity*: measuring how well an output keeps the visual character of its source film. Film Top-1 tests this by a linear probe trained on each film's held-out frames that scores how often an output is recognised as its source film (Sanakoyeu et al., 2018), reported against the probe's ceiling (App. A.2); Style similarity measures it as feature-space closeness to the reference frames, via the DINO/CLIP-I protocol of DreamBooth (Ruiz et al., 2023) with DINOv2 features (Oquab et al., 2024). *(iii) Local structure*: measuring whether fine content structure survives the grade, which those metrics in (i)/(ii) all fail to detect (section 6.2). We use edge F_1: an F_1 between edges re-extracted from the output by a fixed independent detector and the content's, under the condition-reconstruction protocol of Li et al. (2024) (see more in section 6.3).

6 EXPERIMENTS

Competitors. We compare with state of the art style transfer methods: PnP (Tumanyan et al., 2023), InjectFusion (Jeong et al., 2024), DiffuseIT (Kwon & Ye, 2023), StyleAligned (Hertz et al., 2024), InstantStyle (Wang et al., 2024), StyleID (Chung et al., 2024), StyleSSP (Xu et al., 2025). We employ their publicly available implementations in recommended configurations.

Settings. We conducted all experiments with a frozen Stable Diffusion backbone and adopt DDIM inversion and sampling with a total of 50 timesteps ($T = 50$). Unless mentioned separately, we set the deletion strength $w_x = 1$, anchor dose $w_a = 0.5$, $F_a = 8$, $F_b = 1$, where the F_b frame is the medoid one of the F_a frame group; the backbone is frozen throughout and no weights are trained.

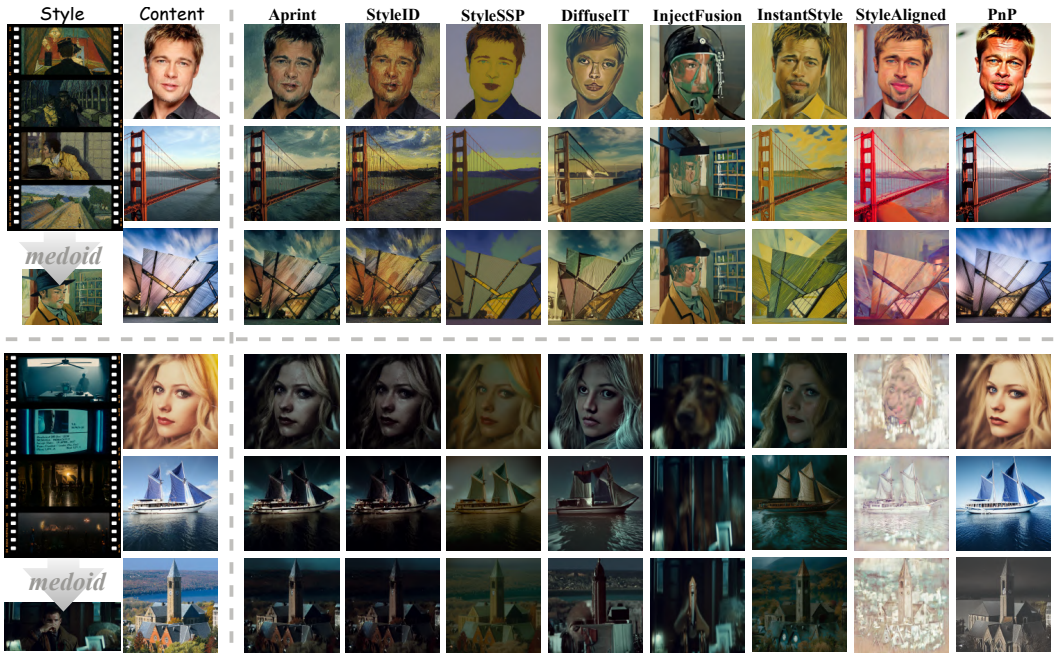

Figure 6: Qualitative comparison with diffusion model baselines.

6.1 Quantitative Comparison

Table 1 reports all comparison experiments on our Reel-30. [2] APrint attains the highest film Top-1 accuracy (0.607) and style similarity (0.222) among all methods, and the best edge F_1 (0.647). On Top-1 accuracy it exceeds our base-model StyleID (0.563) numerically by 7.8%. However, a single metric can always be gamed by a degenerate behaviour, and Reel-30 is designed to expose this by scoring style and structure jointly. On the style side, the ArtFID-based metrics used in prior work fail to rank methods: PnP reaches a film Top-1 of only 0.075 yet spuriously leads on ArtFID because it scarcely alters the input, rewarded by every content-fidelity term. Symmetrically, on the identity side, InjectFusion attains a high film Top-1 by transplanting the style frames almost wholesale, discarding the content entirely. This is exactly what our film-level paired bootstrap makes visible (10,000 resamples over the 30 films; Figure 5, right): APrint's edge F_1 advantage clears zero against every method, so it is the most structure-faithful method by a wide margin over the very baselines it ties on identity (*e.g.*, InjectFusion $\Delta_{\text{film-T}}$=0.02 yet Δ_{edge}=0.54; InstantStyle $\Delta_{\text{film-T}}$=0.08 yet Δ_{edge}=0.23). Reading the two axes together also proves this: among methods that exceed 50% film Top-1, APrint is best on LPIPS (0.458) and CFSD (0.169). We therefore conclude that APrint strikes the most favorable trade-off, acquiring the strongest film characteristics while preserving content structure most faithfully. A complete analysis is given in App. B.

6.2 Evaluation with Human

Film post-production is ultimately a human craft, so we validate APrint against the people it is meant to assist, in collaboration with a professional film academy.

