# PeCA: Palette Context Assisted Inference for Test-Time Paint-Bucket Colourisation on Animation Videos (Supplementary Material)

Dongheng Lin and Jianbo Jiao

The MIx Group, University of Birmingham

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 contents that support 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.* [7] 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 `6189_supp.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.

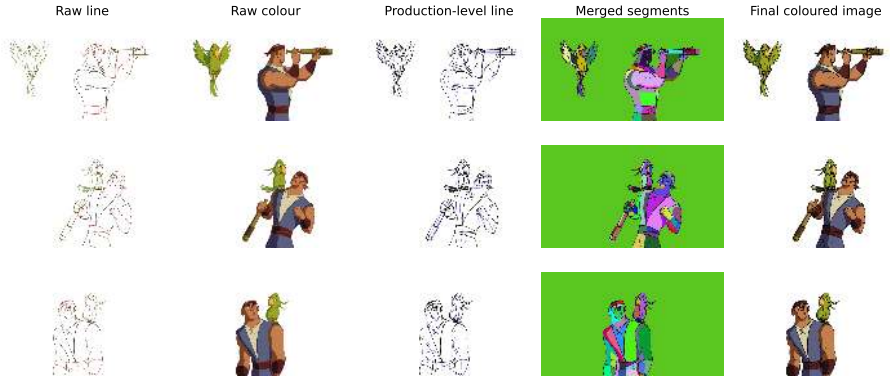



**Fig. S1: Visualisation of Anita-Pirate construction pipeline:** Raw line, Raw colour, Production-level line, Merged segments, and Final coloured image, each corresponds to an intermediate stage of annotation.

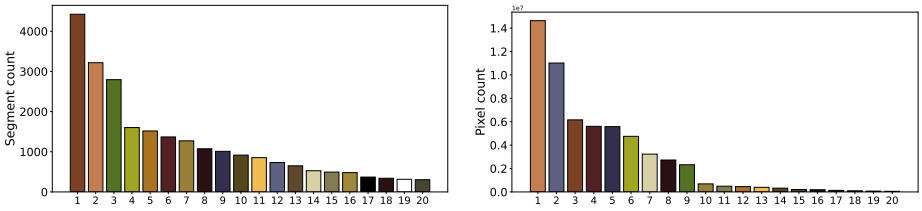
To begin with, we first verified that all frames have properly enclosed regions by applying flood-fill segmentation [24] that reveals enclosed regions, and manually fixed the detected leakages. After that, another issue in the original data is that shadow and highlight boundaries are completely absent 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 suc-

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

| Dataset | PBC-3D [6] | PBC-Real [6] | 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.

cessfully. We then export both the region index map and the segment-to-RGBA mapping, which are directly compatible with PBC datasets [5].

The final Anita-Pirate test set contains 28,773 annotated segments in total. Key dataset-level comparisons to PBC-3D and PBC-Real [6] 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.

## 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 [24], and matched by cosine similarity in $L_2$-normalised feature space. For colourisation-trained methods/backbones, we leverage their official checkpoints provided [5,6,16]. In addition to these previous works, we list the details of other frozen foundation model sources in Tab. S2. (Note that for Stable Diffusion [21], we follow diffusion-feature extraction practice from existing works [16, 25] 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.

### 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 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 | [18] |
| CLIP ViT-L/14 | `ViT-L/14@336px` | 336×336 | visual encoder patch features | [19] |
| DINOv3 ConvNeXT-L | `timm/convnext_large.dinov3_lvd1689m` | 512×512 | `forward_features` map (timm) | [23] |
| SigLIPv2 ViT-B/16 | `timm/vit_base_patch16_siglip_512.v2_webli` | 512×512 | `forward_features` map (timm) | [28] |
| SAM2.1-Large | `facebook/sam2.1-hiera-large` | 512×512 | image predictor visual features | [20] |
| Stable Diffusion 2.1 | `sd2-community/stable-diffusion-2-1` | 768×768 | U-Net first upsampling block, $t = 261/1000$ | [21] |

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

| Transform | Apply Prob. | Parameters |
|---|---|---|
| Horizontal flip ($T_{\text{hflip}}$) | 0.5 | NA |
| Vertical flip ($T_{\text{vflip}}$) | 0.1 | NA |
| 90° rotation ($T_{90^\circ}$) | 0.2 | $k \sim \text{Unif}\{1,2,3\}$, rotate by $90^\circ \times k$ |
| Affine ($T_{\text{affine}}$) | 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 [5, 6, 12].

