PeCA: Palette Context Assisted Inference for Test-Time Paint-Bucket Colourisation on Animation Videos

Dongheng Lin and Jianbo Jiao

The MIx Group, University of Birmingham, Birmingham, United Kingdom
dxl594@student.bham.ac.uk, j.jiao@bham.ac.uk
Project page: https://rathgrith.github.io/PeCA/

Abstract. In animation production, paint-bucket colourisation for hand-drawn animation is a labour-intensive procedure that assigns each enclosed region in line sketches a colour from reference design sheets. Recent automatic paint-bucket colourisation pipelines mirror this workflow via region correspondence, but correspondences can be brittle when regions are ambiguous fragments without proper context. In this paper, we propose Palette Context Assisted (PeCA), a new training-free, plug-and-play framework for animation video colourisation that aims to close this gap at test-time via reasoning over spatial and temporal contexts. Extensive experiments on existing benchmarks and a newly introduced long-video test case show consistent performance boosts.

Keywords: Video Colourisation · Animation · Correspondence

"A colour shines in its surroundings."

Ludwig Wittgenstein

1 Introduction

Hand-drawn animation colourisation is not an unconstrained image synthesis problem. In production, region colours must strictly follow a discrete celluloid palette defined by design sheets, and colours must be correctly assigned despite deformation, occlusion, and changes in the layout of line-enclosed regions [30]. Consequently, transforming line sketches into correctly coloured animation remains a major bottleneck [12]. A particularly labour-intensive stage is the paint-bucket colourisation, where artists meticulously assign colours to a massive number of enclosed regions [30]. This step is repetitive yet unforgiving: even minor colourisation mistakes or boundary leaks can break production-level quality and lead to costly correction. Although automatic colourisation has improved markedly with the advancement of generative computer vision techniques [12,50], fully reliable automated paint-bucket colourisation remains a challenge.

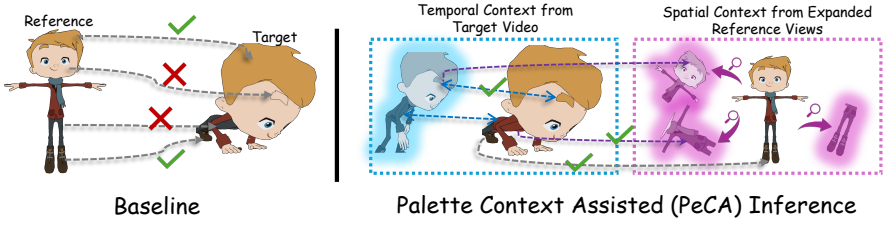


Fig. 1: Conceptual overview of PeCA. Compared to a baseline that colours regions using isolated correspondences, PeCA leverages temporal and spatial context to pick supports for a more reliable colourisation.

Existing automated colourisation approaches can be broadly split into pixel-based generative models and segment-based pipelines. Pixel-generative methods [5, 13, 22, 40, 44, 56, 63, 67], including recent DiT-based generative models [19,58,62,68], can produce seemingly pleasing renders, but often violate production constraints: colours bleed across ink boundaries and discrete colour restriction is weak or absent [50]. Segment-based methods [4,7–9,24,29,37] instead mirror the paint-bucket workflow by treating colourisation as assigning a colour to enclosed regions in the target frame from those in reference frames (*i.e.* coloured key-frames or design sheets) by regional correspondences. This formulation preserves the exact colour and region constraints by construction. The remaining difficulty lies in finding robust correspondences between references and target frame regions that may appear to be completely different.

Most segment-based pipelines solve paint-bucket colourisation by region correspondences, which parse regional similarities to colour label estimations either by top matches or a weighted combination of them [7, 9, 10, 24, 29]. Although some methods introduce structural or temporal constraints, the per-region assignment is still largely driven by local correspondence scores. In practice, failures are often not due to a complete absence of correct matches, but arise when correspondences are ambiguous or noisy due to view/pose variations in animation videos [23, 29]. For example, thin fragments under occlusion can be visually ambiguous on their own, and multiple reference regions may look similarly plausible. This ambiguity persists and becomes a bottleneck. As a result, despite being trained on correspondence-based colourisation, such a direct segment matching & colour propagation pipeline still struggles to deal with spurious correspondences under reference-target appearance gaps [23, 29].

A more robust view is that region identity in animation is rarely resolved from a single isolated match. Humans rely on context when understanding colours in art [55]: both from similar views in reference images, and the temporal continuity in target videos themselves. This suggests a complementary direction to training a stronger backbone: when a direct reference-to-target match is uncertain, spatially similar reference views and temporally neighbouring frames can provide indirect context support that refines ambiguous colour propagation.

To this end, we propose a Palette Context Assisted (PeCA) inference framework that improves segment matching colourisation with test-time context, as shown in Fig. 1. PeCA is training-free at inference; it fits well to either trained colourisation models or frozen foundation backbones. It first constructs a target-conditioned support reference bank, so that reference views are expanded to have better coverage of the current target shot, reducing the visual gap with spatially-close context (Sec. 3.3). We then resolve noisy top correspondences with a soft voting scheme (Sec. 3.4) to reach a consensus that blocks spurious correspondences. Finally, we refine per-region assignments across time, leveraging the continuity between adjacent frames while avoiding unreliable transfers via a gated mechanism (Sec. 3.5). Together, PeCA turns noisy, isolated segment colourisation into context-aware colour assignments without task-specific training. We summarise our contributions as follows:

- We propose a plug-and-play Palette Context Assisted (PeCA) inference framework for paint-bucket colourisation, leveraging context at test-time.
- PeCA improves the default inference that struggles with region ambiguity, by aggregating contexts from both spatially-supportive reference views and temporal continuity in animation videos.
- Extensive experimental analysis on existing benchmarks and a newly introduced long-shot benchmark show consistent gains on both task-trained and frozen foundation backbones, with particularly larger improvements in training-free settings.

2 Related Works

2.1 Automated Paint-bucket Colourisation in Animation

Production paint-bucket colourisation assigns each segmented region a colour from a fixed palette, thus fundamentally relies on region correspondence [30]. Early segment-based methods modelled this problem as geometric or graph-based matching between regions [24, 37]. These approaches typically assume moderate motion and rely on handcrafted similarity measures or motion cues, which fail to handle large appearance gaps and longer videos in production [10]. BasicPBC [9] departs from this simple region matching formulation by explicitly modelling topology with inclusion and subset matching to handle split/merge events, and designed propagation strategies tailored to these cases. Feng *et al*. [10] further extended this direction with a unified pipeline that augments adjacent-frame matching with temporal-structural constraints and additional refinement modules. Compared with earlier methods, they incorporate more structured matching rules and model-specific post-processing to refine colour propagation. But still, these works primarily operate under a temporal-local assumption, where the target frame and reference frames are in the same video.

Key-frame colourisation task relaxes this assumption and considers arbitrary reference–target pairs, such as design sheets and distant shots. BasicPBC-Ref [8] adapts the segment-matching framework to this setting by incorporating stronger

semantic features to bridge pose and layout gaps. DACoN [29] shows that region matching with powerful foundation model descriptors can already serve as a strong baseline for key-frame colourisation, and further improves performance through additional training and model components. Despite such progress, these prior works still report failure modes on extreme views/poses that are visually different from reference shots, and the gains from introducing more reference shots become more marginal as the number of references goes up [29]. This saturation revealed a universal bottleneck identified in many tasks, when models fail to *find direct correspondences under huge visual gaps* [14].

As a response to this bottleneck, our PeCA builds on the simplest segment-matching formulation. Rather than introducing a stronger model for direct correspondence, we focus on improving robustness from a model-agnostic perspective. By keeping the underlying model unchanged, our approach remains compatible with both trained segment matching models and frozen foundation backbones, and isolates the effect of PeCA reasoning from model-specific design choices.

2.2 Visual Correspondence by Foundation Models at Test-time

Large pretrained models provide transferable representations that enable training-free or low-supervision correspondence and region reasoning [16, 31, 35, 38]. A common practice is to pool dense features within regions to build local descriptors, then perform similarity-based retrieval for region matching and propagation [15,42,45]. Beyond such representation reuse, another line of work [18,26,34, 53, 60, 64, 65] improves the performance of various downstream tasks (including generic video colourisation) via test-time adaptation, updating model parameters or refinement at inference steps.

In production-oriented animation paint-bucket colourisation, models must generalise to binarised line sketches with sparse appearance cues [9, 30]. As a result, even strong pretrained region descriptors [31, 35, 38] can be brittle when correspondence is solved purely by similarity retrieval without task-specific training [29]. We thus take a test-time perspective and aim to make inference more reliable with minimal assumptions on the backbone. Instead of treating each region match as an isolated decision, we organise region-level evidence into a palette-space belief and refine it using spatial support from references and temporal support from the target sequence.

3 Methodology

3.1 Problem Formulation

We study production-oriented paint-bucket colourisation for hand-drawn animation, where each enclosed region should be assigned a colour from a discrete palette in reference regions. Given a target video clip of T sketch frames $\{I_t\}_{t=1}^{T}$, each frame is partitioned into closed regions $\mathcal{S}_t = \{s_{t,i}\}_{i=1}^{N_t}$. A reference set $\mathcal{R}$ is provided as $\mathcal{R} = \{(I^{(r)}, \mathcal{S}^{(r)}, Y^{(r)})\}_{r=1}^{R}$, where $Y^{(r)}$ assigns a ground truth

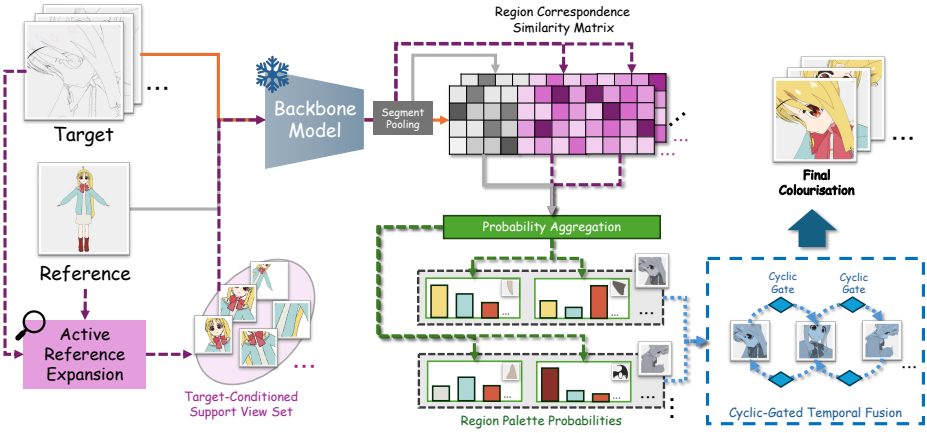


Fig. 2: The proposed PeCA framework overview. Active Reference Expansion (Sec. 3.3) builds a target-conditioned reference support set. **Probability Aggregation** (Sec. 3.4) aggregates noisy matches into per-region palette colour probabilities via soft voting. **Cyclic-gated Temporal Fusion** (Sec. 3.5) fuses colour probabilities across adjacent frames through cycle-consistent temporal links, altogether improving colourisation with **spatial**, **probabilistic**, and **temporal** context.

colour label (from a finite palette $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$) to each reference region in $\mathcal{S}^{(r)}$ (segmented from line-sketches $I^{(r)}$ by flood fill [48]). Our goal is to assign each target region $s_{t,i}$ a colour label $\hat{y}_{t,i} \in \mathcal{C}$ correctly throughout the video.

Following segment-matching pipelines [29, 42], we compute a descriptor for each region by average pooling dense features inside its mask in Eq. (1), and define cosine similarity $S_{(t,i),(r,j)} = \langle \bar{\mathbf{f}}_{t,i}, \bar{\mathbf{f}}_j^{(r)} \rangle$. A retrieval-style baseline in previous SOTA [29] copies the colour label from the top match as in Eq. (2):

$$\mathbf{f}_{t,i} = \mathrm{AvgPool}\big(\{\phi(I_t)[p] \mid p \in s_{t,i}\}\big), \qquad \mathbf{f}_j^{(r)} = \mathrm{AvgPool}\big(\{\phi(I^{(r)})[p] \mid p \in s_j^{(r)}\}\big), \quad (1)$$

$$(r^*, j^*) = \arg\max_{r,j} S_{(t,i),(r,j)}, \qquad \hat{y}_{t,i} = y_{j^*}^{(r^*)}. \quad (2)$$

As discussed in Sec. 1 and Sec. 2, direct correspondence probability $S_{(t,i),(r,j)}$ can be ambiguous for colour propagation. This motivates us to propose the Palette Context Assisted (PeCA) test-time reasoning framework that exploits spatial and temporal context of correspondences and colours during inference.