Blind preference. Trained colorists compare APrint against each baseline in a method-blind, order-randomised, forced-choice protocol, the target film shown as its held-out reference frames. For every pair they answer separately: *(i)Which better matches the film's look and is the more usable grade* and *(ii) Which better holds the raw photograph's structure*. APrint is preferred on look-match in 69% and on structure in 83% of comparisons. Treating *(i)-(ii)* as trained-eye ground truth, our film identity metric and local edge readout (section 5) agree with it in $>60\%$ and $>85\%$ of pairs, validating the instrument beyond the authors' authority.

[2] For fairness, single image baselines are reported with their most favorable frame, and we detail the frame selection procedure in App. C.1.

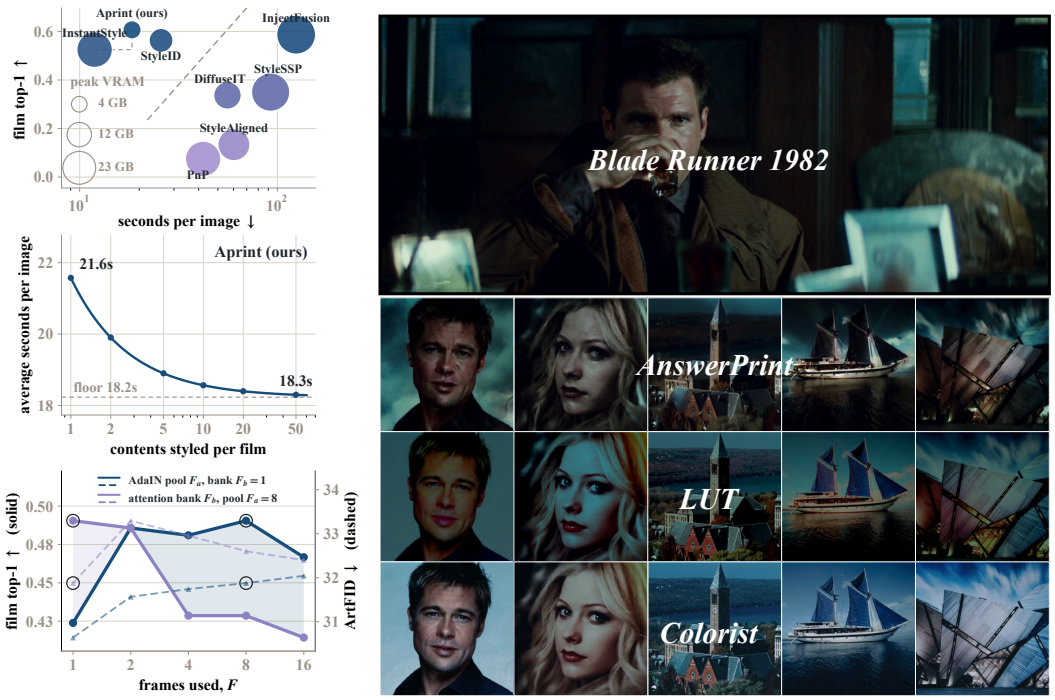

Figure 7: **Left:** Ablation Study on cost analysis and frame selection rule. **Right:** Qualitative comparison between LUTs and our APrint, under the *Blade Runner* reference.

Qualitative comparison. With the blind vote recorded, we unblinded the outputs and asked the colorists to read each method. Figure 6 shows the two regimes they assessed, an animated title (top) and a live-action one (bottom). APrint passed on both, carried the animated title's stylized palette and painterly surface and the live-action title's low-key tonality and grain, while skin tones, facial identity, and object geometry held. The baselines each drew a specific fault. Our base-model StyleID set the palette cleanly but graded to a single frame rather than the reel; StyleSSP posterized the image and crushed the tonal gradation a colorist reads first (App. B.2); InjectFusion let the reference bleed into the plate, importing objects and at times replacing the subject, while DiffuseIT and InstantStyle disturbed the plate itself,*e.g.*, a distorted face, that no secondary could recover; the text-conditioned StyleAligned and PnP they judged unusable, where PnP barely left the source and StyleAligned washed the subject into haze.

Manual grading. We compare APrint against two points from traditional post-production. The first is a fully automatic LUT (Pitié & Kokaram, 2007): it treats the reference and content's colour distributions as RGB Gaussians and applies the closed-form transport map that aligns their full 3×3 covariance. As Figure 7 shows on *Blade Runner* (1982), it aligns the palette convincingly and its teal cast unmistakable, but one affine map per pixel cannot decide where light should sit: saturation rises wherever the content is chromatic and the transform runs out of gamut, so lips turn magenta, hair takes a green rim, and a face is left as bright as the background. The second is a grade built by colorists' hand: the colorists grade a fixed set of shots toward each reference reel under normal working conditions, timed from ungraded plate to an accepted match, secondary corrections included (section 6.3). Drawing on priors about the scene, *e.g.*, the backdrop is a smooth, solid wall, and on their own subjective sense of how the shot should feel, the colorist holds clean and finely graded transitions across the tonal range. This attains finer granularity and a stronger look than any automatic method, but costs $\approx$12 min per shot against APrint's $\approx$20 s, a 36$\times$ gap. Between a fast but rigid LUT and a fine but costly hand grade, APrint balances quality against cost.