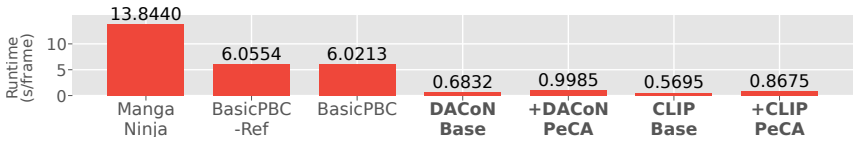



**Fig. S3: Overall latency per frame on Anita-Pirate.** All measurements are obtained on an NVIDIA A100 GPU with batch size 1 and FP32 inference.

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 [13].

## C Further Comparisons to Previous Works

We did not report a direct quantitative comparison to Feng *et al.* [7], 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.* [7] did on 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.* [7].** Training-free indicates whether the backbone model is trained on colourisation tasks. (Note that RAFT [27] used optical flow-based matching.)

| Method | Training-free | Acc ↑ | Acc-Thresh ↑ | Pix-Acc ↑ | Pix-F-Acc ↑ | Pix-B-MIoU ↑ |
|---|---|---|---|---|---|---|
| ToonCrafter [29] | ✗ | 9.69 | 13.11 | 22.48 | 12.92 | 25.64 |
| MangaNinja [12] | ✗ | 14.21 | 14.44 | 53.47 | 24.73 | 57.84 |
| LVCD [10] | ✗ | 26.59 | 28.66 | 58.38 | 42.94 | 60.19 |
| BasicPBC [6] | ✗ | 53.26 | 56.66 | 90.88 | 71.92 | 96.56 |
| Feng *et al.* [7] | ✗ | 68.67 | 72.63 | 95.42 | 87.09 | 97.80 |
| RAFT [27] | ✓ | 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.1 More Recent Pixel-Generative Baselines

We further compare against recent pixel-generative baselines, including both earlier methods already covered in the main paper and newer reference-guided

systems. Since these methods output RGB images rather than region-to-palette assignments, we follow the DACoN post-processing protocol [16]: 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 with the same metrics as the main paper. The raw generated outputs are used in qualitative figures to show the original behaviour of pixel-generative models before this metric conversion. Unless otherwise stated, we use the official/default inference settings for each baseline. For Nano Banana 2, a closed-source model, we use the `gemini-3.1-flash-image-preview` Google Cloud API with the default generation api call and the following prompt:

**Prompt for Nano Banana 2**

```
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 [34] | 9.72 | 10.81 | 50.64 | 9.16 | 57.17 |
| AniDoc [14] | 19.80 | 22.68 | 77.38 | 46.46 | 87.32 |
| Cobra [35] | 15.06 | 17.26 | 69.20 | 19.72 | 82.69 |
| MagicColor [32] | 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 [33] | 37.39 | 40.22 | 85.25 | 59.24 | 90.19 |
| ToonComposer [11] | 29.03 | 31.28 | 32.43 | 48.02 | 22.62 |
| Nano Banana 2 [8] | 47.78 | 52.17 | 90.39 | 71.63 | 98.46 |
| DACoN 1.1 [16] | 65.82 | 69.11 | 94.18 | 80.68 | 98.76 |
| **DACoN 1.1 + PeCA** | **67.64** | **71.29** | **94.70** | **82.11** | **99.48** |

## 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.

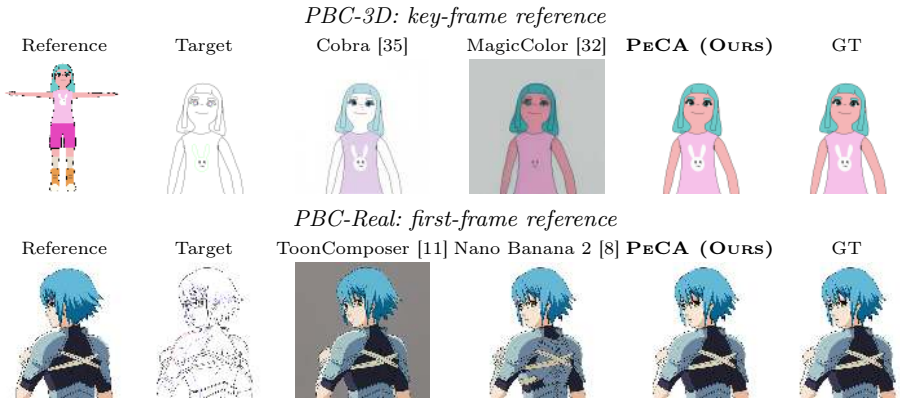



**Fig. S4: Raw qualitative comparison for modern pixel-generative baselines.** We show the references, target line sketches, raw RGB outputs and GT under the PBC-3D key-frame reference protocol and the PBC-Real first-frame reference protocol.

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 [18] 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 [4] (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 $N_{\text{nt}}$ denote the total number of non-transparent segments in the evaluation frames:

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

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

We conduct the above experiments under the one-shot key-frame colourisation setting on PBC-3D [6]. From Fig. S5, we observe three consistent trends be-

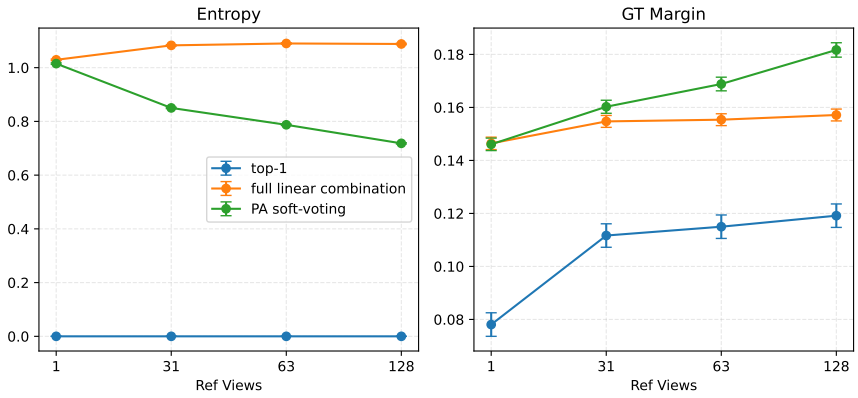



**Fig. S5: Colour probability quality vs. number of reference views $R$.** `Entropy` (↓) measures uncertainty of $P_{t,i}$, and `GT Margin` (↑) measures the probability gap between the ground-truth colour and the strongest competitor. Top-1 hard copy yields near-zero `Entropy` by construction (one-hot), 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.

yond 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, 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 $\tau$, 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 aggregation mechanism to be effectively converted into performance gains rather than through indiscriminate mixing.

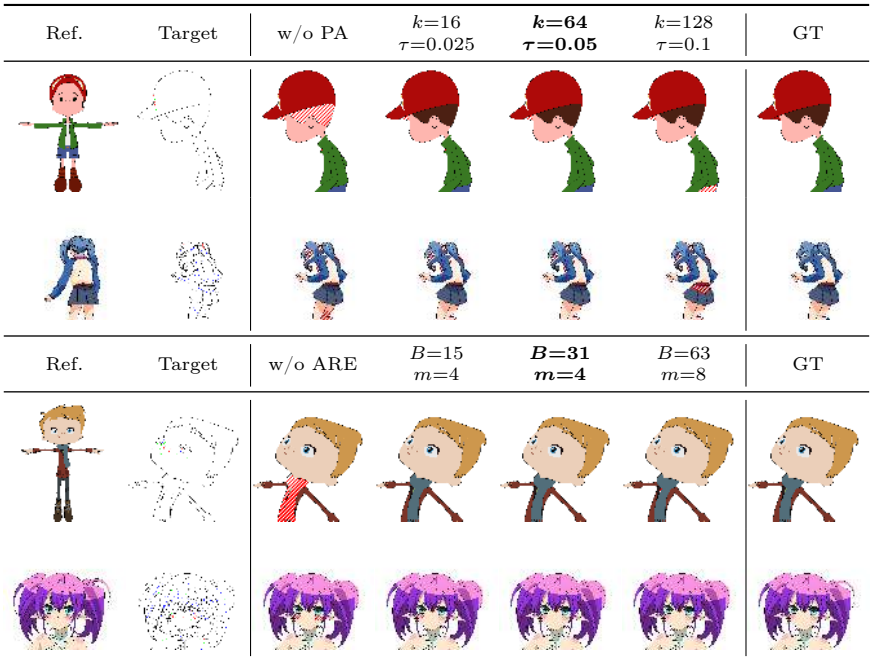


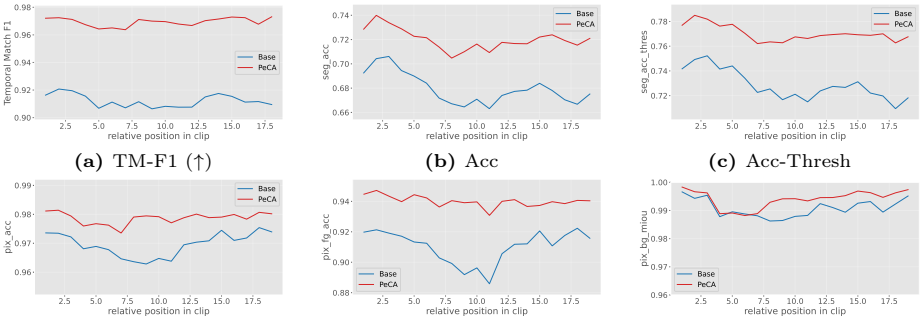