3.2 Palette Context Assisted (PeCA) Framework Overview

We propose PeCA, a training-free and plug-and-play inference framework that improves region matching-based colourisation by constructing and exploiting context at test time (Fig. 2). PeCA first builds a target-conditioned support bank by expanding the given reference shots to a limited support set that maximises spatial coverage to the target video (Sec. 3.3). Then, from the multi-source

correspondences, a soft top-k voting converts multiple plausible matches into per-region palette-colour consensus probabilities (Sec. 3.4). Finally, we refine these probabilities along reliable correspondence links between adjacent frames serving as temporal context (Sec. 3.5).

3.3 Spatially-Supportive Reference Views Selection

Prior works [8, 29] have shown that increasing the number and diversity of reference shots brings consistent gains for segment-matching-based colourisation, largely because it improves the chance that a target region finds a reference region from a similar layout. In real productions, however, we are often given limited reference RGB images due to the labour-intensive nature of colourisation. In this regard, a natural next step is to expand this limited reference bank with test-time augmentation [43], generating new views to reference shots that possibly provide easier "shortcut" matchings for colour propagation.

However, naive test-time augmentation is unlikely to scale efficiently at inference time under this setting [46]. In basic inference pipelines [7, 29], region assignment is resolved from a similarity matrix over all reference-region candidates. Naively augmenting views, which increases the candidate set indiscriminately, may provide useful evidence. But the additional views may also increase exposure to high-similarity distractors [36]. This makes top matches less stable for ambiguous regions, resulting in a limited performance gain (see Tab. 5). This also aligns with a trend in previous works where further adding more references (>5) shows marginal gains [29].

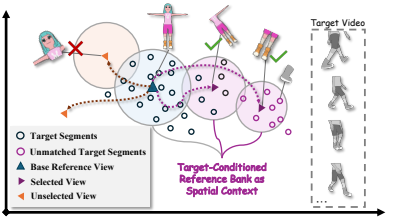


Fig. 3: Active Reference Expansion. Regions are encoded as features (○). There are target segments that hardly match the reference (○). Selecting Target-conditioned views (▶), can maximise coverage to targets, while unselected views ◀ provide limited support to target video.

We therefore designed active reference expansion (Fig. 3) to make additional reference views both budgeted and target-aware, forming a spatial context that best covers the target regions. Starting from the original reference views $\mathcal{V}_0$, we first generate an augmented candidate pool $\mathcal{V}_{\text{aug}}$ by applying geometric transformations (flips, rotations, affine transforms; details in Supp. B.3). Rather than keeping all candidates, we select only B views that best support the target video in the feature space, so the bank is strengthened without indiscriminately enlarging the region candidate set. Specifically, for each candidate view $v \in \mathcal{V}_{\text{aug}}$ with region descriptors $\{\mathbf{f}_j^{(v)}\}_{j=1}^{N_v}$, we measure its support to a target region (t, i) by the similarity of its best-matching region:

$$\text{score}_v(t,i) \;=\; \max_{1\le j\le N_v} \left\langle \bar{\mathbf{f}}_{t,i},\, \bar{\mathbf{f}}_j^{(v)} \right\rangle. \tag{3}$$

We expect the expansion to keep a subset $V \subseteq \mathcal{V}_{\text{aug}}$, such that the resulting support for (t,i) is $\max_{v\in V} \text{score}_v(t,i)$, *i.e.* the best support offered by selected views. We therefore choose B views by maximising a facility-location objective over target regions in uniformly sampled frames $\mathcal{T}_s \subset \{1,\dots,T\}$:

$$F(V) \;=\; \sum_{t\in\mathcal{T}_s}\sum_{i=1}^{N_t} \max_{v\in V} \text{score}_v(t,i), \qquad \max_{V\subseteq\mathcal{V}_{\text{aug}}} F(V) \text{ s.t. } |V| = B. \tag{4}$$

In practice, we form an augmented candidate pool of size $|\mathcal{V}_{\text{aug}}| = mB$ and select B views from it. This selection is performed once per target video, with overhead depending only on mB (see Supp. B.3 and Supp. B.4 for details). Since $F(V)$ is monotone submodular, a greedy algorithm provides a standard approximation guarantee [17]. The resulting support bank $\mathcal{V}_0 \cup V$ improves target-shot coverage under a fixed budget by actively prioritising views that best support the current video. Importantly, this selection step controls the exposure to distractor regions, yielding a more relevant candidate set for subsequent matching. On the other hand, with more reference shots, naturally, for each target region, it introduces additional correspondence candidates from different views. This motivates our next step, which votes multiple high-confidence correspondences to palette colour probabilities (Sec. 3.4).

3.4 Correspondence Candidate Voting for Colour Palette

When more than one reference shot is available, a target region is naturally supported by multiple high-confidence candidates from different sources. To better utilise such probabilistic context, instead of committing to top-1 retrieved region [29] or full linear combinations [7], we wish to resolve the correspondences to colour estimations by *soft-voting* over the top-k correspondence hypotheses. This produces per-region colour consensus probability over the finite palette from the relevant correspondence probabilistic context.

Specifically, for each target region (t,i), let $\mathcal{N}_k(t,i)$ denote the top-k candidate reference regions under similarity $S_{(t,i),(r,j)}$. We convert similarities into normalised weights with temperature $\tau \in (0,1]$:

$$p_{t,i}(r,j) = \frac{\exp\left(S_{(t,i),(r,j)}/\tau\right)}{\sum_{(r',j')\in\mathcal{N}_k(t,i)} \exp\left(S_{(t,i),(r',j')}/\tau\right)}, \qquad (r,j)\in\mathcal{N}_k(t,i). \tag{5}$$

We then vote these weighted matches into the palette colour space:

$$P_{t,i}(c) = \sum_{(r,j)\in\mathcal{N}_k(t,i)} p_{t,i}(r,j)\cdot \mathbb{1}\left[y_j^{(r)} = c\right], \qquad c\in\mathcal{C}. \tag{6}$$

Here, $P_{t,i} \in \Delta^{|\mathcal{C}|}$ summarises the correspondence evidence as colour-level probabilities, where $\Delta^{|\mathcal{C}|}$ denotes the probability simplex over $|\mathcal{C}|$ colour entries. The temperature τ controls the sharpness of this soft vote. When $k{=}1$ and $\tau\to 0$,

it reduces to hard copying as in baseline, while moderately larger k or τ pool evidence across matches and reduce sensitivity to spurious correspondence.

Serving as an interface from region correspondence to per-region colourisation results, the aggregation has two practical benefits for spatial/temporal context: First, restricting aggregation to top-k candidates avoids the dilution effect of naive global mixing when the candidate pool gets larger after active expansion in Sec. 3.3, leading to robust colourisations from soft voting by multiple candidates. Second, colour probabilities live in a fixed label simplex shared by all frames, making them directly comparable and therefore suitable for supporting each other as temporal contexts in Sec. 3.5, unlike matching probabilities that may change across regions with the same colour identity but in different frame pairs.

3.5 Temporal Cycle-Consistency for Colour Refinement

Now each frame has per-region palette-colour probabilities constructed from a strengthened reference support. In addition, animation videos also provide a temporal cue, in which adjacent frames within the same video look more similar due to temporal continuity of videos [6, 50, 54]. Motivated by this, if a direct match between reference and target failed, an easier transitive colour-propagation shortcut can be parsed from the adjacent frames' correspondences (a temporal context) to refine the colourisation.

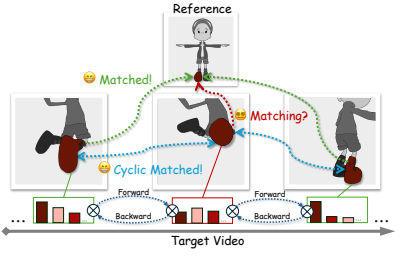


Fig. 4: Cyclic-Gated Temporal Fusion utilises temporal context by fusing per-region colour probabilities only along cycle-consistent matches between adjacent frames, refining colours while avoiding unreliable temporal fusions.

However, region matching is not always reliable under changes between adjacent frames. If we indiscriminately fuse information across time, a single spurious match can be propagated to other frames and amplified. We therefore refine colour probabilities only along cycle-consistent temporal links, so that temporal context is used when it is reliable and ignored otherwise, as shown in Fig. 4.

Specifically, between adjacent frames, we compute adjacent-frame region similarity $A_t[i,j] = \langle \bar{\mathbf{f}}_{t,i}, \bar{\mathbf{f}}_{t-1,j} \rangle$. We define the forward nearest-neighbour match from frame t to $t-1$ as $\pi_t(i) = \arg\max_j A_t[i,j], \quad i \in \{1,\dots,N_t\}$, and symmetrically, the backward match from frame $t-1$ to t as $\rho_t(j) = \arg\max_i A_t[i,j], \quad j \in \{1,\dots,N_{t-1}\}$. We treat a temporal link as cyclic stable only if it is bidirectional:

$$(t,i) \text{ is stable if } \rho_t\big(\pi_t(i)\big) = i. \tag{7}$$

This cycle-check conservatively filters unreliable correspondences, which is crucial as the matching between adjacent frames can still be noisy (see Tab. 6). For cyclic stable links, we fuse colour probabilities with a product update:

$$\tilde{P}_{t,i}(c) \propto P_{t,i}(c) \cdot P_{t-1,\pi_t(i)}(c), \qquad P_{t,i} \leftarrow \text{Normalise}(\tilde{P}_{t,i}), \quad c \in \mathcal{C}. \tag{8}$$

Intuitively, Eq. (8) reinforces colour propagation for hard cases by leveraging palette probabilities carried by regions in other frames of the same video, while a cycle-consistency gate prevents error propagation. We perform one forward sweep ($t = 2 \to T$) and one backward sweep ($t = T - 1 \to 1$). In the backward sweep, we apply the same matching, cycle check, and fusion with indices swapped. In practice, the two passes are complementary: each direction conditions on different neighbouring frames that may already have more reliable estimates, so bidirectional sweeping improves robustness and yields more stable colourisation.

4 Experiments

4.1 Experiment Setup

We evaluate on the existing PaintBucket benchmarks [8, 9] sourced from CG-Rendered and hand-drawn videos. We further test our method against previous works on a newly constructed long-shot dataset, with 10–20× the length of target video, compared to previous datasets. Further experiments of segment matching on generic video dataset and ablations are provided in Supp. F.

PaintBucket-Character (PBC-3D) dataset contains 22 characters with 9–16 reference design-sheets per character; following prior work [8,29], we report results on the official test split of 3,000 frames.

PaintBucket-Real (PBC-Real) is a hand-drawn test set collected from professional animation, with 200 frames in total (20 short clips). Since it is not provided with design-sheet references, we use the first coloured frame from the clips as references, following previous works [9, 10, 29].

Anita-Pirate. We annotated a new long-shot benchmark featuring a hand-drawn 206-frame sequence, with raw data from the Anita dataset [1], which is more challenging than PBC-Real due to its 10–20× longer time horizon with roughly 140 segments per frame. It comes with frames with complex region layouts over a long time horizon. This new challenging test set is available on our project page, with annotation protocol and licensing details.

Evaluation protocols: We consider several production-relevant settings. Firstly, we evaluate standard design-sheet key-frame colourisation, following previous works [8,29], in Tab. 1. This setting provides a few coloured design sheets without assuming any temporal relation to the target video, and is therefore the most general reference protocol [8, 29]. We also evaluate video colourisation under the same-video key-frame colourisation, in which references are sampled from the target video itself. Note that this is the only reference type available for PBC-Real and Anita-Pirate without design sheets. We report two protocols from prior works. 1) First-frame reference (Tab. 3): only the first frame is given as an RGB reference [29]. 2) Two-sided reference (in-between): only the first and last frames are given as RGB references [10] (Tab. 4). Following previous works [8–10, 29], we report both segment- and pixel-level metrics: `Acc`, `Acc-Thresh` (segments>10 pixels), `Pix-Acc`, `Pix-F-Acc` (for foreground), and `Pix-B-MIoU` (for background). All metrics are reported in percentages, and larger values mean better performance. See Supp. B.1 for further details.

Table 1: One-shot key-frame (design-sheet) colourisation on PBC-3D. We follow prior work and use a single design-sheet reference image to colourise the target video. We compared the **PeCA** framework on both the trained DACoN 1.1 Model and other frozen foundation models with the **base** setting that matches the segmented region features for prediction. **Training-free** indicates whether the backbone is used without further training (✓) or requires training (✗) on colourisation tasks.

