COLOR ANNOTATIONS IN TEXT-TO-IMAGE PRETRAINING: AN EMPIRICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Numerical color annotations provide a route to precise color control in text-to-image pretraining, but how their design affects generation with and without explicit color inputs remains unclear. We investigate this question by training diffusion transformers from scratch on a shared BLIP3-o subset, keeping the architecture and training hyperparameters fixed. We compare region-level palettes extracted from image pixels with object colors predicted jointly with color-enriched descriptions. With benchmark colors explicitly supplied, our pixel-annotated model reaches 86.97% accuracy on the evaluated GenColorBench numerical-color task. Its scores exceed published NumColor results in the three color categories compared, under the respective evaluation settings. Within our palette-trained configurations, pixel-extracted annotations give the strongest supplied-color fidelity, including prompts requiring multiple colors on multiple objects. This ranking reverses when palettes are withheld from matched prompts while descriptions, target colors, and scoring remain fixed. The pixel-annotated model also gains more in image quality from receiving palettes, whereas predicted annotations support stronger texture binding and compositional generation. These results demonstrate the effectiveness of color-annotated pretraining for precise numerical control. They also show that, in this comparison, the strongest control of supplied colors accompanies the greatest palette dependence and does not translate into the strongest color generation from descriptions.

1 INTRODUCTION

Design, branding, and illustration require specified colors to appear on the intended objects. Text-to-image models still struggle with numerical color requests and multi-object color binding (Butt et al., 2026a; Shomer-Chai et al., 2026). Adapters, guidance, and editing improve control after pretraining (Butt et al., 2024; Laria et al., 2026; Shum et al., 2025; Lobashev et al., 2025; Aharoni et al., 2026; Yin et al., 2026). We study an earlier choice: how color is represented in pretraining annotations, and how that choice relates to numerical control and color generation from language.

Structured captions can describe an object's color in words and attach a numerical palette to it (Ideogram AI, 2026). A palette extracted from the pixels of a bounding box summarizes surface color, illumination, shadows, and background. A palette predicted alongside the object's description instead aims to identify the object's own colors. Both annotations come from the training image, but one records pixel statistics directly while the other interprets them semantically. Does the annotation that produces the strongest numerical control also produce the strongest color generation when only the description is available?

We train diffusion transformers from scratch on a shared BLIP3-o subset under two annotation regimes: pixel-extracted palettes (PIXEL) and captioner-predicted palettes with color-enriched descriptions (PREDICTED). A third configuration (PREDICTED-LEAF) uses the predicted annotations with leaf-only palettes and removal before text encoding (Section 3). We evaluate numerical-color fidelity, paired generation with and without palettes, color generation from descriptions, and quality under palette perturbation (Sections 4–7). Figure 1 summarizes the central comparisons.

Three findings emerge. First, with benchmark colors supplied directly, PIXEL reaches 86.97% accuracy on the evaluated GenColorBench NCU prompts, with scores above the published NumColor references in the three color categories compared under the respective evaluation settings (Section 5.3).