**Fig. S6: Qualitative hyperparameter comparisons.** Rows 1–2 vary PA settings, and rows 3–4 vary ARE selection budget. ▨ marks wrong-colour foreground regions under each setting.

We further provide qualitative examples for extreme and default hyperparameter settings in Fig. S6. The examples are selected from the automatic sweep outputs by foreground accuracy changes. The extremes produce interpretable degradations: 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 [6] dataset. We report five standard per-frame quality metrics, together with an additional temporal stability metric defined below.

***Temporal stability metric definition.*** Since exact ground-truth segment trajectories between adjacent frames are unavailable, we evaluate temporal stability




**(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.

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 [16]:

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

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:

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

This completes the calculation of the surrogate temporal stability metric TM-F1. Higher TM-F1 indicates better temporal consistency, *i.e.* fewer "colour

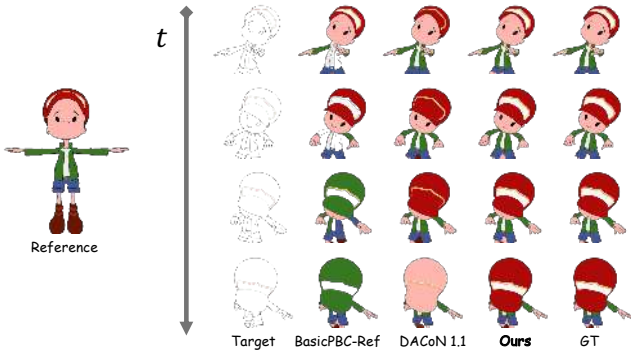



**Fig. S8: Visualisation of temporal stability.** Compared with previous methods (BasicPBC-Ref [5], *Base* DACoN 1.1 [16]), our method shows significantly fewer "colour flickers" over time. More qualitative results are provided in the accompanying video.

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 [15]. 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 maximum 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

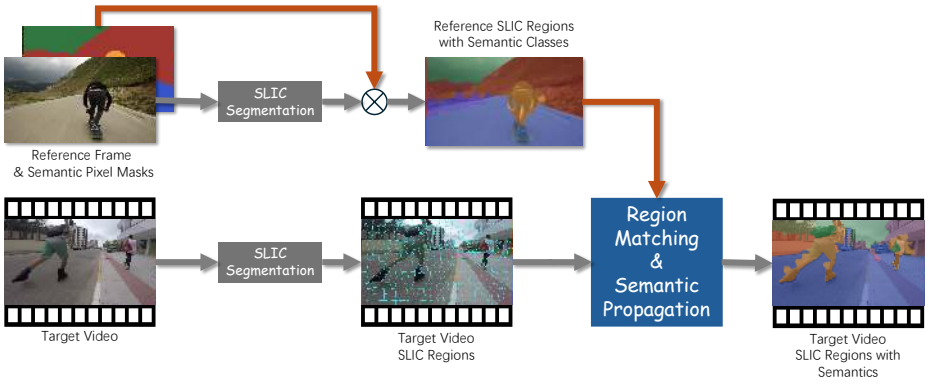



**Fig. S9: Task formulation for reference-guided region label propagation for natural videos.** Given an external reference frame with panoptic semantic masks, we compute SLIC superpixels and assign each reference superpixel a semantic class according to the original panoptic annotation. For the target 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.

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 [18] and SAM2.1-Large [20]. Metrics are computed per frame and averaged over all evaluation frames following Sec. B.1.

**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** |

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

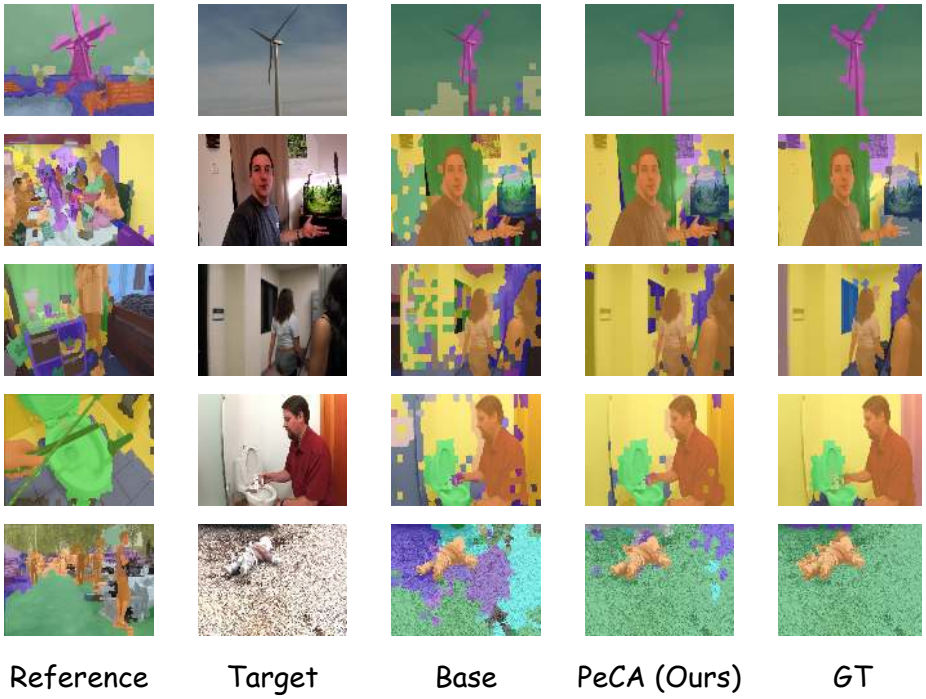


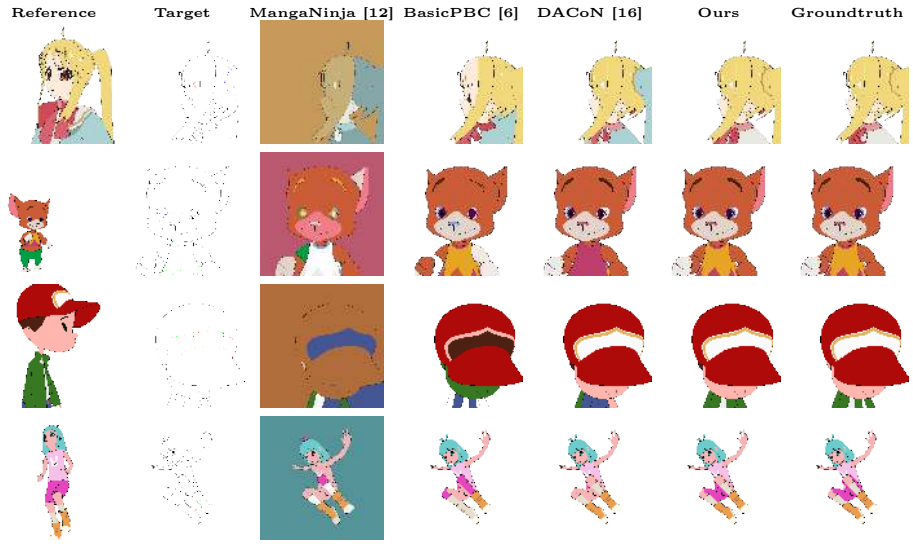