Method / Backbone	Training-free	Acc (%)	Acc-Thresh (%)	Pix-Acc (%)	Pix-F-Acc (%)	Pix-B-MIoU (%)
ColorFlow [67]	✗	9.72	10.81	50.64	9.16	57.17
MangaNinja [22]	✗	14.86	16.73	7.11	28.52	0.00
AniDoc [27]	✗	19.80	22.68	77.38	46.46	87.32
Cobra [68]	✗	15.06	17.26	69.20	19.72	82.69
MagicColor [62]	✗	21.48	24.81	16.34	44.04	7.63
BasicPBC-Ref [8]	✗	52.55	56.73	90.53	72.33	94.56
DACoN [29]	✗	67.87	72.58	96.99	91.00	99.08
DACoN 1.1 [29]	✗	68.01	72.87	96.97	91.03	99.11
DACoN 1.1 + PeCA	✗	**72.04** (4.03↑)	**77.08** (4.21↑)	**97.90** (0.93↑)	**94.04** (3.01↑)	**99.42** (0.31↑)
SAM2.1-Large (Base) [38]	✓	34.54	38.95	86.76	54.12	88.37
SAM2.1-Large + PeCA	✓	**46.65** (12.11↑)	**49.92** (10.97↑)	**88.70** (1.94↑)	**66.96** (12.84↑)	**96.70** (8.33↑)
DINOv3 ConvNeXT-L (Base) [47]	✓	34.90	36.35	71.32	49.79	75.93
DINOv3 ConvNeXT-L + PeCA	✓	**45.88** (10.98↑)	**46.97** (10.62↑)	**80.13** (8.81↑)	**60.15** (10.36↑)	**85.38** (9.45↑)
SigLIPv2 ViT-B/16 (Base) [52]	✓	48.64	51.68	89.24	70.05	91.03
SigLIPv2 ViT-B/16 + PeCA	✓	**55.34** (6.70↑)	**58.88** (7.20↑)	**92.48** (3.24↑)	**80.37** (10.32↑)	**93.88** (2.85↑)
DINOv2 ViT-L/14 (Base) [31]	✓	57.49	61.86	95.35	87.24	97.45
DINOv2 ViT-L/14 + PeCA	✓	**61.38** (3.89↑)	**65.58** (3.72↑)	**96.25** (0.90↑)	**89.31** (2.07↑)	**98.62** (1.17↑)

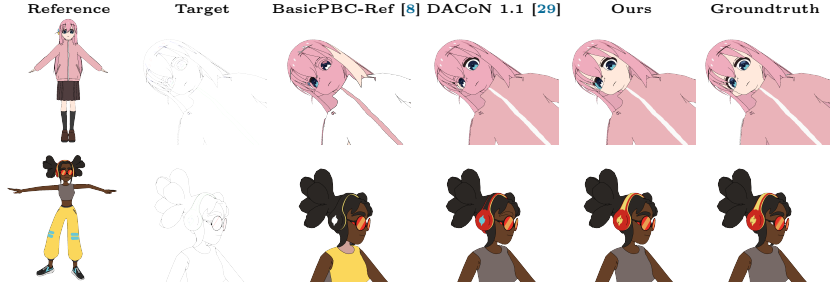
Implementation details: We compare against representative pixel-generative methods [22, 27, 67] and segment-based pipelines [8, 9, 29] under official protocols [8, 29]. We build a *Base* inference for SOTA [29] and frozen foundation models that perform colourisation using region correspondence, following prior works. We then apply PeCA as a plug-and-play inference framework to various backbone models, either with or without colourisation task training, and fixed hyperparameters (top-k=64, τ=0.05 and #views $B = 31$, m=4.) across all experiments. Further implementation details, computational costs, and hyperparameter settings are provided in the Supp. B.

4.2 Main Experimental Results

Results on design-sheet key-frame colourisation. Tab. 1 reports one-shot key-frame results with design-sheet references on PBC-3D. On task-trained models, the Palette Context Assisted (PeCA) framework further improves the current SOTA on all metrics consistently. Notably, PeCA yields substantially larger gains on training-free backbones, indicating that PeCA's test-time context reasoning unlocks region matching animation colourisation even for foundation models without colourisation training. Such gain also persists when more key-frame references are given, as shown in Tab. 2, which reports 5-shot and max-shot results on PBC-3D. This further shows that more reference shots do not dilute PeCA's contribution to the task. Fig. 5a shows representative qualitative results, showing our method can help the model overcome a huge appearance gap between references and target frames when previous base inference failed to.

Table 2: Key-frame (design-sheet) colourisation on PBC-3D with more references. PeCA show consistent gains under multi-reference protocols from [8, 29].

# of Refs	Method / Backbone	Training-free	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
5-shot refs	ColorFlow [67]	✗	12.64	14.37	54.51	15.26	61.22
	BasicPBC-Ref [8]	✗	–	64.59	96.12	83.17	98.67
	DACoN [29]	✗	73.25	77.44	97.74	93.70	99.13
	DACoN 1.1 [29]	✗	73.91	78.23	97.84	94.28	98.92
	DACoN 1.1 + PeCA	✗	**77.73** (3.82↑)	**82.39** (4.16↑)	**98.87** (1.03↑)	**97.02** (2.74↑)	**99.45** (0.53↑)
	SAM2.1-Large (Base) [38]	✓	43.80	46.59	87.66	62.25	96.75
	SAM2.1-Large + PeCA	✓	**57.23** (13.43↑)	**60.96** (14.37↑)	**91.50** (3.84↑)	**76.52** (14.27↑)	**97.18** (0.43↑)
	DINOv2 ViT-L/14 (Base) [31]	✓	62.65	66.42	96.77	91.54	97.96
	DINOv2 ViT-L/14 + PeCA	✓	**66.46** (3.81↑)	**70.01** (3.59↑)	**97.73** (0.96↑)	**93.57** (2.03↑)	**98.83** (0.87↑)
max-shot refs	DACoN [29]	✗	74.31	78.48	98.04	94.27	99.10
	DACoN 1.1 [29]	✗	75.05	79.23	98.19	94.79	99.16
	DACoN 1.1 + PeCA	✗	**79.03** (3.98↑)	**83.43** (4.20↑)	**99.01** (0.82↑)	**97.21** (2.42↑)	**99.55** (0.39↑)
	SAM2.1-Large (Base) [38]	✓	46.40	49.30	87.98	63.27	96.59
	SAM2.1-Large + PeCA	✓	**56.88** (10.48↑)	**60.50** (11.20↑)	**91.94** (3.96↑)	**77.49** (14.22↑)	**97.29** (0.70↑)
	DINOv2 ViT-L/14 (Base) [31]	✓	63.84	67.67	97.07	91.70	98.28
	DINOv2 ViT-L/14 + PeCA	✓	**67.28** (3.44↑)	**70.82** (3.15↑)	**97.71** (0.64↑)	**93.63** (1.93↑)	**98.59** (0.31↑)



(a) Key-frame (Design-sheet) Reference Colourisation Results

(b) First-frame Reference Colourisation Results

Fig. 5: Qualitative comparison under two production formulations. We show one-shot reference and target line sketch, followed by results from **BasicPBC(-Ref)** [8, 9], **DACoN 1.1** [29], our **PeCA** on DACoN 1.1, and colour ground-truth (right).

Results on same-video key-frame colourisation. Tab. 3 reports colourisation results on PBC-3D and PBC-Real using only the first frame as a reference to colourise the rest of the video. Including PeCA inference provides consistent improvements on the trained DACoN 1.1 pipeline on both domains, and

Table 3: First-frame Colourisation on PBC-3D and PBC-Real. We include both colourisation-trained methods (✗) and results using frozen backbones (✓). Additional analyses of more baselines [11, 19, 63] are provided in Supp. C.2.

Method / Backbone	Training-free	PBC-3D					PBC-Real				
		Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
BasicPBC [9]	✗	56.28	60.14	93.00	77.25	97.19	59.31	62.00	91.84	72.50	98.39
BasicPBC [9] (Online*)	✗	53.18	58.28	93.57	79.92	96.19	57.28	60.47	92.74	74.92	98.35
DACoN [29]	✗	69.91	73.59	97.30	-	-	65.85	69.15	93.50	-	-
DACoN 1.1 [29]	✗	70.34	74.04	97.30	91.13	99.17	65.82	69.11	94.18	80.68	98.76
DACoN 1.1 + PeCA	✗	**74.41**	**78.08**	**98.11**	**94.06**	**99.50**	**67.64**	**71.29**	**94.70**	**82.11**	**99.48**
StableDiffusion 2.1 (Base) [39]	✓	32.93	34.52	87.38	58.70	94.40	46.45	48.84	89.91	64.13	97.96
StableDiffusion 2.1 + PeCA	✓	**40.50**	**42.01**	**90.87**	**71.01**	**96.51**	**48.11**	**49.70**	**90.89**	**67.45**	**98.18**
SAM2.1-Large (Base) [38]	✓	49.10	52.46	91.64	72.40	97.38	55.63	58.31	90.32	69.21	98.73
SAM2.1-Large + PeCA	✓	**58.98**	**62.89**	**93.65**	**79.72**	**98.11**	**60.41**	**63.44**	**93.25**	**75.99**	**99.00**

* Online setting: the first frame uses the ground-truth reference, and each subsequent frame is colourised using the previous frame's prediction as the reference.

Fig. 6: Qualitative results of in-between colourisation on Anita-Pirate. It compares how our proposed PeCA performs against existing methods when dealing with a 10× longer target frame sequence.

again yields larger gains on training-free backbones. For in-between colourisation, Tab. 4 shows that our PeCA consistently improves multiple backbones on both PBC-3D (20-frame clips) and the new long-shot Anita-Pirate dataset. In the latter, only the first and last reference frames are provided with colours, while all the remaining 204 frames have to be coloured, given such a limited reference. All these suggest that test-time context reasoning with PeCA can substantially strengthen existing models' performance with different model types and supervisions. Such improvements persist when we naturally have better feature coverage by the temporal proximity of reference-target frames. Corresponding qualitative comparisons are provided in Fig. 5b and Fig. 6.

4.3 Ablations and Analysis

We analyse the three steps in the Palette Context Assisted (PeCA) inference framework on both task-trained and frozen foundation backbone under different key-frame reference types. Quantitative results are shown in Tab. 5.

Three steps are jointly contributing. Across both reference types and both backbones, we observe a consistent pattern. Active Reference Expansion (ARE) only gives a minor performance improvement. That means expanding the reference bank, even if in a target-aware way, can still lead to confusion, as it may

Table 4: In-between colourisation results. We report results on short video clips on both the existing PBC-3D [9] testset (20 frames) and a new challenging testset, Anita-Pirate, with 10× longer frame sequence containing multiple complex subjects.

Method/Backbone	Training-free	PBC-3D					Anita-Pirate				
		Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
BasicPBC [9]	✗	63.38	67.77	94.84	84.20	97.54	28.54	28.97	88.52	39.77	96.63
BasicPBC [9] (Online*)	✗	53.97	59.13	93.74	80.62	96.32	7.71	7.94	32.97	17.00	35.93
DACoN 1.1 [29]	✗	78.02	82.11	98.48	95.51	99.47	38.16	39.36	94.29	61.65	99.16
DACoN 1.1 + PeCA	✗	**80.80**	**84.82**	**99.00**	**97.18**	**99.58**	**41.24**	**42.16**	**94.29**	**62.78**	**99.43**
DINOv2 ViT-L/14 (Base) [31]	✓	66.25	70.17	97.73	93.36	98.89	28.55	29.30	93.06	53.88	99.40
DINOv2 ViT-L/14 [31] + PeCA	✓	**69.29**	**72.60**	**98.23**	**94.49**	**99.33**	**31.01**	**31.81**	**93.39**	**57.18**	**99.49**

* Online setting: the first frame uses the ground-truth reference, and each subsequent frame is colourised using the previous frame's prediction as the reference.

Table 5: Ablation study on colourisation under different reference and model types. DACoN 1.1 is pretrained on colourisation; DINOv2 (ViT-L/14) is used frozen in a zero-shot manner. ARE: Active Reference Expansion in Sec. 3.3, PA: Probability Aggregation in Sec. 3.4, CT: Cyclic-gated Temporal-fusion in Sec. 3.5.