6.3 Ablation Study

Component Analysis. To assess each component, we partition edge agreement into two regions—inside the anchor graph and outside it—and report the change in each over the baseline (Table 2). This split separates a genuine structural effect from indiscriminate copying: the former raises agree-

Table 2: Ablation studies. *Secondary* and *Correction* are independent additions to the primary grade, and ANSWERPRINT combines both. We sweep the anchor dose w_a over the sparse graph, with w_a=0.50 as our default. Partition columns give the relative change in edge F_1 inside vs. outside the operator.

setting	film top-1 ↑	edge F_1 ↑	LPIPS ↓	ArtFID ↓	CFSD ↓	ΔF_1 inside	In/Out
Primary (no anchor)	0.618	0.610	0.486	31.80	0.181	–	–
+ *Secondary*: anchor dose w_a (sparse graph)							
$w_a = 0.25$	0.597	0.624	0.474	31.62	0.174	+6.2%	8.5
$w_a = 0.50$ (default)	0.580	0.635	0.465	31.41	0.169	+11.4%	**8.3**
$w_a = 0.75$	0.545	0.645	0.458	31.13	0.164	+15.4%	7.7
+ *Correction*	**0.637**	0.621	0.480	31.10	0.182	+1.7%	0.9
ANSWERPRINT	0.607	**0.647**	**0.458**	**30.77**	**0.169**	+13.3%	4.1

ment only where it acts, the latter lifts every region alike. The gain of secondary pass concentrates where it operates, at a ratio of 8.3 inside to outside, evidence that the recovered structure is anchored, not copied wholesale. This effect is moreover controllable: sweeping the anchor dose w_a trades identity against structure monotonically, film Top-1 falling $0.618 \rightarrow 0.545$ as edge F_1 rises $0.610 \rightarrow 0.645$ (App. C.2), and we adopt w_a=0.5 as the setting that recovers structure while largely preserving the transferred look. Our process correction, a CA deletion, lifts both regions near-uniformly (1.7% inside vs. 1.9% outside), the signature of removing a content-independent offset; it accordingly acts on identity rather than structure, raising film Top-1 $0.618 \rightarrow 0.637$. Combining both, APrint tops every structural metric and reaches a film Top-1 of 0.607, realising the trade-off.

Cost Analysis. In film post-production a single reference is applied to many shots, so the quantity that matters is the marginal cost of one image, measured against a colorist's per-shot minutes (section 6.2). Since the inversion is reused and only sampling runs per style-content pair, this cost falls as the grid grows while peak memory stays constant (Figure 7, medium-left). Under each baseline's reported configuration (Figure 7, top-left), InstantStyle is the cheapest at 11.9s per image and our APrint takes 18.4s but improves film Top-1 by 15.6% relative. The memory gap is far larger: our 3.9GB peak is about one fifth of InstantStyle's (19.1GB) and half of StyleID's (7.6GB), which is the lowest peak memory footprint, running within a 4 GB budget. DiffuseIT and StyleAligned take 55.9s and 60.3s yet score only 0.337 and 0.133, while StyleSSP is the most expensive at 92.5s and 22.0GB. We attribute this efficiency to our architecture: APrint needs no style-specific preprocessing and reuses the attention outputs from content inversion, so its passes add almost no cost.

Where a reel enters? In section 4.1, a set of frames enter in either place: pooled into the AdaIN statistics of the starting latent, or concatenated into the attention bank. Chung et al. (2024) find that attention injection carries texture and local pattern but leaves colour tone to the initial latent's channel statistics. As shown in Figure 7, we sweep one axis at a time on Reel-30 over selected contents and decide where a set may be aggregated: pooling more frames raises film Top-1 from 0.424 to 0.491, whereas widening the bank moves it the other way, to 0.414. This suggests that tone is distributional, so averaging over accumulating frames cancels the palette bias of any single scene and estimates it more accurately. By contrast, texture such as grain or brushstroke is already captured from one or two frames through attention, with no gain from more. We therefore pool F_a frames and bank the single most central one as F_b, with the full frame-selection rule in App. C.1.

7 CONCLUSION

We introduce a new problem *film style transfer* (FST): applying a whole film's look to raw shots. We design AnswerPrint, a training-free FST method. Following real workflow, it performs the primary look grade from the reel, re-instates fine structure at content-defined keypoints the global grade would erase, and lifts a content-independent contamination at every step. We created Reel-30, a benchmark that defines a style with a reel. Assessed both by evaluation metrics and by film industry professionals, our method is shown to be superior and useful in practice.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, AI-assisted coding, and scholarly navigation, such as identifying related work. We did not use generative AI tools for research ideation or to draft any figures and sections of the paper. We reviewed all AI-assisted work, including checking that every LLM-polished passage preserved our intended meaning and verifying AI-assisted code for correctness before using it in our experiments. We take responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our new benchmark is built from publicity and gallery stills of thirty copyrighted motion pictures, obtained through lawful access and used solely for non-commercial academic research. Copyright in the original audiovisual works remains with the respective rights holders. We redistribute no film frames and no video. Our release contains only a manifest – for every still, its filename, dimensions – together with the film and frame selection procedures, the evaluation code, and the preprocessing scripts needed to reproduce the benchmark; readers obtain the stills themselves through lawful channels and can verify by digest that they are running on our exact inputs. The content photographs are standard public style-transfer test images. Code is released under a permissive licence; the manifest and annotations we authored are licensed separately, as we hold no rights in the depicted works.

REPRODUCIBILITY STATEMENT

We make every effort to ensure the reproducibility of our work. The main paper and appendix provide detailed descriptions of the proposed method, experimental settings, implementation details, and evaluation protocols. We will release the complete source code of our method and the necessary configuration files and scripts for reproducing the reported results. We introduce a new benchmark for film-look style transfer in this work. Due to copyright restrictions, the copyrighted film stills cannot be redistributed; instead, we will provide a complete manifest containing the filename and image dimensions of each still, together with the film and frame selection procedures, preprocessing scripts, evaluation code, and benchmark annotations. All released code and benchmark resources will be made publicly available upon publication.

REFERENCES

Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7941–7950, 2021.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8795–8805. IEEE, 2024.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr^2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11316–11326. IEEE, 2022.

Encyclopædia Britannica. Answer print. `https://www.britannica.com/technology/answer-print`, 2026. Accessed: 2026-09-22.