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

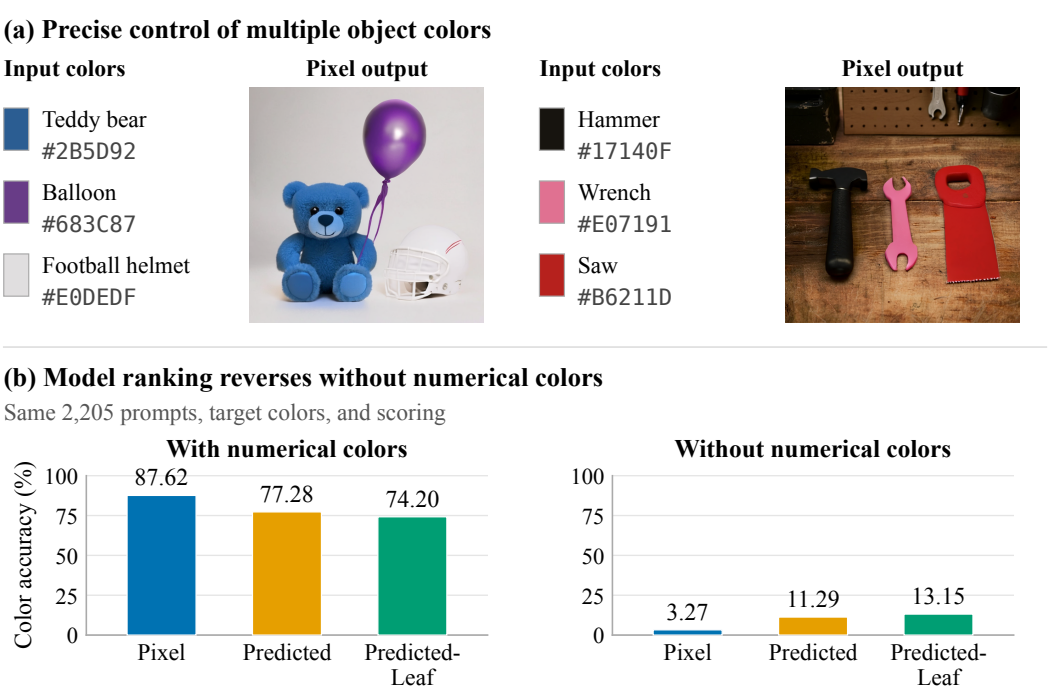

Figure 1: **Precise color control and the effect of annotation design.** (a) Supplied object colors and PIXEL outputs on two GenColorBench MOC examples; the corresponding three-model comparison is in Appendix E. (b) Target-independent numerical accuracy ($\Delta E \leq 5$) on the same 2,205 prompts and targets, with and without numerical annotations (color words retained). The accuracy ranking reverses within our training configurations.

Its supplied-color advantage within our configurations extends to multi-object requests. Second, removing palettes from matched prompts reverses the numerical-color ranking: PIXEL scores lowest without them and gains most in paired image quality from adding them. PREDICTED has stronger texture-binding and compositional scores. Third, PREDICTED does not show smaller average perturbation losses than PIXEL, whereas PREDICTED-LEAF has the smallest losses under training-consistent palette scopes, alongside lower numerical fidelity. The palette-removal and perturbation comparisons concern the three palette-trained configurations rather than external methods or isolated training factors.

2 RELATED WORK

Color control in text-to-image generation. Adapters (Butt et al., 2024; Aharoni et al., 2026), sampling guidance (Laria et al., 2026; Lobashev et al., 2025), and editing (Yin et al., 2024; 2026) improve color control after pretraining. Rich-text generation associates RGB colors with text spans (Ge et al., 2023), while NumColor learns embeddings for numerical color specifications (Butt et al., 2026b). Our focus is how color annotation during pretraining shapes both palette execution and generation without palettes.

Caption design for pretraining. Recaptioning changes the supervision available to a generator (Betker et al., 2023; Chen et al., 2024). Structured captions already include per-element boxes and HEX palettes (Ideogram AI, 2026), and palette–text–image datasets enable studies of palette representations (Qiu et al., 2025). We study the behavioral consequences of pixel-statistical and semantically predicted annotations on the same training images, using these established conditioning representations.

Reliance on conditioning. Condition dropout (Ho & Salimans, 2022) and noise-conditioning augmentation (Ho et al., 2022; Saharia et al., 2022) address missing or imperfect diffusion conditions.

Shortcut learning provides a broader perspective on dependence on predictive cues (Geirhos et al., 2020). Our palette interventions empirically characterize reliance on color conditioning.

Color evaluation. GenColorBench (Butt et al., 2026a), CompColor (Shomer-Chai et al., 2026), and ColorConceptBench (Ruan et al., 2026) evaluate color generation; ColorBench evaluates color understanding in vision-language models (Liang et al., 2025). We combine color-fidelity tests with PRISM-Bench's image-quality evaluation (Fang et al., 2026) to examine palette execution and reliance (Sections 5–7).

3 COLOR-ANNOTATED PRETRAINING

We compare two ways to annotate the same training images: direct pixel statistics (PIXEL) and jointly predicted object colors and descriptions (PREDICTED). The reduced-exposure configuration (PREDICTED-LEAF) uses PREDICTED annotations with a different palette scope and removal stage. Below we define the shared representation and the differences among these three configurations.

3.1 REGION-LEVEL COLOR CONDITIONS

A global palette lists colors without assigning them to objects, so we associate descriptions and palettes with regions in a hierarchical JSON annotation (Appendix A). The annotation combines scene-level information with region descriptions, spatial locations, parent–child relations, and palettes of sRGB HEX colors. Each region's description and palette jointly condition its appearance.