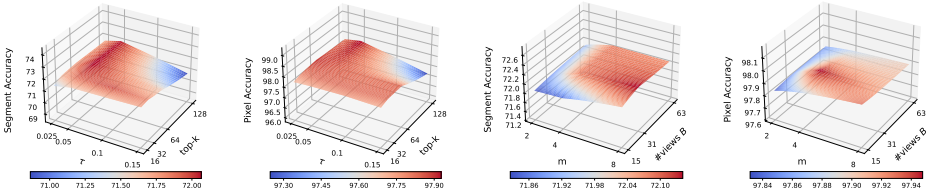
Backbone	ARE	PA	CT	Design-sheet (one-shot)					First-frame (one-shot)				
				Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
DACoN 1.1	✗	✗	✗	68.01	72.87	96.97	91.03	99.11	70.34	74.04	97.30	91.13	99.17
	✓	✗	✗	69.04	74.02	97.22	91.67	99.24	70.65	74.54	97.29	91.34	99.25
	✓	✓	✗	70.61	75.40	97.43	92.52	99.28	72.50	76.17	97.64	92.09	99.38
	✗	✓	✓	70.30	75.27	97.21	92.47	99.38	72.78	76.56	97.83	92.87	99.35
	✓	✗	✓	69.02	74.07	97.19	91.68	99.26	70.87	74.80	97.37	91.44	99.31
	✓	✓	✓	**72.04**	**77.08**	**97.90**	**94.04**	**99.42**	**74.41**	**78.08**	**98.11**	**94.06**	**99.50**
DINOv2 ViT-L/14	✗	✗	✗	57.49	61.86	95.35	87.24	97.45	59.50	62.99	96.25	87.95	98.47
	✓	✗	✗	58.74	63.43	95.96	88.14	98.57	60.92	64.53	96.52	88.87	98.64
	✓	✓	✗	60.64	64.66	95.94	88.32	98.60	62.51	65.61	96.80	89.00	99.06
	✗	✓	✓	58.68	62.35	94.56	84.52	97.99	61.00	63.77	96.54	88.76	99.07
	✓	✗	✓	58.51	63.29	95.89	88.08	98.52	61.16	64.85	96.70	89.10	98.97
	✓	✓	✓	**61.38**	**65.58**	**96.25**	**89.31**	**98.62**	**63.28**	**66.47**	**97.13**	**90.45**	**99.22**

dilute the exact match with more region candidates. Therefore, further adding Probability Aggregation (PA) yields substantial gains by soft-voting across multiple plausible matches, which reduces sensitivity to spurious matches. Cyclic-gated Temporal-fusion (CT) then provides additional improvements by exploiting temporal context inside the video, refining per-region colour probabilities. Combining all steps gives the strongest results, while removing any of them results in a considerable performance drop, indicating these context cues are complementary and reciprocal as expected, instead of isolated heuristics.

Step-wise contributions differ by reference types. The relative contributions of ARE and CT differ between the two reference types. Under design-sheet references, errors are often driven by spatial mismatch because design sheets can be far from the target shot in pose and region layout; correspondingly, ARE tends to contribute more by improving target-conditioned reference support. Under first-frame references, the reference comes from the same video, and the appearance gap is smaller, so temporal continuity becomes a stronger cue; CT therefore tends to provide more noticeable gains. The balance also depends on the backbone: frozen features benefit more from improved support and aggregation, whereas the task-trained model benefits more from temporal refinement, consistent with its stronger correspondence quality.

Table 6: Additional one-shot key-frame (design-sheet) colourisation ablations on PeCA's internal designs. ARE Selection uses ✗=*random*, ✓=*greedy* (Eq. (4)); CT Cycle Gate uses ✗=*off*, ✓=*on* (Eq. (7)).

Backbone	Selection	Cycle Gate	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
SAM2.1-Large [38]	✗	✓	47.90	51.41	87.60	63.84	96.74
	✓	✗	45.92	49.11	88.19	65.49	97.18
	✓	✓	**50.69**	**54.59**	**89.10**	**69.05**	**97.39**
CLIP ViT-L/14 [35]	✗	✓	46.45	48.70	88.61	68.66	90.44
	✓	✗	40.30	41.83	88.64	67.45	91.37
	✓	✓	**48.58**	**50.94**	**90.51**	**73.77**	**92.19**



(a) Segment-Acc (top-k vs. τ). **(b)** Pixel-Acc (top-k vs. τ). **(c)** Segment-Acc (#views B vs. m). **(d)** Pixel-Acc (#views B vs. m).

Fig. 7: Hyperparameter sensitivity for PeCA inference under key-frame (one-shot design-sheet referenced) colourisation. We visualise the sensitivity of (top-k, τ) in PA (Sec. 3.4) and (#views B vs. m) in ARE Sec. 3.3, reported under Segment Accuracy and Pixel Accuracy.

Ablation on view selection and cyclic gating. We additionally ablate two module-internal designs in Tab. 6. For ARE, greedy facility-location selection (Eq. (4)) consistently outperforms keeping all the random views under the same budget, supporting that the gain comes from coverage-aware selection rather than merely adding more augmented views. For CT, disabling the cyclic gate (Eq. (7)) leads to a substantial performance drop, indicating that gating is necessary to prevent unreliable temporal matches from propagating errors over time.

Hyperparameter sensitivity. Fig. 7a and Fig. 7b visualise the sensitivity of PA (top-k, τ) and ARE (#views B, exploration factor m) under key-frame colourisation. Across a wide range of top-k and τ, all metrics vary mildly, indicating stable behaviour. The observed degradations match expected trends. As $\tau \to 0$ and top-$k \to 1$, probability aggregation degenerates to simple nearest-neighbour [29]. Increasing top-k and τ makes aggregation less selective and approaches a global mixing of candidates, similar in spirit to linear combination schemes used in prior work [7], which can overly flatten the colour distribution and let low-quality matches dilute the prediction. Thus, we select the default setting (τ=0.05, top-k=64) that retains multiple plausible matches while keeping the palette colour probability mass concentrated on high-confidence candidates.

Meanwhile, Fig. 7c and Fig. 7d show that in ARE, the number of selected views B trades off coverage against cost, while m controls the size of the explored

candidate pool (mB) for greedy selection. Performance saturates quickly as B increases, and increasing m yields only marginal gains beyond moderate search breadth. We therefore adopt B=31 and m=4 as a cost-effective operating point.

5 Conclusion

To sum up, we introduced a training-free and plug-and-play Palette Context Assisted (PeCA) inference framework for production-level paint-bucket colourisation, where each region must follow a correct colour in a predefined palette. PeCA improves region matching by constructing and validating test-time context: it strengthens spatial context with a target-aware reference expansion, unifying noisy matching evidence in colour space, and uses adjacent frames as temporal context to support prediction. In particular, such context can provide additional support for colour propagation paths when direct matches are ambiguous. Experiments on existing benchmarks and a new long-video test case showed consistent gains on both task-trained models and frozen backbones.

Limitations. Similar to prior paint-bucket colourisation methods that rely on region segmentation and a well-defined palette, PeCA cannot fully resolve failures caused by line leakage or by colours/regions missing from the available references. This limitation is inherent to palette-constrained paint-bucket colourisation, which requires strict, discrete colour assignment from reference palette to target sketch regions. This construction prevents the system from hallucinating colours that are not supported by reference evidence. We provide further discussions and failure case analysis of the task formulation and current methods in Supp. H.

Beyond these limitations, a promising direction is to move from a fixed-reference setting to an interactive or open-world reference-growth setting. Instead of relying only on the provided key frames or design sheets, future systems may actively acquire additional colour evidence through artist correction feedback, generated reference views, or external character-sheet retrieval. Such extensions would help cover missing poses, parts, and colours, while preserving the strict palette constraints required by paint-bucket colourisation. We view this as a natural next step for extending the concept of test-time context reasoning introduced by PeCA toward practical production toolkits.

Acknowledgement

This project is partially supported by an Amazon Research Award, the EPSRC Doctoral Landscape Award (DLA) Collaborative Studentships with Industry and Allsee Technologies Ltd. The computations in this research were partially performed using the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP\T022221\1) and the Digital Research Infrastructure programme (EP\W032244\1) and is operated by Advanced Research Computing at the University of Birmingham.

Supplementary Material

This supplementary document provides additional details on dataset construction, implementation, computation cost, diagnostic analyses, extended experiments, qualitative results, and limitations. For ease of navigation, we provide a roadmap of the supplementary content that supports the main paper:

- **Sec. A: Details on Annotation and Statistics of the New Test Data.** This section expands the details on the newly introduced Anita-Pirate data, with the complete construction pipeline, licensing information, and dataset statistics/comparisons that justify its role as a long-video stress test.
- **Sec. B: Additional Implementation Details.** This section provides further details for reproduction: metric definitions, backbone configurations, detailed view expansion steps and end-to-end runtime comparisons.
- **Sec. C: Further Comparisons to Previous Works.** This section reports an additional evaluation under the same shorter-clip in-between protocol related to Feng *et al*. [10] and additional details on pixel-generative baselines.
- **Sec. D: Additional Analysis on Reference Expansion and Probability Aggregation.** This section supports the motivation of the main paper (Sec. 3.3–3.4) by analysing how the "quality" of colour probabilities changes as the number of reference views and correspondence aggregation differs.
- **Sec. E: Additional Temporal Stability Analysis.** This section further proves that our method achieved better video-level temporal consistency. We report an additional temporal stability metric together with curves and qualitative results, revealing how performance evolves over time.
- **Sec. F: Extension to Natural Video Region Label Propagation.** This section supports the generality of PECA beyond paint-bucket colourisation task by conceptually extending it to a new task of semantic label propagation over regions from a generic video segmentation dataset.
- **Sec. G: More Qualitative Results.** This section provides additional visual comparisons under different reference settings and backbones, complementing the representative examples shown in the main paper.
- **Sec. H: Limitations.** This section expands the limitations discussion by detailing failure modes related to imperfect line segmentation and incomplete reference coverage, together with future directions.

We also include an accompanying video named `PeCA.mp4` with additional qualitative results.

A Details on Annotation and Statistics of Anita-Pirate

For the same-video colourisation experiments in the main paper, we introduce Anita-Pirate, a stress-test case study with a much longer video, to validate the long-horizon stability of our method. To construct this stress test, we select the longest continuous sequence from Anita [1], which provides hand-drawn line sketches paired with intermediate colourisation targets. The source animation

video is licensed under a CC-BY licence. Our goal is to convert the raw data into production-ready line sketches paired with region-level colour annotations for standard paint-bucket evaluation. This requires addressing several mismatches between the raw data format and production-style annotation.

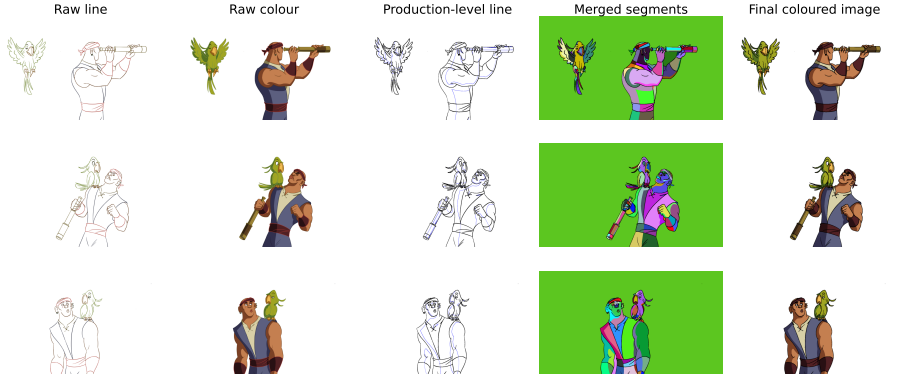


Fig. S1: Visualisation of Anita-Pirate construction pipeline: Raw line, Raw colour, Production-level line, Merged segments, and Final coloured image, each corresponds to an intermediate stage of annotation.

To begin with, we first verified that all frames have properly enclosed regions by applying flood-fill segmentation [48] that reveals enclosed regions, and manually fixed the detected leakages. After that, another issue in the original data is that shadow and highlight boundaries are completely absent from the raw line sketches (first column in Fig. S1). We therefore infer such boundaries from colour discontinuities by running flood-fill segmentation on coloured images (second column in Fig. S1) and merging the boundaries into the line map. In the resulting line sketches, original artist lines are preserved in black, while newly introduced shadow/highlight separators are encoded in pure blue following industrial convention, yielding near production-level line sketches in the third column of Fig. S1.

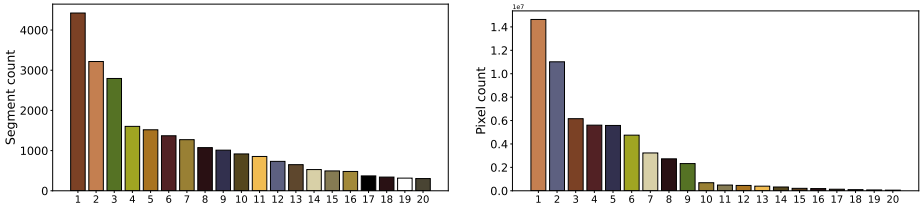
After line merging, we performed flood-fill segmentation again on the constructed production-level lines and assigned the majority colour to each region as illustrated in the last two columns of Fig. S1. It can be seen that the resulting frames recover the appearances and colours from the raw coloured frames successfully. We then export both the region index map and the segment-to-RGBA mapping, which are directly compatible with PBC datasets [8].