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.

Edward J Giorgianni and Thomas E Madden. *Digital color management: Encoding solutions*. John Wiley & Sons, 2009.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4775–4785. Ieee, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6840–6851, 2020.

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.

Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5139–5149. IEEE, 2024.

Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *International Conference on Learning Representations (ICLR)*, 2023.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision (ECCV)*, pp. 129–147. Springer, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

François Pitié and Anil Kokaram. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *4th European Conference on Visual Media Production (CVMP)*, pp. 1–9, 2007.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

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.

Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. RB-Modulation: Training-free personalization of diffusion models using stochastic optimal control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*, pp. 698–714, 2018.

Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pp. 593–600. IEEE, 1994.

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.

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

Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved ArtGAN for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2018.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930. IEEE, 2023.

Alexis Van Hurkman. *Color correction handbook: professional techniques for video and cinema*. Pearson Education, 2014.

Rafael Grompone Von Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):722–732, 2008.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7643–7655. IEEE, 2023.

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

Ruojun Xu, Weijie Xi, Xiaodi Wang, Yongbo Mao, and Zach Cheng. StyleSSP: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

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

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10146–10156, 2023.

A REEL-30: THE COMPLETE FILM ROSTER

In the era of photochemical film, the spool that carried the developed footage was called a reel. Now, we borrow the name for the short, continuous clips of striking imagery that we draw from films.

Compared with the disparate and hard-to-define style image used by single image protocols, a reel is a far more meaningful reference for real post-production work. For instance, a colorist of the *Dune* might grade the source footage toward the cool, tactile palette of a desert epic for the final shot.

As previously mentioned (section 1), we observe that an individual frame of a reel rarely captures the static stylistic character of the film as a whole. As shown in Figure A1, a single image is confounded by scene, subject, lighting, and similar factors, and when it enters the model as a reference the transferred style drifts away from the true one. This is exactly the failure case in which a colorist or concept artist cannot pin the style down. Our idea is that sampling frames uniformly across a clip and averaging over several of them synthesises what the frames share rather than what is idiosyncratic to any one of them, so that these confounders are largely suppressed or removed. It is much as a colorist finds it easier to pin a style down after working through many examples of the same kind.



Figure A1: The visual comparison between the single-image protocol and Reel-30.

We build **Reel-30**, a benchmark of 30 films spanning eight categories, from animation to live action and from monochrome to saturated. Each film comes with a curated set of frames, from which the

protocol draws F frames as the style set. The benchmark pairs these with content images from MSCOCO (Lin et al., 2014) and follows the standard training-free style-transfer protocol (Chung et al., 2024), while holding out the frames not drawn as references so that no leakage occurs at evaluation time. The full list of films, together with per film statistics and category grouping, is given in Table A1.

Concretely, the 30 films of Reel-30 are selected from 38 candidates by an exact drop-8 diversity optimisation (section A.1, $\tau^\star$=0.611). The categories are a reporting device rather than a selection criterion. For each film we report its category, the mean pairwise compactness of its own reference frames in the CSD embedding (higher values indicate stronger self-consistency and hence a more reliable reference set), its standalone identifiability (the recall of a trained linear probe that distinguishes the film from the others, between 0.86 and 1.00 across candidates), and the number of curated frames available. Since the protocol draws $F \in \{1, 4, 8, 16\}$ style frames from each pool, the three films with fewer than 50 frames are still covered adequately.

Table A1: The thirty films of Reel-30, grouped by reported category. *Comp.* is mean reference-frame compactness; *Ident.* is stand-alone identifiability; *Frames* is the size of the curated frame pool.

#	film	category	comp.	ident.	frames
1	Days of Heaven	arthouse	0.705	1.00	50
2	Lawrence of Arabia	arthouse	0.583	0.94	50
3	Moonrise Kingdom	arthouse	0.630	0.92	50
4	The Fall	arthouse	0.559	0.92	50
5	The Grand Budapest Hotel	arthouse	0.624	0.94	50
6	Suspiria	arthouse	0.581	0.98	50
7	Feeling from Mountain and Water	cn-animation	0.781	1.00	26
8	Havoc in Heaven	cn-animation	0.708	1.00	26
9	Farewell My Concubine	cn-liveaction	0.565	0.94	50
10	Hero	cn-liveaction	0.523	0.96	50
11	In the Mood for Love	cn-liveaction	0.673	0.94	50
12	Raise the Red Lantern	cn-liveaction	0.688	1.00	50
13	Shadow	cn-liveaction	0.764	1.00	50
14	Punch-Drunk Love	crime-drama	0.628	0.98	50
15	The Godfather	crime-drama	0.562	0.94	50
16	The Social Network	crime-drama	0.705	1.00	50
17	AKIRA	jp-animation	0.691	0.90	50
18	Paprika	jp-animation	0.640	0.92	50
19	Citizen Kane	monochrome	0.748	0.98	50
20	Blade Runner (1982)	scifi	0.618	0.96	50
21	Dune	scifi	0.558	0.96	50
22	Pacific Rim	scifi	0.628	1.00	50
23	The Matrix	scifi	0.727	0.98	50
24	Transformers	scifi	0.657	0.93	30
25	Coraline	west-animation	0.553	0.90	50
26	Isle of Dogs	west-animation	0.520	0.94	50
27	Klaus	west-animation	0.733	1.00	50
28	Loving Vincent	west-animation	0.794	1.00	50
29	Spider-Man Into the Spider-Verse	west-animation	0.614	1.00	50
30	The Triplets of Belleville	west-animation	0.738	1.00	50

A.1 SELECTION.