3.2 TWO ANNOTATION REGIMES

We annotate the BLIP3-o subset with an in-house image-to-JSON captioner: a Qwen3.5-4B multimodal model fine-tuned to emit the schema of Section 3.1 from the image alone. The two regimes use two variants of this captioner, which share the base model and fine-tuning recipe and differ only in their targets (Figure 2; Appendix B).

PIXEL. The captioner is fine-tuned on layout-only records and predicts the element inventory, boxes, and descriptions without color fields. Palettes are then filled by a rule: for each element we run k-means (k=5) on the RGB pixels inside its box and keep the cluster centers ranked by pixel share; the global palette is extracted the same way over the whole image with a larger number of clusters (Table 7). A PIXEL palette is therefore a deterministic function of the target pixels, $P = f(x_0)$, and summarizes whatever the box contains, including illumination, shadows, and background.

PREDICTED. The captioner is fine-tuned on the same records after two additional steps: rule-extracted palettes are attached as in PIXEL, and a commercial multimodal model (Gemini 3.5 Flash-Lite; Google DeepMind, 2026) revises each of the 6M fine-tuning records so that palettes name the object's own color rather than the box's pixel statistics and descriptions mention that color. The resulting captioner predicts descriptions and palettes jointly with the layout, $\tilde{P} = h(x_0)$; Gemini is not run on the training set. PREDICTED palettes are model predictions of object color, typically one to three entries.

Each captioner predicts its own element inventory and boxes, so the two annotation sets need not share a layout. On 500 held-out training images, mean element-palette length drops from 4.98 to 1.86, and the share of descriptions containing a color word rises from 59.6% to 82.8% (Appendix B.2; Table 7).

3.3 THREE TRAINING CONFIGURATIONS

For an image x_0, PIXEL conditions the denoiser $v_\theta(x_t, t, c_{\text{pix}}, P)$ on the descriptions c_{pix} and the palette $P = f(x_0)$, and PREDICTED conditions it on c_{pred} and $\tilde{P} = h(x_0)$; both enter through the text condition (Section 3.4). PIXEL and PREDICTED train the same denoiser on the same subset under their respective annotations with identical hyperparameters and identical color-field dropout.

PREDICTED-LEAF trains on the PREDICTED annotations with lower exposure to palettes. Only leaf elements carry palettes: the global palette and every non-leaf element palette are removed from the JSON before it reaches the text encoder. Palettes selected by stochastic dropout are removed from the JSON text in the same way, rather than masked after encoding as in PIXEL and PREDICTED, so the

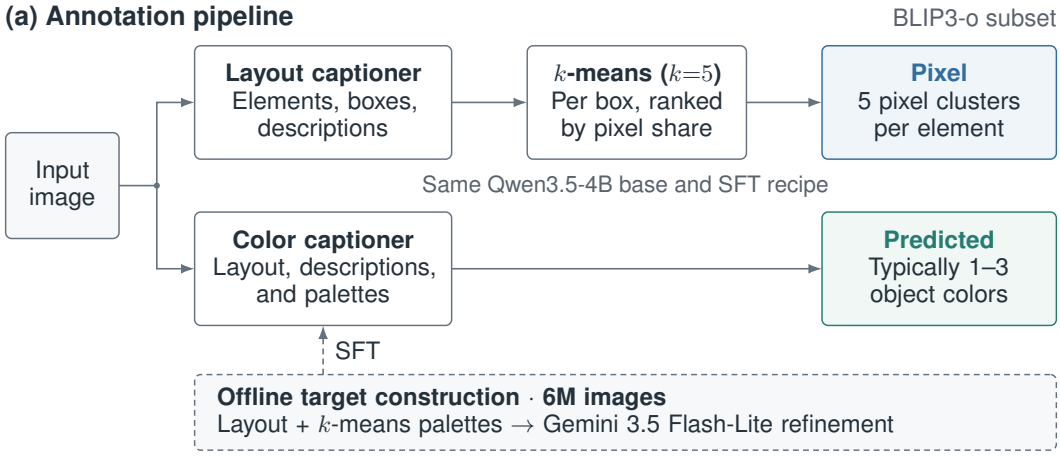

Figure 2: **Annotation regimes.** (a) Two captioners annotate the same BLIP3-o subset using a shared Qwen3.5-4B base and SFT recipe, with different targets. PIXEL adds pixel-cluster palettes to predicted layouts and descriptions; PREDICTED predicts layout, descriptions, and palettes jointly. Gemini is used only to construct the color captioner's 6M fine-tuning targets (dashed path). (b) An element-aligned example with an identical box in both records; layouts can differ in general. Shared description words are gray; differing words are bold and branch-colored.