The final Anita-Pirate test set contains 28,773 annotated segments in total. Key dataset-level comparisons to PBC-3D and PBC-Real [9] are summarised in Tab. S1. Beyond these per-frame statistics, we further visualise the palette distribution of Anita-Pirate in Fig. S2. These statistics indicate that the newly introduced test case Anita-Pirate features a longer sequence, denser region layouts, and a richer colour palette.

Table S1: Statistical comparison of Anita-Pirate against existing test sets.

Dataset	PBC-3D [9]	PBC-Real [9]	Anita-Pirate (Ours)
Avg. Video Length (Frame)	20	10	**206**
Frame Resolution	1024 × 1024	mixed†	1920 × 1080
Avg. Regions per Frame	68.26	89.19	**139.67**
Unique RGBA Colours	196	271	**366**

†PBC-Real contains mixed resolutions: 512 × 512, 1024 × 1024, 1280 × 1280, and 1600 × 1600.



(a) Top-20 colours by segment count. **(b)** Top-20 colours by pixel count.

Fig. S2: Non-transparent colour-distribution visualisation for Anita-Pirate.

B Additional Implementation Details

B.1 Evaluation Metrics

In the main paper experiment Sec.4, we evaluate at both segment level and pixel level using exact discrete RGBA equality after palette colour decoding. Let N be the number of valid target segments in a frame, a_i be the pixel area of segment i, y_i and $\hat{y}_i$ be ground-truth and predicted RGBA labels, and $\alpha(\cdot)$ denote the alpha channel. We list the complete metric calculations below:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i]. \tag{i}$$

$$\text{Acc-Thresh} = \frac{1}{|I_{>10}|}\sum_{i\in I_{>10}} \mathbb{1}[\hat{y}_i = y_i], \quad I_{>10} = \{i \mid a_i > 10\}. \tag{ii}$$

$$\text{Pix-Acc} = \frac{\sum_i a_i\, \mathbb{1}[\hat{y}_i = y_i]}{\sum_i a_i}. \tag{iii}$$

$$\text{Pix-F-Acc} = \frac{\sum_i a_i\, \mathbb{1}[\alpha(y_i) > 0]\, \mathbb{1}[\hat{y}_i = y_i]}{\sum_i a_i\, \mathbb{1}[\alpha(y_i) > 0]}, \tag{iv}$$

where the background is defined by transparency ($\alpha = 0$). Let $b_i = \mathbb{1}[\alpha(y_i) = 0]$ and $\hat{b}_i = \mathbb{1}[\alpha(\hat{y}_i) = 0]$. The background IoU at pixel level is:

$$\text{Pix-B-MIoU} = \frac{\sum_i a_i\, \mathbb{1}[b_i = 1 \wedge \hat{b}_i = 1]}{\sum_i a_i\, \mathbb{1}[b_i = 1 \vee \hat{b}_i = 1]}. \tag{v}$$

All final metrics are averaged over all per-frame metrics evaluated. Here, `Acc` measures segment-wise exact RGBA accuracy (all segments equally weighted), while `Acc-Thresh` excludes tiny segments to reduce noise. `Pix-Acc` is equivalent to pixel-value accuracy across all pixels, `Pix-F-Acc` restricts `Pix-Acc` to foreground (non-transparent) regions, and `Pix-B-MIoU` measures the IoU of predicted vs. ground-truth background (transparent) pixels.

B.2 Details on Foundation Model Usages

In the main text Sec.4, we have run several experiments using various vision backbone models. For these foundation models or colourisation-pretrained models, segmented region descriptors are obtained by region pooling over dense feature maps within region masks obtained by flood-fill [48], and matched by cosine similarity in L_2-normalised feature space. For colourisation-trained methods/backbones, we leverage their official checkpoints provided [8,9,29]. In addition to these previous works, we list the details of other frozen foundation model sources in Tab. S2. (Note that for Stable Diffusion [39], we follow diffusion-feature extraction practice from existing works [29,49] with prompt "a photo of an anime character.") As shown, our experiments cover a broad spectrum of models with different capacities, supervisions, usages and resolutions, demonstrating the model-agnostic generality of PeCA.

Performance differences between backbones should therefore be interpreted as differences in descriptor quality and pretraining task bias. Self-supervised features tend to provide stronger region identity, while visual-language and generative diffusion model features are less directly optimised for local region correspondences. PeCA uses the same region-pooling and matching interface for all backbones, so its gains are measured relative to each backbone's base inference.

Table S2: Foundation backbones and configurations used in our inference pipeline.

Backbone	Checkpoint / Model ID	Input Size	Feature Used	Source
DINOv2 ViT-L/14	`facebookresearch/dinov2:dinov2_vitl14`	518×518	final patch-token feature map	[31]
CLIP ViT-L/14	`ViT-L/14@336px`	336×336	visual encoder patch features	[35]
DINOv3 ConvNeXT-L	`timm/convnext_large.dinov3_lvd1689m`	512×512	`forward_features` map (timm)	[47]
SigLIPv2 ViT-B/16	`timm/vit_base_patch16_siglip_512.v2_webli`	512×512	`forward_features` map (timm)	[52]
SAM2.1-Large	`facebook/sam2.1-hiera-large`	512×512	image predictor visual features	[38]
Stable Diffusion 2.1	`sd2-community/stable-diffusion-2-1`	768×768	U-Net first upsampling block, $t = 261/1000$	[39]

B.3 More Details on Active Reference Expansion (ARE).

In main Sec. 3.3, we mentioned that ARE candidates are generated by applying joint palette-preserving geometric transforms to the reference triplet (line image, segment map, colour image). Specifically, the transform order follows:

$$T = T_{\text{affine}} \circ T_{90^\circ} \circ T_{\text{vflip}} \circ T_{\text{hflip}}. \tag{vi}$$

Table S3: Geometric transformation parameters used in Active Reference Expansion Step (Main text Sec. 3.3).

Transform	Apply Prob.	Parameters
Horizontal flip (T_{hflip})	0.5	NA
Vertical flip (T_{vflip})	0.1	NA
90° rotation (T_{90°)	0.2	$k \sim \text{Unif}\{1,2,3\}$, rotate by $90^\circ \times k$
Affine (T_{affine})	1.0	angle $\theta \sim \text{Unif}[-30^\circ, 30^\circ]$; translation $(\Delta x, \Delta y)$ up to ±50% of image width/height; scale $s \sim \text{Unif}[0.5, 2.0]$; shear = 0

Each transform is independently determined by probability p. For affine, we use rotation + uniform scale + translation (no shear). Given an original pixel in homogeneous coordinates $\mathbf{p} = [x, y, 1]^\top$, the transformed point is

$$\mathbf{p}' = \mathbf{A}\mathbf{p}, \quad \mathbf{A} = \begin{bmatrix} s\cos\theta & -s\sin\theta & \Delta x \\ s\sin\theta & s\cos\theta & \Delta y \\ 0 & 0 & 1 \end{bmatrix}, \tag{vii}$$

where $\theta \sim \mathcal{U}[-30^\circ, 30^\circ]$, $s \sim \mathcal{U}[0.5, 2.0]$, and $(\Delta x, \Delta y)$ is sampled translation. Following standard image affine warping, the transform is applied around the image centre (c_x, c_y):

$$\mathbf{A} = \mathbf{T}(\Delta x, \Delta y)\,\mathbf{T}(c_x, c_y)\,\mathbf{R}(\theta)\,\mathbf{S}(s)\,\mathbf{T}(-c_x, -c_y). \tag{viii}$$

We fix these transformation parameters (Tab. S3) for all experiments. When multiple reference images are provided, we split the budget B evenly across references: we generate $mB/|\mathcal{R}|$ candidates and select $B/|\mathcal{R}|$ views per reference. Lastly, the selection is computed against $|\mathcal{T}_s| = 20$ target frames uniformly subsampled from the target video to bound the selection cost.

B.4 Computation Cost

We report overall runtime on Anita-Pirate, which exhibits the highest per-frame segment complexity among benchmarks (Sec. A) using different backbone models with PECA. All measurements are obtained on a single NVIDIA A100 GPU with batch size 1 and FP32 inference. We report the total runtime of each method under the same evaluation setting. As shown in Fig. S3, although PECA introduces additional test-time computation, the overall pipeline remains substantially faster than earlier diffusion-based or inclusion-matching pipelines [8,9,22].

Overall, these results suggest that the full method remains efficient. Note that all these runtimes are practical in production settings, where manual colouring typically requires minutes per tens of frames [25].

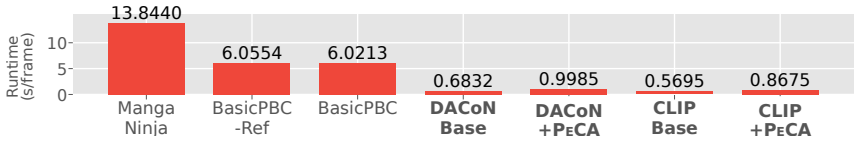


Fig. S3: Overall latency per frame on Anita-Pirate. All measurements are obtained on an NVIDIA A100 GPU with batch size 1 and FP32 inference.

C Further Comparisons to Previous Works

C.1 Comparison to a Unified Framework [10]

We did not report a direct quantitative comparison to Feng *et al.* [10], since their code, checkpoints, and new evaluation data are not publicly available at the time of submission. Moreover, their in-between evaluation protocol focuses on short clips (lengths of 3, 5, and 10 frames), rather than a complete video, which is not directly comparable to our in-between setting, which considers the rest of the whole video as the target. Nevertheless, we managed to test our method under shorter frame sequences (10 frames) as Feng *et al.* [10] did on the PBC-3D dataset, with results shown in Tab. S4. The simplest *Base* training-free feature-matching baseline already achieves competitive performance with previous works. Adding PECA on top of this baseline further improves the results.

Table S4: Short-sequence inbetweening colourisation on PBC-3D with shorter clips of 10 frames, following Feng *et al.* [10]. Training-free indicates whether the backbone model is trained on colourisation tasks. (Note that RAFT [51] used optical flow-based matching.)

Method	Training-free	Acc ↑	Acc-Thresh ↑	Pix-Acc ↑	Pix-F-Acc ↑	Pix-B-MIoU ↑
ToonCrafter [56]	✗	9.69	13.11	22.48	12.92	25.64
MangaNinja [22]	✗	14.21	14.44	53.47	24.73	57.84
LVCD [13]	✗	26.59	28.66	58.38	42.94	60.19
BasicPBC [9]	✗	53.26	56.66	90.88	71.92	96.56
Feng *et al.* [10]	✗	68.67	72.63	95.42	87.09	97.80
RAFT [51]	✓	32.06	36.07	60.74	52.88	88.80
SAM2.1	✓	67.92	72.13	96.40	89.33	98.49
SAM2.1 + PECA	✓	**73.18**	**77.17**	**97.16**	**92.48**	**98.73**

C.2 Details on More Recent Pixel-Generative Baselines

Here we provide further details and additional comparisons against more recent pixel-generative baselines. Since these methods output RGB images (as shown in Fig. S4) rather than region-to-palette assignments, we follow the DACoN post-processing protocol [29]. *I.e.*, We resize each generated image to the target resolution, replace each pixel with the nearest colour from the reference palette,

and unify each line-enclosed segment to its most frequent projected colour. The resulting palette-preserving images are then evaluated by the same metrics to previous works [8, 29].

Unless otherwise stated, we use the official/default inference settings for each baseline. For Nano Banana 2, a closed-source model, we use the Google Cloud API (`gemini-3.1-flash-image-preview`) [11] with the default generation api call and the following prompt:

Prompt for Nano Banana 2 (`gemini-3.1-flash-image-preview`) [11]

```
Colorize the target line art using the colored reference character image. The first
image is the colored reference. The second image is the target line art. Preserve
the target line art, pose, composition, and plain background. Use only colors
visible in the reference image. Return only the final colored image.
```

Table S5: Modern pixel-generative baseline comparisons. Left: PBC-3D under the one-shot design-sheet key-frame reference protocol; right: PBC-Real under the first-frame reference protocol. RGB generation baselines are evaluated after DACoN-style post-processing before computing paint-bucket metrics.