Our goal in selecting the roster is simply to avoid redundancy: no two films should be near-duplicates in style, so that each contributes a distinct look. From 38 candidates we keep 30, dropping the 8 that would otherwise form near-duplicate pairs. Concretely, we lower a similarity ceiling τ until only 30 films remain mutually below it; this yields $\tau^\star$=0.611. However, several equally valid rosters satisfy this bound, and no single film is forced out; among them we take the one whose films are

most self-consistent (highest mean compactness) as the final list in Table A1. This selection uses only pairwise style similarity in the CSD embedding for redundancy removal, and is blind to our method's outputs, so it never tunes the benchmark toward our results.

The 8 dropped films. Joker, Puss in Boots: The Last Wish, Seven (1995), Ghost in the Shell, Spirited Away, Schindler's List, Arcane, and Howl's Moving Castle. As shown in Table A2, each is dropped only to break a redundant pair above $\tau^\star$, never for being weak on its own: all 38 candidates are individually identifiable at 0.86–1.00 recall.

Table A2: The ten redundant film pairs above $\tau^\star$, ranked by centroid cosine similarity in the CSD embedding.

film A	film B	centroid cosine
Howl's Moving Castle	Spirited Away	0.926
AKIRA	Ghost in the Shell	0.792
Citizen Kane	Schindler's List	0.743
Paprika	Spirited Away	0.699
Blade Runner	Seven	0.691
Blade Runner	Joker	0.686
Howl's Moving Castle	Paprika	0.674
Seven	The Social Network	0.669
Arcane	Klaus	0.654
Puss in Boots	Into the Spider-Verse	0.650

Roster observations. First, Japanese animation collapses from five candidate films to two: Howl's↔Spirited (0.926), Paprika↔Spirited (0.699), Howl's↔Paprika (0.674), and AKIRA↔Ghost in the Shell (0.792) are all mutually redundant above the ceiling everything else respects, so keeping three or more would tolerate a redundancy the rest of the set does not. Second, the tightest admissible pair is Farewell My Concubine↔Raise the Red Lantern at 0.611, which sets $\tau^\star$. Finally, three films carry fewer than fifty frames (Feeling from Mountain and Water and Havoc in Heaven at 26, Transformers at 30); the first two constitute the cn-animation category and rank among the most self-consistent in the benchmark on compactness.

A.2 Evaluation Protocol.

Comparability Metrics. ArtFID (Wright & Ommer, 2022) evaluates overall style transfer performances with consideration of both content and style preservation. Specifically, ArtFID is computed as (ArtFID =(1 + LPIPS) · (1 + FID)). LPIPS (Zhang et al., 2018) measures content fidelity between the stylized image and the corresponding content image, and FID (Heusel et al., 2017) assesses the style fidelity between the stylized image and the corresponding style image. Following StyleID (Chung et al., 2024), we additionally report Content Feature Structural Distance (CFSD), a distance that depends only on the spatial correlation between image patches. For an image I, let $\boldsymbol{F} \in \mathbb{R}^{hw\times c}$ be the `conv3` feature map of VGG19; we form the patch-similarity map $\boldsymbol{M} = \boldsymbol{F}\boldsymbol{F}^\top \in \mathbb{R}^{hw\times hw}$ and turn each row into a distribution, $\boldsymbol{S} = [\mathrm{softmax}(\boldsymbol{M}_i)]_{i=1}^{hw}$, where $\boldsymbol{M}_i \in \mathbb{R}^{1\times hw}$ is the i-th patch against all others. CFSD is the KL divergence between the correlation maps of the content ($\boldsymbol{S}^c$) and the stylized image ($\boldsymbol{S}^{cs}$),

$$\mathrm{CFSD} = \frac{1}{hw}\sum_{i=1}^{hw} D_{\mathrm{KL}}\big(\boldsymbol{S}_i^c \,\big\|\, \boldsymbol{S}_i^{cs}\big). \tag{A1}$$

CSD embedding and film-identity probe. All film-identity measures (subset selection, compactness, redundancy, and Film Top-1) use a single frozen CSD encoder (Somepalli et al., 2024). We use the public `CSD-ViT-L` checkpoint: an OpenAI CLIP ViT-L/14 backbone loaded through `open_clip`, with the visual projection disabled so that the 1024-dimensional pooled feature is passed through the model's style head to a 768-dimensional style embedding, L_2-normalised at output. Inputs follow the standard CLIP evaluation transform at 224×224. The encoder is fixed throughout and is never fine-tuned on our data.

Film Top-1 is a multinomial logistic-regression probe (scikit-learn, C=10) over these L_2-normalised embeddings. The probe is trained only on real held-out frames, never on generated images: for each film we exclude the frames used for injection and fit on the remainder. To report the probe's ceiling we take a 75/25 stratified split of the real frames (`random_state`= 0), giving 0.987 against a chance level of 0.033; the probe is then refit on all real frames and applied once to the generated outputs. This ceiling is the reference against which every Film Top-1 in the paper is read.

B Results Analysis

B.1 Text-Conditioned Methods

One baseline in section 6 takes no image input to define style, as a consequence of its architecture: Plug-and-Play injects features under a target prompt, so a film can reach it only through the text channel. To place such a text-to-image (T2I) baseline on Reel-30 without altering how to define style, we automatically generate an identical caption from the same still it would otherwise receive. Captions come from Qwen2.5-VL-3B-Instruct with greedy decoding, which makes the image to sentence mapping deterministic and reproducible. The instruction asks only for the visual style of the film still:

> *"Describe only the visual style of this film still: its colour palette, contrast, lighting, film grain and texture. Do not name any object, person or place. Answer with one short phrase beginning with the word 'style'."*

This yields 30 descriptions of 8 to 36 words each (median 24), for instance rendering *Akira* as "style The film still exhibits a gritty, dark colour palette with high contrast, dramatic lighting, and a grainy texture that evokes a sense of tension and urgency". Nothing here is hand written and every caption shares the same format, so no gain can be attributed to differences between the frame and its sentence. Even so, the T2I results show that style transfer through this channel fails: the descriptions remain an incomplete substitute for the reference image, confirming that text, as a medium for human communication, cannot stand in for the full visual understanding that an image provides.