**Fig. S10: Qualitative results on VIPSeg superpixel region label propagation.** From left to right: first external reference frame with label map, target RGB frame (input), *Base*, PeCA, and ground truth. Compared with direct matching, PeCA produces more spatially coherent predictions with fewer fragmented labels and better object coverage. Ground truth denotes the superpixel semantic labels induced from the VIPSeg [15] annotations.

that PeCA captures a more general form of reference-guided region matching that transfers to natural videos.

## 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 [12], we follow the post-processing steps in previous work [16] to convert the raw results to palette-preserving paint-bucket colourisations. In general, our method shows superior performance under challenging test cases.


**(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** [12], **BasicPBC(-Ref)** [5,6], **DACoN 1.1** [16], our **PECA** on DACoN 1.1, and colour ground-truth (right).

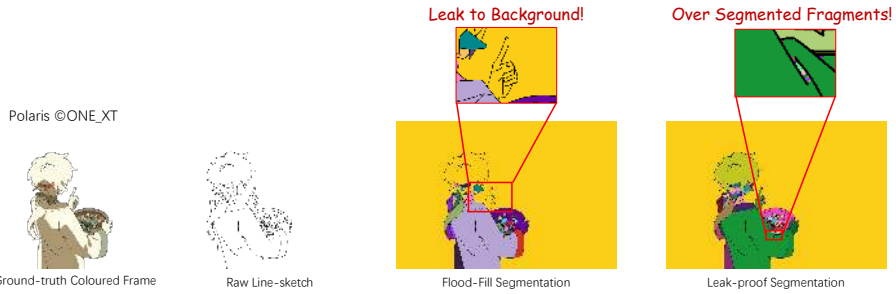



**Fig. S12: Segmentation failure modes on amateur-level line sketches from [1].** From left to right: coloured reference, raw sketch, standard flood-fill [24] segmentation, and leakage-robust segmentation. Standard flood-fill may leak into the background when strokes are not fully closed, while leakage-robust 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 [5, 6] and industry-level animation workflows [17] assume.

## H Limitations

Like previous paint-bucket colourisation methods [5–7, 16], PECA expects line sketches with sufficiently enclosed regions, so that simple flood-fill segmentation [24] can be applied. This assumption is consistent with standard animation production workflows [17, 26] 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 [22, 30] and leakage-robust region segmentation [3, 31]. 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 [9].

Another limitation comes from incomplete reference coverage. Like other reference-guided colourisation methods [5–7,16], 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 inexpensive in-plane spatial support, which further reduce moderate pose or layout gaps, but do not synthesize 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 ARE selects and reweights from the supplied reference pool, it can also benefit from expanded pools provided by artist interaction, retrieval, or generated novel-view references 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.