PBC-3D: key-frame reference

Method	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
ColorFlow [67]	9.72	10.81	50.64	9.16	57.17
AniDoc [27]	19.80	22.68	77.38	46.46	87.32
Cobra [68]	15.06	17.26	69.20	19.72	82.69
MagicColor [62]	21.48	24.81	16.34	44.04	7.63
DACoN 1.1 + PECA	**72.04**	**77.08**	**97.90**	**94.04**	**99.42**

PBC-Real: first-frame reference

Method	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
AnimeColor [63]	37.39	40.22	85.25	59.24	90.19
ToonComposer [19]	29.03	31.28	32.43	48.02	22.62
Nano Banana 2 [11]	47.78	52.17	90.39	71.63	98.46
DACoN 1.1 [29]	65.82	69.11	94.18	80.68	98.76
DACoN 1.1 + PECA	**67.64**	**71.29**	**94.70**	**82.11**	**99.48**

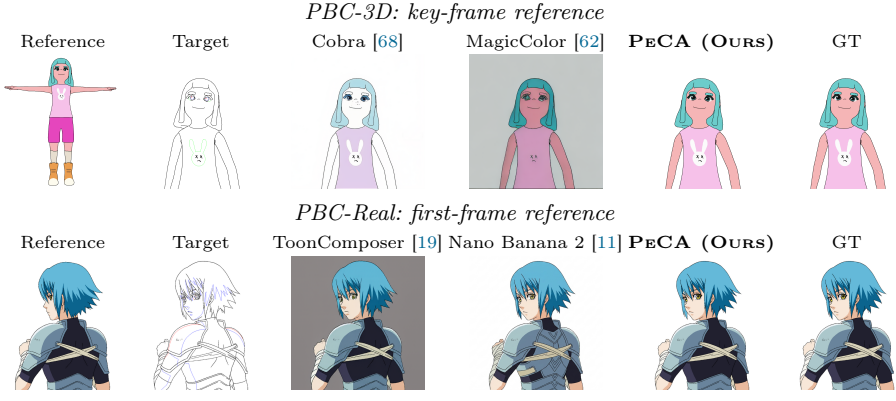


Fig. S4: Raw qualitative comparison for modern pixel-generative baselines. We show the references, target line sketches, raw RGB outputs and GT under the PBC-3D key-frame reference protocol and the PBC-Real first-frame reference protocol.

The evaluation results in Tab. S5 and Fig. S4 show that pixel generative methods, even when using much more powerful model structures [32, 39] failed

to provide precise colourisation over the simple line sketch inputs. All of them appear to be altering the original line structure with some artefacts. Specifically, a stronger closed-source Nano Banana 2 model [11] hallucinates significant patterns that never existed in input line sketches. Apart from these, some methods [19, 62] show background bias, which failed to predict emptiness for background, this may be inherited from their pretraining dataset Sakuga-42M [66], which always has rich backgrounds from completed animation.

D Additional Analysis on Reference Expansion and Probability Aggregation

As discussed in the main paper Sec. 3.3 and Sec. 3.4, increasing the number of reference views has two opposing effects. On the one hand, more views improve target coverage and increase the chance of retrieving the correct correspondence, which is the main motivation behind Active Reference Expansion (ARE). On the other hand, a larger reference pool also introduces more distractor matches, so the benefit of additional views can only be realised when the aggregation rule is sufficiently selective. This is exactly the role of our Probability Aggregation (PA). In other words, ARE enlarges the pool of potentially useful colour evidence, while PA is needed to convert that larger evidence pool into actual gains without suffering from dilution.

To make this connection more explicit, we analyse how the quality of the induced colour distribution changes as the number of reference views increases. We use DINOv2 [31] as the frozen backbone, keep all other settings fixed, and vary the total number of random reference views included as $R \in \{1, 31, 63, 128\}$.

We compare three inference-time colour propagation rules from correspondence probabilities: top-1 hard copy as in main text Eq. (2), full linear combination over all source matches [7] (can be viewed as a special case for Eq. (6) from main text with $k = +\infty, \tau = 1$), and the PA soft-voting introduced in main paper Sec. 3.4. Note that PA soft-voting uses the default hyperparameters in Main Eq. (6).

To characterise the quality of the colourisation probability $P_{t,i}$, we measure it from two complementary aspects. First, *Uncertainty* is measured by `Entropy`, which quantifies how concentrated the predicted colour distribution is (lower is better). Second, *Discriminability* is measured by `GT Margin`, which quantifies how strongly the ground-truth colour is separated from the strongest competing colour (higher is better). We compute both metrics on non-transparent segments only ($\alpha(y^{gt}_{t,i}) > 0$). Let N_{nt} denote the total number of non-transparent segments in the evaluation frames:

$$\text{Entropy} = \frac{1}{N_{\text{nt}}} \sum_{\alpha(y^{gt}_{t,i})>0} \left[-\sum_{c \in \mathcal{C}} P_{t,i}(c) \log P_{t,i}(c) \right], \tag{ix}$$

$$\text{GT Margin} = \frac{1}{N_{\text{nt}}} \sum_{\alpha(y^{gt}_{t,i})>0} \left[P_{t,i}(y^{gt}_{t,i}) - \max_{c \neq y^{gt}_{t,i}} P_{t,i}(c) \right]. \tag{x}$$

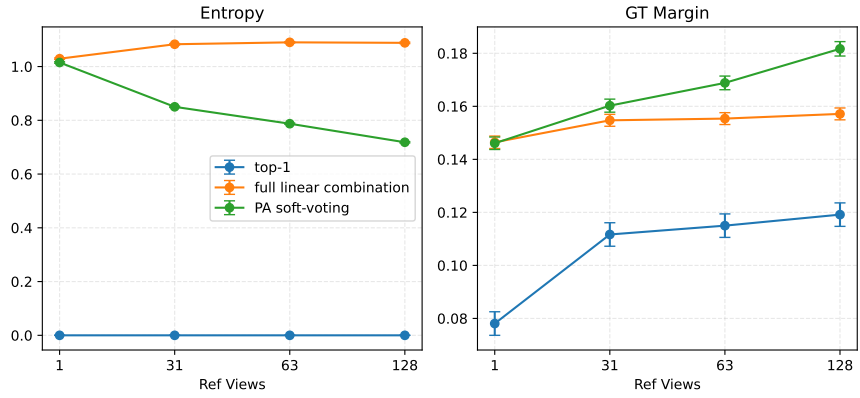


Fig. S5: Colour probability quality vs. number of reference views R. `Entropy` (↓) measures uncertainty of $P_{t,i}$, and `GT Margin` (↑) measures the probability gap between the ground-truth colour and the strongest competitor. Top-1 hard copy yields near-zero `Entropy` by construction (one-hot), but has the lowest `GT Margin`, indicating less robust decisions under ambiguous matches. As R increases, PA soft-voting improves both confidence (lower `Entropy`) and discriminability (higher `GT Margin`), while full linear combination saturates due to evidence dilution.

We conduct the above experiments under the one-shot key-frame colourisation setting on PBC-3D [9]. From Fig. S5, we observe three consistent trends beyond the performance gains already reported in the ablations (main text Tab. 5). First, top-1 prediction keeps near-zero `Entropy` across all R, which is expected from its one-hot prediction, but it also yields the lowest `GT Margin`, indicating limited robustness when the best match is ambiguous. Second, a full linear combination shows the opposite trend: as R grows, `Entropy` further increases and then saturates at a relatively high level, while `GT Margin` improves only mildly. This indicates that simply adding more views is not sufficient; without the selective soft-voting, the additional evidence is increasingly diluted by distractor correspondences. This observation is also consistent with the hyperparameter analysis in the main paper (Fig. 7(a,b)): enlarging the aggregation range by increasing top-k or τ, thereby moving the behaviour closer to full combination, leads to clear performance drop, while reducing top-k, which pushes the model towards top-1 direct matching, also weakens the benefit of expanding reference set.

In contrast, PA soft-voting exhibits the desired behaviour for leveraging larger reference pools: as R increases, `Entropy` decreases while `GT Margin` increases steadily, and the gap to both top-1 and full linear combination becomes more pronounced at larger R. Taken together, these results support both the motivation that expanding reference views indeed provides more useful matching evidence, but such possibly noisy evidence requires a controlled aggrega-

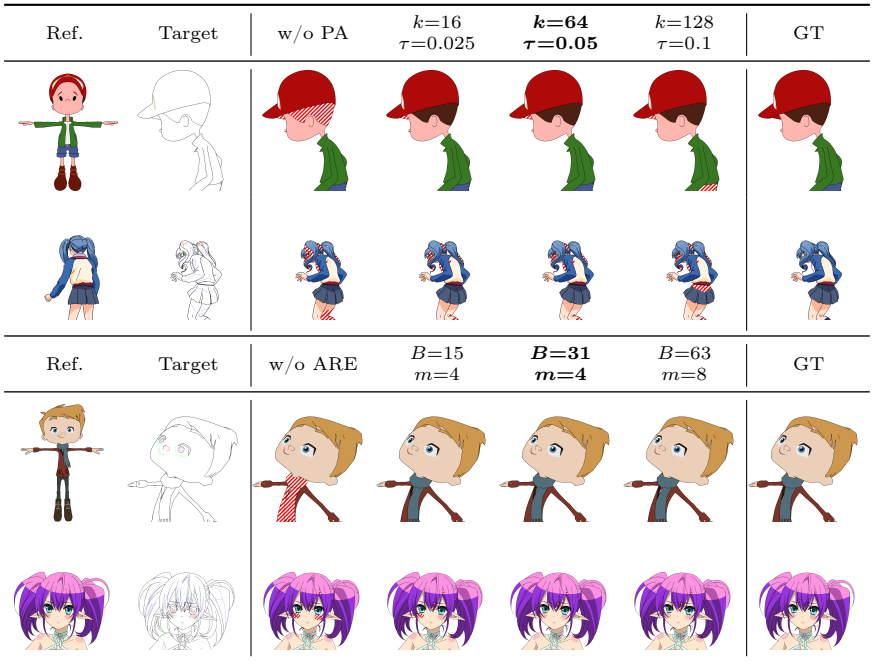


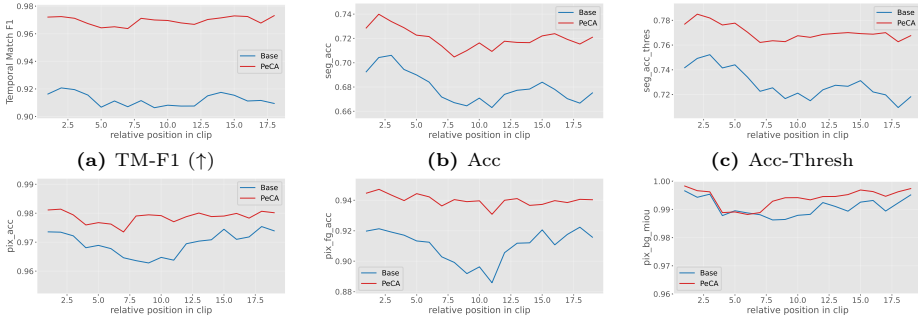
Fig. S6: Qualitative hyperparameter comparisons. Rows 1–2 vary PA settings, and rows 3–4 vary ARE selection budget. ▨ marks wrong-colour foreground regions under each setting.

tion mechanism to be effectively converted into performance gains rather than through indiscriminate mixing.

The numerical analysis above therefore provides explanations to the provided qualitative examples for extreme and default hyperparameter settings in Fig. S6. The results show a trend aligning with quantitative hyperparameter analysis. Overly hard PA under-aggregates useful evidence, while broader PA moves closer to indiscriminate colour mixing; very small selection budgets reduce reference support, whereas larger budgets saturate around the fixed default. These qualitative behaviours are consistent with the quantitative hyperparameter surfaces in the main paper.

E Additional Temporal Stability Analysis

To further validate robustness throughout the video, we compare *Base* and PeCA using aggregate curves over relative clip position under the one-shot design-sheet key-frame colourisation setting for PBC-3D [9] dataset. We report five standard per-frame quality metrics, together with an additional temporal stability metric defined below.



(a) TM-F1 (↑) **(b)** Acc **(c)** Acc-Thresh

(d) Pix-Acc **(e)** Pix-F-Acc **(f)** Pix-B-MIoU

Fig. S7: Metric curves over relative video clip position (*Base* vs. PeCA). PeCA improves temporal consistency (TM-F1) and consistently improves segment/pixel quality over time, mitigating long-horizon error accumulation.