In addition, two image-conditioned baselines need a prompt, but for them the prompt describes the content and never the film: InstantStyle receives the content caption plus the authors' own quality suffix and negative prompt, and takes the film through the IP-Adapter's CLIP image embedding of the same banked still. StyleAligned needs two prompts because it generates the reference and the target in one batch: the reference row is prompted with a caption of what is *in* the film still (*e.g.*, *"Man drinking at bar"* for *Blade Runner*, *"Three books in an open suitcase"* for *Moonrise Kingdom*), and the target row with the content caption, while the film style reaches the target through shared attention rather than through either sentence. We caption the content photographs with BLIP-2 (20 in total), keeping these identical across all styles, which is the property that matters here.

Notably, our APrint is an reel-to-image model and uses no caption of its own, so its text condition is always empty. This confines our style transfer entirely to the visual domain.

B.2 Misleading results

Metrics won by inaction. The clearest evidence comes from baselines (Table 1). Plug-and-Play takes the best value in the table on LPIPS, while its film top-1 is 0.075 against a chance of 0.033 and its style similarity is 0.026. It is a text-conditioned editor and on this benchmark it edits almost nothing, so every measure of "the content fidelity" is maximised by an output that is nearly the input. Moreover, ArtFID, a comprehensive indicator for previous style-transfer methods, is still won this way: $\text{ArtFID} = (1 + \text{LPIPS})(1 + \text{FID})$, where PnP performs best. The same reward appears among arms that do transfer: StyleAligned transplants the style reference almost wholesale into the generated image, damaging the content so severely that little of the original semantics remains, with a visually poor result. Yet it records the best FID in the table (17.92). Both metrics mislead, and they mislead oppositely, which is why no single read-out suffices.

Animation and Distortion. Some memory heavy style transfer methods that are ostensibly powerful instead introduce a strong sense of distortion, particularly when the style end is a live action film, where they still render an animated look. StyleSSP is one example. It alters the image markedly: the uneven skin texture on a woman's face is smoothed away, the trusses of a bridge merge into a single continuous mass, and a cluttered water surface becomes clean and coherent. These changes are unwarranted: the method hallucinates a smooth, coherent structure that the content never had, erasing real detail rather than transferring style.

B.3 Error analysis

Of the thirty films, four are recognised with low film Top-1 accuracy, and they fail in different ways.



Figure A2: The eight frames Reel-30 drew for Hero, which span four disjoint colour regimes.

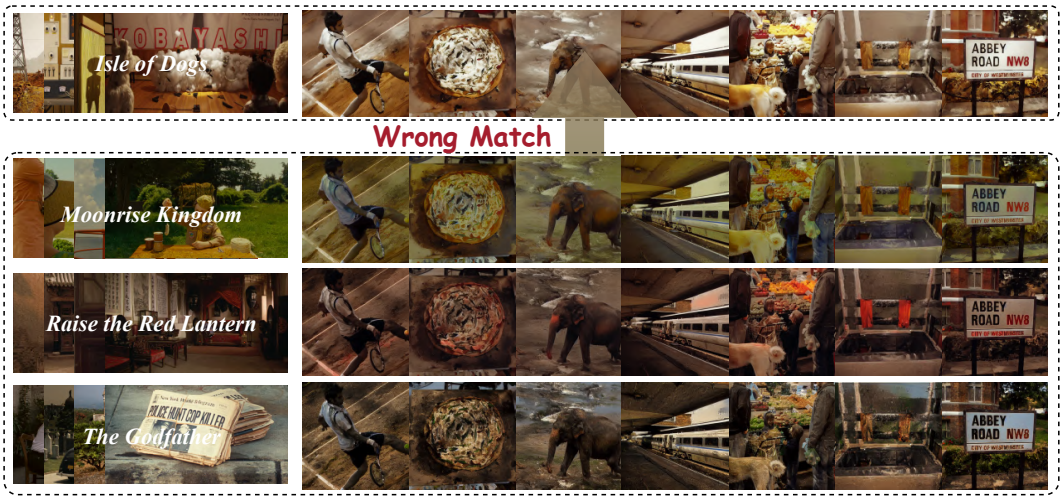

Figure A3: One-way collapse onto a single attractor.

Films without a single style. *Hero* is the first case: its twenty outputs scatter across eight different films with no single winner. This dispersion indicates that the probe fails to form a class for the film at all, rather than confusing it with a nearby one. Figure A2 shows the reason. The eight frames that Reel-30 draws for *Hero* span four essentially disjoint colour regions: a blue and white snowfield, two intensely coloured deep red interiors, a green courtyard, and a lantern lit hall that is almost entirely black. The film narrates through colour as an independent element of its visual design, so its frames do not fluctuate around any single unified style. This is precisely the regime our reel-level assumption does not cover: when a film spans several deliberately distinct colour regimes, no single set of frames averages into a look the film ever actually presents, and pooling across them estimates a style that exists in none of the reels. Hero is thus a limit of the clip-level definition rather than a failure of the transfer itself.

Films a machine confuses but a human separates. The remaining three films show the opposite failure mode: each collapses onto the same neighbouring film, which we call an attractor. *Moonrise Kingdom*, *The Godfather*, and *Raise the Red Lantern* each assign more than half of their transferred outputs to *Isle of Dogs*. This collapse does not reflect genuine visual similarity: a human classifying by style would never group Wes Anderson, Zhang Yimou, and Francis Ford Coppola into one. The three films remain plainly distinct to the eye, yet the probe folds them onto a single attractor. Figure A3 gives the full picture of this misclassification. The outputs are clearly separable to the eye through fine detail, yet the CSD probe loses this discrimination at a film Top-1 accuracy of 0.607. At the level of film category, however, accuracy reaches 0.920: even when it cannot identify the individual film, the system almost always recovers the coarse category it belongs to. This points to our next step, matching film style to the visual scene in time axes so as to achieve fine grained film style transfer.