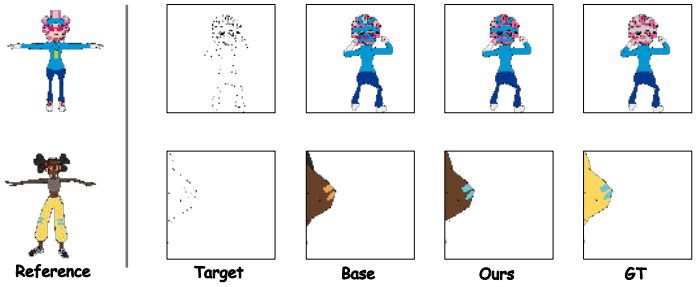



**Fig. S13: Failure cases caused by incomplete reference coverage.** From left to right: reference, target line sketch, *Base* and PeCA predictions, and ground truth. Top: missing reference coverage for the target pose leads to incorrect colour assignment on the helmet back regions. Bottom: a very small 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. 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)
5. 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)

6. 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)
7. 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)
8. 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]
9. Guajardo, J., Bursalioglu, O., Goldman, D.B.: Generative ai for 2d character animation. In: ACM SIGGRAPH 2024 Posters, pp. 1–2 (2024)
10. 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)
11. 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)
12. 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)
13. 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)
14. 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)
15. 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)
16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. 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)

22. 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)
23. 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
24. Smith, A.R.: Tint fill. In: Proceedings of the 6th Annual Conference on Computer Graphics and Interactive Techniques. p. 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
25. 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)
26. 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)
27. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
28. 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)
29. 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)
30. 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)
31. Zhang, L., JI, Y., Liu, C.: Danbooregion: An illustration region dataset. In: European Conference on Computer Vision (ECCV) (2020)
32. Zhang, Y., Ma, Y., Wang, B., Chen, Q., Wang, Z.: Follow-your-color: Multi-instance sketch colorization (2025), https://arxiv.org/abs/2503.16948
33. 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)
34. 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)
35. 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)