Temporal stability metric definition. Since exact ground-truth segment trajectories between adjacent frames are unavailable, we evaluate temporal stability on surrogate temporal correspondences, following the same strategy used in CT. Concretely, for each target segment i at frame t, we first define its nearest temporal link to frame $t-1$ by feature matching using DACoN 1.1 features [29]:

$$j_t^*(i) = \arg\max_j \ \left\langle \bar{\mathbf{f}}_{t,i}, \bar{\mathbf{f}}_{t-1,j} \right\rangle . \tag{xi}$$

Not all nearest-neighbour links are reliable, so we further retain only cycle-consistent links as stable:

$$\text{stable}(t,i) = \mathbb{1}\left[i = \arg\max_{i'} \ \left\langle \bar{\mathbf{f}}_{t-1,j_t^*(i)}, \bar{\mathbf{f}}_{t,i'} \right\rangle \right] . \tag{xii}$$

This filtering removes spurious matches and restricts evaluation to reliably trackable regions. On each stable link, we then evaluate whether the ground-truth label is temporally consistent, and whether the prediction preserves the same consistency pattern:

$$g_{t,i} = \mathbb{1}\left[y_{t,i} = y_{t-1,j_t^*(i)} \right] , \qquad \hat{g}_{t,i} = \mathbb{1}\left[\hat{y}_{t,i} = \hat{y}_{t-1,j_t^*(i)} \right] . \tag{xiii}$$

Intuitively, $g_{t,i} = 1$ means the true colour should stay the same along the link, while $\hat{g}_{t,i} = 1$ means the method predicts no colour change. We then compute per-step precision and recall over stable links:

$$P_t = \frac{\sum_i \mathbb{1}[\text{stable}(t,i)]\, \hat{g}_{t,i} g_{t,i}}{\sum_i \mathbb{1}[\text{stable}(t,i)]\, \hat{g}_{t,i}}, \qquad R_t = \frac{\sum_i \mathbb{1}[\text{stable}(t,i)]\, \hat{g}_{t,i} g_{t,i}}{\sum_i \mathbb{1}[\text{stable}(t,i)]\, g_{t,i}}. \tag{xiv}$$

Finally, the temporal stability metric is defined as the per-step F1 score:

$$\text{TM-F1}_t = \frac{2 P_t R_t}{P_t + R_t}, \qquad t = 2, \ldots, T. \tag{xv}$$

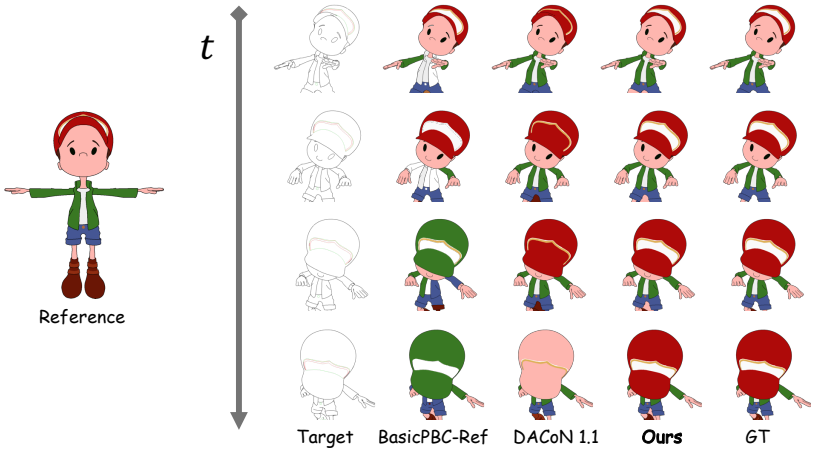


Fig. S8: Visualisation of temporal stability. Compared with previous methods (BasicPBC-Ref [8], *Base* DACoN 1.1 [29]), our method shows significantly fewer "colour flickers" over time. More qualitative results are provided in the accompanying video.

This completes the calculation of the surrogate temporal stability metric TM-F1. Higher TM-F1 indicates better temporal consistency, *i.e.* fewer "colour flickers" in the output video. As shown in Fig. S7, PECA remains consistently above the base inference across nearly the entire clip for all metrics. In particular, TM-F1 improves from around 91% (*Base*) to around 97% (PECA), indicating substantially better temporal stability throughout the sequence. Overall, these curves suggest that PECA not only improves average colourisation quality, but also mitigates error accumulation over time. We provide additional qualitative visualisations for such stability in Fig. S8 and in the accompanying video included.

F Extension to Natural Video Region Label Propagation

To test whether PECA generalises beyond palette-based colour assignment, we evaluate it on a reference-guided Region Label Propagation task built on the panoptic video segmentation dataset VIPSeg [28]. Given a target video and a small set of external reference frame(s), the goal is to assign each target superpixel a semantic label from its correspondences to reference superpixels. Ground-truth superpixel labels are induced from VIPSeg panoptic annotations via maximum overlap, and we evaluate predictions using Segment-wise accuracy (`Seg-Acc`), as well as pixel-level accuracy and Mean IoU (`Pix-Acc, Pix-MIoU`).

Specifically, we use the VIPSeg validation split (343 videos, 8,255 frames) as targets. For each target frame, we over-segment the RGB image into SLIC superpixels [2]. Each target superpixel is assigned a semantic category by max-

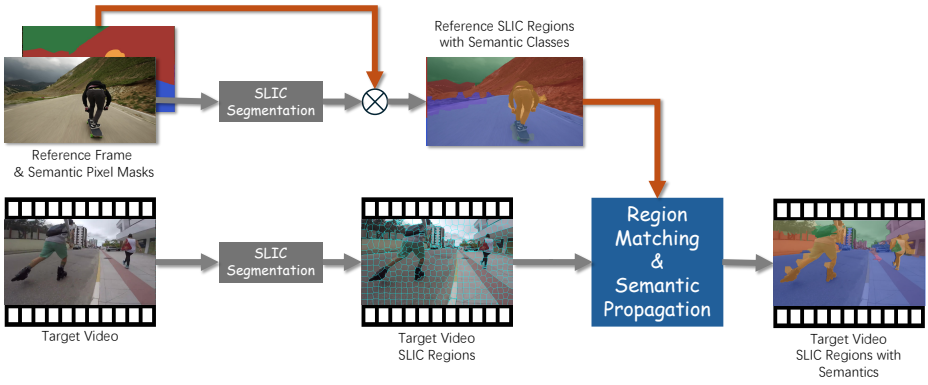


Fig. S9: Task formulation for reference-guided region label propagation for natural videos. Given an external reference frame with panoptic semantic masks, we compute SLIC superpixels and assign each reference superpixel a semantic class according to the original panoptic annotation. For the target video, we compute SLIC superpixels for each frame. Region matching then propagates semantic labels from reference to target SLIC regions, producing per-region semantic predictions for evaluation.

imum overlap with the panoptic mask, which serves as the ground-truth label for evaluation.

For this diagnostic experiment, we use the ground-truth panoptic labels to identify the semantic classes present in each target video, and then greedily select a small set of external reference frames from the VIPSeg training split whose union covers these classes. We apply the same SLIC over-segmentation and maximum-overlap label assignment to the selected reference frames, yielding reference superpixels with semantic labels.

With these reference–target pairs constructed, we compare two inference pipelines as in Sec. 4. *Base* uses backbone features and direct matching (main paper Eq. (2)) to perform nearest-neighbour superpixel matching and hard label transfer. PeCA enables the full inference pipeline with the same default hyperparameters as in the colourisation experiments, aggregating matched semantic labels to obtain the final prediction. We report results with two generic pretrained backbones, DINOv2 ViT-L/14 [31] and SAM2.1-Large [38]. Metrics are computed per frame and averaged over all evaluation frames following Sec. B.1.

As shown in Tab. S6, PeCA yields consistent gains across both backbones and all three metrics, with qualitative examples in Fig. S10. Although this task is outside our main scope of paint-bucket colourisation, the improvements suggest that PeCA captures a more general form of reference-guided region matching that transfers to natural videos.

Fig. S10: Qualitative results on VIPSeg superpixel region label propagation. From left to right: first external reference frame with label map, target RGB frame (input), *Base*, PeCA, and ground truth. Compared with direct matching, PeCA produces more spatially coherent predictions with fewer fragmented labels and better object coverage. Ground truth denotes the superpixel semantic labels induced from the VIPSeg [28] annotations.

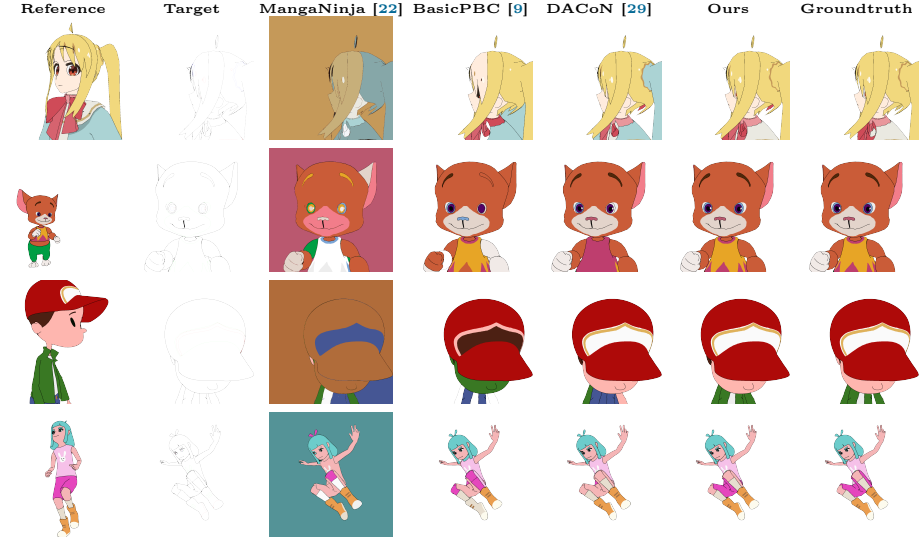
Table S6: VIPSeg Region Label Propagation results. All numbers are frame-wise averages. PeCA consistently improves over direct hard matching across both backbones and all metrics.

Backbone	Pipeline	Seg-Acc (%)	Pix-Acc (%)	Pix-MIoU (%)
SAM2.1-Large	*Base*	33.35	33.05	6.78
	PeCA (ours)	**38.95**	**38.79**	**10.85**
DINOv2 ViT-L/14	*Base*	44.12	44.03	12.68
	PeCA (ours)	**52.47**	**52.38**	**19.23**

G More Qualitative Results

We provide further qualitative results on various settings in separate figures below. Specifically, we visualise a more comprehensive set of test samples under different methods and references in Fig. S11. Note that for the pixel-generative method [22], we follow the post-processing steps in previous work [29] to convert

(a) Key-frame (Design-sheet) Reference Colourisation Results



(b) First-frame Reference Colourisation Results

Fig. S11: Additional qualitative comparison under different references. We show the one-shot reference and target line sketch, followed by the results from **MangaNinja** [22], **BasicPBC(-Ref)** [8,9], **DACoN 1.1** [29], our **PECA** on DACoN 1.1, and colour ground-truth (right).

the raw results to palette-preserving paint-bucket colourisations. In general, our method shows superior performance under challenging test cases.

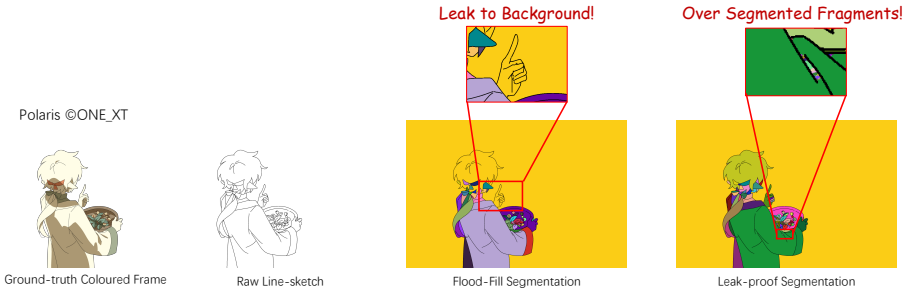


Fig. S12: Segmentation failure modes on amateur-level line sketches from [1]. From left to right: coloured reference, raw sketch, standard flood-fill [48] segmentation, and leakage-robust segmentation. Standard flood-fill may leak into the background when strokes are not fully closed, while leakage-robust segmentation reduces leakage, but often over-segments the drawing into fragments. These results highlight the gaps between amateur-level and production-ready line sketches that paint-bucket colourisation task formulations [8, 9] and industry-level animation workflows [30] assume.

H Limitations

Like previous paint-bucket colourisation methods [8–10, 29], PeCA expects line sketches with sufficiently enclosed regions, so that simple flood-fill segmentation [48] can be applied. This assumption is consistent with standard animation production workflows [30, 50] and is therefore largely inherited from the paint-bucket formulation rather than introduced by our method. When applied to raw drafts or amateur sketches with broken strokes and leakage, region extraction may fail and subsequently affect matching and colour assignment.