Table 1: **Training configurations.** All three share the denoiser, the training subset, and the dropout probabilities of Appendix C.2.

	PIXEL	PREDICTED	PREDICTED-LEAF
Captioner	layout-only	end-to-end color	end-to-end color
Element palette	pixel k-means, 5 entries	predicted, 1–3 entries	as PREDICTED, leaf elements only
Descriptions	color word in 59.6%	color word in 82.8%	as PREDICTED
Dropped palettes	masked after encoding	masked after encoding	removed from the text

encoder never sees a dropped palette. Dropout probabilities are unchanged. Table 1 summarizes the three configurations.

3.4 COLOR-FIELD DROPOUT AND INJECTION

All palette-trained models share the stochastic color-field dropout rates and classifier-free guidance dropout (Ho & Salimans, 2022); palette scope and the stage at which palettes are removed differ as described above (Appendix C.2). Textual conditioning and region locations are supplied to an

in-house 24-layer single-stream diffusion transformer (Peebles & Xie, 2023). All three configurations train from scratch with the same recipe on a 13M-image subset of BLIP3-o (Chen et al., 2025) at 512×512 (Appendix C).

4 Experimental Setup

We evaluate numerical-color fidelity and generation from color descriptions on GenColorBench, attribute binding and composition on T2I-CompBench++, and image quality with missing or altered palettes on PRISM (Fang et al., 2026). PRISM's official automated judge scores Aesthetic and Alignment; Combined is their mean.

A prompt-expansion (PE) language model converts benchmark prompts to structured JSON. When palettes are supplied, numerical-color and compositional comparisons use the same all-scope PE JSON across our configurations. The NoColor baseline shares the architecture, training subset, and training settings, omitting color fields during training. We use 40k exponential-moving-average (EMA) checkpoints for diagnostic comparisons and 90k EMA for official NCU and T2I-CompBench++ comparisons. Full protocols appear in Appendices C–D.

5 Color Fidelity and Object Appearance

GenColorBench tests one numerical color per object (NCU) and multiple colors and objects (MOC). We first evaluate fidelity to supplied colors, then examine attribute binding and appearance.

5.1 A target-independent protocol for color fidelity

The official NCU scorer guides pixel selection with the target and permits color-neighborhood fallback. Our target-independent protocol segments the object with GroundingDINO (Liu et al., 2024) and SAM2 (Ravi et al., 2024), extracts its dominant color by clustering its pixels, and compares it with the PE-declared HEX using CIEDE2000 (Sharma et al., 2005), without target-guided selection or fallback. Accuracy requires $\Delta E \leq 5$. For MOC, one-to-one palette matching must place every target color within this threshold; image accuracy requires every object to pass. Localization and matching failures count as incorrect (Appendix C.4).

5.2 Color fidelity on NCU and MOC

Pixel leads both benchmarks under common Pixel-format inputs and the target-independent scorer (Table 2). Its NCU accuracy is 86.95%, against 77.85% for Predicted and 73.16% for Predicted-Leaf. The advantage extends to simultaneous object–color binding: requiring every object to pass gives image accuracies of 72.84%, 57.53%, and 50.69%, respectively.

Table 2: **Color fidelity on NCU and MOC** under the target-independent protocol. Bold marks the best value per column.

	NCU		MOC		
Configuration	Acc. ↑	Mean ΔE ↓	Object acc. ↑	Image acc. ↑	Mean ΔE ↓
NoColor	14.78	13.134	23.68	5.72	7.811
Pixel	**86.95**	**4.186**	**86.34**	**72.84**	**2.512**
Predicted	77.85	4.988	77.14	57.53	3.196
Predicted-Leaf	73.16	5.681	71.95	50.69	3.686

5.3 Official benchmarks and external comparisons

Official numerical-color results. With the official NCU scorer, **A** evaluates PE-JSON generations against benchmark colors, **B** scores the same images against PE-declared colors, and **C** injects benchmark colors into the JSON before generation and scoring.

Table 3: **NCU accuracy and external references** (%, ↑). Published results retain NumColor's original column labels (Butt et al., 2026b); inputs and evaluation settings differ from ours. Bold compares our configurations within A or C. Full A/B/C results: Table 11.