B.4 VISUALIZATION.

We provide additional qualitative results on Reel-30. Figure 2 shows APrint's style/content pairs, and figure F (attached at the end) gives a larger gallery. Please zoom in to view more details.

Table A3: Frame sensitivity on Reel-30: StyleID performance across the eight frames.

Style reference	film top-1 ↑	category ↑	style sim. ↑	edge F_1 ↑	ArtFID ↓
rank 0 *(medoid)* ‡	0.563	0.905	0.213	0.582	30.65
rank 1	0.530	0.878	0.196	0.588	30.62
rank 2	0.447	0.845	0.200	0.576	32.24
rank 3	0.435	0.882	0.189	0.542	30.90
rank 4	0.443	0.892	0.199	0.593	30.93
rank 5	0.442	0.882	0.204	0.590	30.24
rank 6	0.403	0.848	0.168	0.593	29.60
rank 7 *(least central)*	0.387	0.857	0.159	0.438	29.39
mean ± sd over the eight	0.456 ± 0.056	0.874 ± 0.020	0.191 ± 0.017	0.563 ± 0.050	30.57 ± 0.83
APRINT (ours, a frame *set*)	**0.607**	**0.920**	**0.222**	**0.647**	30.77

‡ The frame used as F_b bank in our main experiment (Table 1)

C HYPERPARAMETER

C.1 FRAME SELECTION

Rule. Our experiments (section 6.3) show that the two routes for obtaining film look from the reel behave very differently: feeding more frames into the initial statistics improves film recognisability and saturates at around eight frames, whereas enlarging the attention bank trades recognisability for structure. A frame set helps because it estimates a central statistic, not because it offers more fragments to match against, so the bank gains nothing from a second frame and is in fact hurt by admitting one. We therefore merge all supplied frames for the initial statistics but keep a single one for the bank, the frame closest to the centre of the supplied set:

$$f^\star = \arg\min_{f \in \mathcal{S}} \left\| \hat{\boldsymbol{d}}^{(f)} - \bar{\hat{\boldsymbol{d}}} \right\|_2, \tag{A2}$$

where $\hat{\boldsymbol{d}}^{(f)}$ is the normalised colour descriptor of frame f over the supplied set, and $\bar{\hat{\boldsymbol{d}}}$ is the set mean. The centre $f^\star$ is taken over the F_a frames, so the rule never needs access to the full film.

Frame Sensitivity. For a fair comparison, every single-image baseline is given the same reference still: for each film we deterministically draw F frames and select the most colour-central one (Eq. A2) as the shared reference, so no baseline gains from a luckier frame.

Then we measure how strongly a single-image baseline depends on the chosen still. Fixing the draw and ordering the eight frames by centrality, we run our base-model, StyleID, once per rank on Reel-30 and score it. As Table A3 shows, film Top-1 falls from 0.563 at the most central frame to 0.387 at the most distant one, with a mean of 0.456 and a standard deviation of 0.056. This spread corroborates our argument regarding "reel-level" characteristics: even when all frames are drawn from the same film, transferred results vary significantly depending on the choice of still, because a single frame conflates the film's look with the idiosyncrasies of its own scene, lighting, and framing (cf. figure A1). Given that, we take the most central frame in the main experiments (Table 1) in favour of the baseline: the reported figure is higher than a truly random choice of frame would give.

C.2 ANCHOR DOSE w_a

As shown in Table A4, the anchor dose w_a trades identity against local structure monotonically. Going from no anchoring through w_a=0.25 to w_a=0.75, film Top-1 falls $0.618 \to 0.597 \to 0.580 \to 0.545$ while edge F_1 rises $0.610 \to 0.624 \to 0.635 \to 0.645$. We adopt w_a=0.5 as the reported setting, since the structural gain is exactly what the anchors are there to provide.

Meanwhile, the straight member graph is almost empty on organic content such as horses or portraits, where there are few long lines and sharp corners to anchor to, so we add a curved contour layer to cover these objects. As shown in Table A4, it does reach the intended objects, but at the cost of film identity. We finally keep the straight graph at w_a=0.5 as the default for its lower cost, and record the curved layer as an ablation here.

Table A4: Ablation on anchor dose. w_a blends the straight anchor graph back at the anchor tokens; cw_a the same graph with a curved contour layer added. Best in each column in **bold**.

anchor	film T1 ↑	sty. ↑	edge F_1 ↑	ArtFID ↓	CFSD ↓
No Anchor	**0.618**	**0.226**	0.610	31.80	0.181
$w_a = 0.25$	0.597	0.222	0.624	31.62	0.174
$w_a = 0.50$	0.580	0.218	0.635	31.41	0.169
$w_a = 0.75$	0.545	0.212	0.645	31.13	0.164
$cw_a = 0.25$	0.553	0.209	0.644	31.32	0.161
$cw_a = 0.50$	0.452	0.186	0.673	31.07	0.145
$cw_a = 0.75$	0.377	0.164	**0.695**	**30.74**	**0.135**

Table A5: The single-image protocol. Rows are cumulative: each adds one component to the row above, so the last row is the full APrint. *inside* is edge F_1 restricted to the anchor footprint. Best in each column in **bold**.