Possible ways to relax this assumption include draft-line gap closing [41, 57] and leakage-robust region segmentation [3, 59]. The latter preserves the original drawing and therefore has been preferred as additional lines may break the original structure of the target animation. But it often produces more fragmented regions, which can make matching less stable and increase manual colourisation workload (with many more fragments to colour) as shown in Fig. S12. Bridging the gap between raw or amateur-level sketches and production-ready line art is therefore a promising and largely orthogonal direction for future work [12].

Another limitation comes from incomplete reference coverage. Like other reference-guided colourisation methods [8–10, 29], PeCA can only propagate colours that are represented in the available references. Missing views, colours, or part appearances may lead to systematic colour confusion as shown in Fig. S13. Accordingly, the geometric transformations in PeCA should be understood as lightweight in-plane spatial support, which further reduces moderate pose or layout gaps, but do not synthesise out-of-plane 3D rotations or colours for surfaces never observed in the reference pool.

A possible practical direction is a more interactive reference-selection workflow that allows dynamic reference growth (as an online setting) in paint-bucket

colourisation process: instead of simply increasing the number of manually coloured keyframes (which shifts the workload back to the user and thus reduces the benefit of automation), artists may choose to colour a small subset of the most informative keyframes or design-sheet views for downstream automatic colourisation. Because PECA selects and reweights from the supplied reference pool, it can also benefit from expanded pools provided by artist interaction [30], external knowledge base/prompt retrieval [20, 61], or generated novel-view references [21, 33] without changing the paint-bucket output interface. An efficient selection strategy based on layout complexity or feature coverage, as we already attempted with PECA, may be a valuable direction to further accelerate automation. To sum up, we view this gap as a promising orthogonal direction for future work.

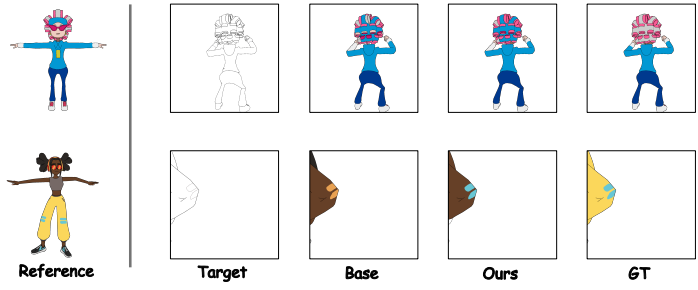


Fig. S13: Failure cases caused by incomplete reference coverage. From left to right: reference, target line sketch, *Base* and PECA predictions, and ground truth. Top: missing reference coverage for the target pose leads to incorrect colour assignment on the helmet back regions. Bottom: a very small visible part in the target frame lacks sufficient colour evidence in the reference, resulting in local colour confusion.

References

1. Anita dataset. https://zhenglinpan.github.io/AnitaDataset_homepage/, accessed: 2024-06-24
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels. Technical report, EPFL (06 2010)
3. Allen, B., Maejima, A., Anjyo, K.: Fast leak-resistant segmentation for anime line art. In: SIGGRAPH Asia 2024 Technical Communications. SA '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3681758.3698003, https://doi.org/10.1145/3681758.3698003
4. Cao, R., Mo, H., Gao, C.: Line art colorization based on explicit region segmentation. In: Computer Graphics Forum. vol. 40, pp. 1–10. Wiley Online Library (2021)
5. Cao, Y., Meng, X., Mok, P., Lee, T.Y., Liu, X., Li, P.: AnimeDiffusion: Anime diffusion colorization. IEEE Transactions on Visualization and Computer Graphics **30**(10), 6956–6969 (2024)

6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Casey, E., Pérez, V., Li, Z.: The animation transformer: Visual correspondence via segment matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11323–11332 (2021)
8. Dai, Y., Li, Q., Zhou, S., Luo, Y., Li, C., Loy, C.C.: Paint bucket colorization using anime character color design sheets. arXiv preprint arXiv:2410.19424 (2024)
9. Dai, Y., Zhou, S., Li, Q., Li, C., Loy, C.C.: Learning inclusion matching for animation paint bucket colorization. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 25544–25553 (2024)
10. Feng, X., Huang, T., Wang, P., Huang, Z., Haihang, Z., Zou, Y., Li, D., Zou, K.: A unified framework for industrial cel-animation colorization with temporal-structural awareness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19301–19310 (October 2025)
11. Google: Nano Banana 2: Combining Pro capabilities with lightning-fast speed — blog.google. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, [Accessed 21-06-2026]
12. Guajardo, J., Bursalioglu, O., Goldman, D.B.: Generative ai for 2d character animation. In: ACM SIGGRAPH 2024 Posters, pp. 1–2 (2024)
13. Huang, Z., Zhang, M., Liao, J.: LVCD: reference-based lineart video colorization with diffusion models. ACM Transactions on Graphics (TOG) **43**(6), 1–11 (2024)
14. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker: It is better to track together. In: European Conference on Computer Vision. pp. 18–35. Springer (2024)
15. Kim, S., Park, D., Shim, B.: Semantic-aware superpixel for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1142–1150 (2023)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
17. Krause, A., Golovin, D.: Submodular function maximization. Tractability **3**(71-104), 3 (2014)
18. Li, J., Liang, Q., Li, Q., Gang, R., Fang, J., Lin, C., Feng, S., Liu, X.: RTTLC: Video colorization with restored transformer and test-time local converter. pp. 1722–1730 (06 2023). https://doi.org/10.1109/CVPRW59228.2023.00173
19. Li, L., Wang, G., Zhang, Z., Li, Y., Li, X., Dou, Q., Gu, J., Xue, T., Shan, Y.: Tooncomposer: Streamlining cartoon production with generative post-keyframing. arXiv preprint arXiv:2508.10881 (2025)
20. Liu, R., Pei, J., Zhu, J.: From pixels to personas: Tracking the evolution of anime characters. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 20, pp. 1488–1504 (2026)
21. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: 2023 IEEE/CVF International Conference on Computer Vision. pp. 9264–9275. IEEE (2023)
22. Liu, Z., Cheng, K.L., Chen, X., Xiao, J., Ouyang, H., Zhu, K., Liu, Y., Shen, Y., Chen, Q., Luo, P.: Manganinja: Line art colorization with precise reference following. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5666–5677 (2025)

23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (Nov 2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94, https://doi.org/10.1023/B:VISI.0000029664.99615.94
24. Maejima, A., Kubo, H., Funatomi, T., Yotsukura, T., Nakamura, S., Mukaigawa, Y.: Graph matching based anime colorization with multiple references. In: ACM SIGGRAPH 2019 Posters, pp. 1–2 (2019)
25. Maejima, A., Kubo, H., Shinagawa, S., Funatomi, T., Yotsukura, T., Nakamura, S., Mukaigawa, Y.: Anime character colorization using few-shot learning. In: SIGGRAPH Asia 2021 Technical Communications, pp. 1–4 (2021)
26. Manli, S., Weili, N., De-An, H., Zhiding, Y., Tom, G., Anima, A., Chaowei, X.: Test-time prompt tuning for zero-shot generalization in vision-language models. In: Advances in Neural Information Processing Systems (2022)
27. Meng, Y., Ouyang, H., Wang, H., Wang, Q., Wang, W., Cheng, K.L., Liu, Z., Shen, Y., Qu, H.: AniDoc: Animation creation made easier. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18187–18197 (2025)
28. Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., Yang, Y.: Large-scale video panoptic segmentation in the wild: A benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2022)
29. Nagata, K., Kaneko, N.: DACoN: Dino for anime paint bucket colorization with any number of reference images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17899–17908 (2025)
30. Nakanishi, H., Shichijo, N., Sugi, M., Ogata, T., Hara, T., Ota, J.: Modeling the process of animation production. Int. J. Autom. Technol. **7**(4), 439–450 (2013)
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
33. Peng, H.Y., Zhang, J.P., Guo, M.H., Cao, Y.P., Hu, S.M.: Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. ACM Transactions on Graphics (TOG) **43**(4), 1–13 (2024)
34. Qiao, R., Tan, Q., Yang, M., Dong, G., Yang, P., Lang, S., Wan, E., Wang, X., Xu, Y., Yang, L., et al.: V-thinker: Interactive thinking with images. arXiv preprint arXiv:2511.04460 (2025)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
36. Radovanovic, M., Nanopoulos, A., Ivanovic, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of machine learning research **11**(sept), 2487–2531 (2010)
37. Ramassamy, S., Kubo, H., Funatomi, T., Ishii, D., Maejima, A., Nakamura, S., Mukaigawa, Y.: Pre-and post-processes for automatic colorization using a fully convolutional network. In: SIGGRAPH Asia 2018 Posters, pp. 1–2 (2018)
38. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF

conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (June 2022)
40. Sadihin, B.C., Meng, Y., Wang, M.H., Chen, M.J., Su, H.: TimeColor: Flexible reference colorization via temporal concatenation. arXiv preprint arXiv:2601.00296 (2026)
41. Sasaki, K., Iizuka, S., Simo-Serra, E., Ishikawa, H.: Joint gap detection and inpainting of line drawings. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5725–5733 (2017)
42. Schuurmans, M., Berman, M., Blaschko, M.B.: Efficient semantic image segmentation with superpixel pooling. arXiv preprint arXiv:1806.02705 (2018)
43. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: Better aggregation in test-time augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1214–1223 (2021)
44. Shi, M., Zhang, J.Q., Chen, S.Y., Gao, L., Lai, Y.K., Zhang, F.L.: Reference-based deep line art video colorization. IEEE Transactions on Visualization and Computer Graphics **29**(6), 2965–2979 (2022)
45. Shlapentokh-Rothman, M., Blume, A., Xiao, Y., Wu, Y., TV, S., Tao, H., Lee, J.Y., Torres, W., Wang, Y.X., Hoiem, D.: Region-based representations revisited. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 17107–17116 (2024)
46. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
47. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
48. Smith, A.R.: Tint fill. In: Proceedings of the 6th Annual Conference on Computer Graphics and Interactive Techniques. pp. 276–283. SIGGRAPH '79, Association for Computing Machinery, New York, NY, USA (1979). https://doi.org/10.1145/800249.807456, https://doi.org/10.1145/800249.807456
49. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36**, 1363–1389 (2023)
50. Tang, Y., Guo, J., Liu, P., Wang, Z., Hua, H., Zhong, J.X., Xiao, Y., Huang, C., Song, L., Liang, S., et al.: Generative AI for cel-animation: A survey. arXiv preprint arXiv:2501.06250 (2025)
51. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
52. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: SIGLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
53. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
54. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)
55. Wittgenstein, L.: Remarks on Colour. University of California Press (1977), https://books.google.co.uk/books?id=xQhbwgEACAAJ

56. Xing, J., Liu, H., Xia, M., Zhang, Y., Wang, X., Shan, Y., Wong, T.T.: Tooncrafter: Generative cartoon interpolation. ACM Transactions on Graphics (TOG) **43**(6), 1–11 (2024)
57. Xu, P., Hospedales, T.M., Yin, Q., Song, Y.Z., Xiang, T., Wang, L.: Deep learning for free-hand sketch: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(1), 285–312 (2022)
58. Yang, Y., Fan, L., Lin, Z., Wang, F., Zhang, Z.: Layeranimate: Layer-level control for animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10865–10874 (October 2025)
59. Zhang, L., JI, Y., Liu, C.: Danbooregion: An illustration region dataset. In: European Conference on Computer Vision (ECCV) (2020)
60. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
61. Zhang, X., Du, H., Wei, X., Li, Q.: Omnicolor: A unified framework for multi-modal lineart colorization (2026), https://arxiv.org/abs/2603.27531
62. Zhang, Y., Ma, Y., Wang, B., Chen, Q., Wang, Z.: Follow-your-color: Multi-instance sketch colorization (2025), https://arxiv.org/abs/2503.16948
63. Zhang, Y., Wang, L., Wang, H., Wu, D., Lin, Z., Wang, F., Song, L.: Animecolor: Reference-based animation colorization with diffusion transformers. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6682–6690 (2025)
64. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: MMICL: Empowering vision-language model with multi-modal in-context learning. In: The Twelfth International Conference on Learning Representations
65. Zhao, Y., Zheng, H., Luo, J., Lam, E.Y.: Improving video colorization by test-time tuning. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 166–170. IEEE (2023)
66. Zhenglin Pan, Yu Zhu, Y.M.: Sakuga-42m dataset: Scaling up cartoon research. arXiv preprint arXiv:2405.07425 (2024)
67. Zhuang, J., Ju, X., Zhang, Z., Liu, Y., Zhang, S., Yuan, C., Shan, Y.: ColorFlow: Retrieval-augmented image sequence colorization. arXiv preprint arXiv:2412.11815 (2024)
68. Zhuang, J., Li, L., Ju, X., Zhang, Z., Yuan, C., Shan, Y.: Cobra: Efficient line art colorization with broader references. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)