arm	edge F_1 ↑	CFSD ↓	LPIPS ↓	inside ↑	ArtFID ↓
StyleID	0.641	0.230	0.505	0.733	29.13
Primary	0.640	0.232	0.507	0.734	29.18
+ *Correction*	0.649	0.231	0.498	0.737	**29.01**
+ *Secondary*	**0.669**	**0.213**	**0.471**	**0.796**	30.15

Algorithm 1: The construction of the anchor graph. Reported configuration: N_c=80, $\ell_{\min}$=64, N_ℓ=20, ρ=6, $e_{\min}$=48, w_a=0.5; MERGE tolerances α_{tol}=10°, γ=16, $\delta_\perp$=4.

input : content image I; layers $\mathcal{L}$; dose w_a
output: anchored output ϕ^{cs} at every ℓ

$g \leftarrow \text{grey}(I)$
$S \leftarrow \texttt{Merge}\big(\texttt{LSD}(g);\ \alpha_{\text{tol}}, \gamma, \delta_\perp\big)$
$S \leftarrow$ the N_ℓ longest of $\{s : \|s\| \geq \ell_{\min}\}$
$V \leftarrow \{p \in \texttt{Corners}(g; N_c) : \min_s \text{dist}(p, s) \leq \rho\}$
$\mathcal{G} \leftarrow S \cup V$
$\mathcal{G} \leftarrow \mathcal{G} \setminus \{k \in \texttt{Components}(\mathcal{G}) : \text{extent}(k) < e_{\min}\}$
$R \leftarrow \texttt{Raster}(\mathcal{G})$
foreach $\ell \in \mathcal{L}$ **do**
  $M_\ell \leftarrow$ reduce R onto ℓ's grid by equation 5
  $\phi^{cs} \leftarrow$ blend ϕ^{cs} with ϕ^c at M_ℓ by equation 6
end

D SINGLE IMAGE PROTOCOL

To validate the performance of APrint, we also test it under the standard single image protocol used in prior work (Deng et al., 2022; Chung et al., 2024; Xu et al., 2025). The protocol uses the same content photographs as Reel-30, together with forty style paintings from WikiArt (Tan et al., 2018), with every input centre cropped to 512×512 pixels, producing 800 stylised images in total.

As Table A5 shows, primary grade matches our base-model StyleID implementation on all axes, confirming that our pipeline reduces exactly to StyleID when its operations are disabled. The process correction moves the perceptual and edge terms while leaving CFSD essentially unchanged, consistent with its role of suppressing spurious injection. Secondary pass adjusts all three metrics and by a larger margin, reaching edge F_1 0.669, CFSD 0.213, and LPIPS 0.471, with F_1 score inside the anchor graph rising from 0.734 to 0.796. However, the distributional side worsens (FID from 18.35 to 19.50, which in turn raises ArtFID), because pulling structure back toward the content necessarily moves the output distribution away from the referenced set of paintings. On a painting styled benchmark this is a genuine loss, whereas on Reel-30, where the references are film stills and structure is part of the visual effect rather than opposed to it, the same operator improves ArtFID instead.

E ANCHOR GRAPH INJECTION

The anchor graph $\mathcal{G}$, constructed by Alg. 1, can be rendered as a line drawing, so it is natural to ask whether that drawing can be *injected* rather than only used to locate anchors. We tried several routes: blending its query at every step or on an early-schedule subset, substituting it as the source for decoder-feature injection, forming a structural cross-attention, and appending its keys and values to the bank. All were negative. The content's own query features are already the best available description of its geometry, and a line drawing is only a lossy compression of them. The graph therefore survives as the definition of *where* to anchor and as a human-legible representation, not as an injected signal.

Coverage. At the default setting of section 4.2, the graph anchors on average 9.98% of tokens at the finer decoder grid and 16.0% at the coarser one, varying by image because it anchors long structure, which a building offers far more of than an animal. The operator is thus a sparse intervention on both grids, making the inside/outside split of section 6.3 decidable.

F LIMITATIONS

Our approach has several limitations that also mark directions for future work.

- We keep the diffusion backbone frozen and add no auxiliary weights, which makes the method lightweight and training-free but caps its ceiling at the generator's own capacity: content the backbone renders poorly, such as fine high-frequency texture, high resolutions, or domains under-represented in its prior, cannot be recovered by any of our passes. A stronger or fine-tuned generator would likely lift this ceiling, which our training-free design deliberately leaves unexplored.
- Our reel-level formulation assumes a film shares a consistent tone and texture across frames. This holds for most stylistically driven films, but not universally: some titles narrate through deliberately disjoint colour regimes (figure A2), or are driven by story rather than a distinctive visual signature, *e.g.*, popcorn movie, so no single look exists to transfer. On such inputs the task itself is under-defined rather than the method failing, marking the boundary of where FST applies.
- **FST is a still-image problem.** By design we recover only the static tone and texture a still-image method can reach, and do not target temporal consistency. Higher-level cinematic language, including audiovisual rhythm, camera motion, and narrative structure, lies beyond this scope, so a considerable gap remains before the method is ready for film post-production. Bridging it demands a genuine understanding of the cinematic scene, which is the focus of our ongoing work.

"It's easier to make a colour look good than it is for colour to service a story."

— Roger Deakins


1133 1132 1131 1130 1129 1128 1127 1126 1125 1124 1123 1122 1121 1120 1119 1118 1117 1116 1115 1114 1113 1112 1111 1110 1109 1108 1107 1106 1105 1104 1103 1102 1101 1100 1099 1098 1097 1096 1095 1094 1093 1092 1091 1090 1089 1088 1087 1086 1085 1084 1083 1082 1081 1080
Feeling from Mountain and Water
Havoc in Heaven
The Triplets of Belleville
Loving Vincent
Paprika
Punch-Drunk Love
The Grand Budapest Hotel
Days of Heaven
Pacific Rim
ABBEY ROAD NW8