Method	Input / protocol	CSS3/X11	ISCC-L1	ISCC-L3
Published reference results				
FLUX	prose	5.9	13.8	7.7
FLUX + NumColor	parsed color	52.0	55.7	48.0
SD3	prose	4.6	11.8	5.9
SD3 + NumColor	parsed color	47.2	49.2	39.5
SD3.5	prose	5.8	14.9	7.5
SD3.5 + NumColor	parsed color	46.9	51.7	43.4
PixArt-α	prose	4.2	9.5	5.4
PixArt-α + NumColor	parsed color	38.5	44.9	36.2
Our configurations		HEX-CSS	RGB-L1	RGB-L3
PIXEL	PE JSON (A)	**89.26**	43.23	39.22
PREDICTED	PE JSON (A)	86.89	**47.38**	**45.07**
PREDICTED-LEAF	PE JSON (A)	82.48	46.92	44.52
PIXEL	injected JSON (C)	**93.07**	**85.08**	**85.30**
PREDICTED	injected JSON (C)	91.52	82.46	83.15
PREDICTED-LEAF	injected JSON (C)	84.67	75.69	73.56

With benchmark colors supplied directly (protocol C), PIXEL reaches 86.97% pooled accuracy across the four evaluated NCU subsets (Table 11). Its scores exceed the published NumColor reference results in all three color categories compared in Table 3, where FLUX + NumColor is the strongest published baseline (Butt et al., 2026b). These are reported scores under the respective evaluation settings; the table preserves the source's category labels.

Within our configurations, PIXEL also leads when scored against PE-declared colors (B), at 92.47%. Under the full prompt-conversion pipeline (A), PREDICTED leads (55.24% vs. 52.38%). PE can replace unusual requested colors with canonical ones, making requested and received targets differ. The A/B difference does not isolate PE error, since the official scorer also uses the target during pixel selection (Appendix C.5).

Numerical precision and attribute binding favor different configurations. With identical PE JSON, PREDICTED and PREDICTED-LEAF have higher eight-category T2I-CompBench++ (Huang et al., 2025a) means than PIXEL (0.561 and 0.560 vs. 0.538; Table 4). Among our configurations, PREDICTED leads texture binding and Color; scoring the Color prompts against PE-declared colors or generating from color-injected JSON preserves this ordering (Table 16). Relative to the published OSPO-7B, PREDICTED is higher on 2D relations and numeracy and lower on color, shape, texture, and complex composition (Table 4).

In Figure 3's texture example (second row), PIXEL shows less visible fabric detail than PREDICTED and PREDICTED-LEAF for the same PE JSON and seed. We next compare palette dependence within our three training configurations.

6 PALETTE DEPENDENCE AND COLOR GENERATION FROM LANGUAGE

We first measure paired quality gains from adding palettes to the same 700 PIXEL-format PRISM prompts, retaining description color words, then test color generation without palettes. With-palette inputs follow each model's training scope: all-scope for PIXEL/PREDICTED and leaf-only for PREDICTED-LEAF. Table 5 summarizes the paired gains alongside the perturbation results of Section 7.

PIXEL gains most from the palette. Adding the original palette raises PRISM Combined for all three configurations. PIXEL rises from 6.51 to 7.15, a gain of 0.63, against 0.29 for both PREDICTED and PREDICTED-LEAF. Paired 95% intervals support positive gains for all three and a larger gain for

Table 4: **T2I-CompBench++ scores and external references.** Published references use their original inputs and settings (Huang et al., 2025b; Oh et al., 2026). Our configurations share PE JSON (Color: protocol A). Bold compares our configurations; the 8-category mean is unweighted. Complete results: Table 15.

				Spatial			Non-		8-cat.
Method	Color	Shape	Texture	2D	3D	Numeracy	spatial	Complex	mean
Published reference results									
SDXL	0.588	0.469	0.530	0.213	0.357	0.499	0.312	0.324	0.411
DALL-E 3	0.779	0.621	0.704	0.287	0.374	0.588	0.300	0.377	0.504
OSPO (7B)	0.857	0.639	0.773	0.356	0.393	0.591	0.316	0.415	0.542
Our configurations; PE JSON inputs									
PIXEL	0.800	0.584	0.693	0.432	0.381	0.715	0.313	0.389	0.538
PREDICTED	**0.825**	**0.615**	**0.736**	0.454	0.396	0.747	0.313	**0.405**	**0.561**
PREDICTED-LEAF	0.817	0.606	0.724	**0.455**	**0.402**	**0.763**	0.313	0.400	0.560

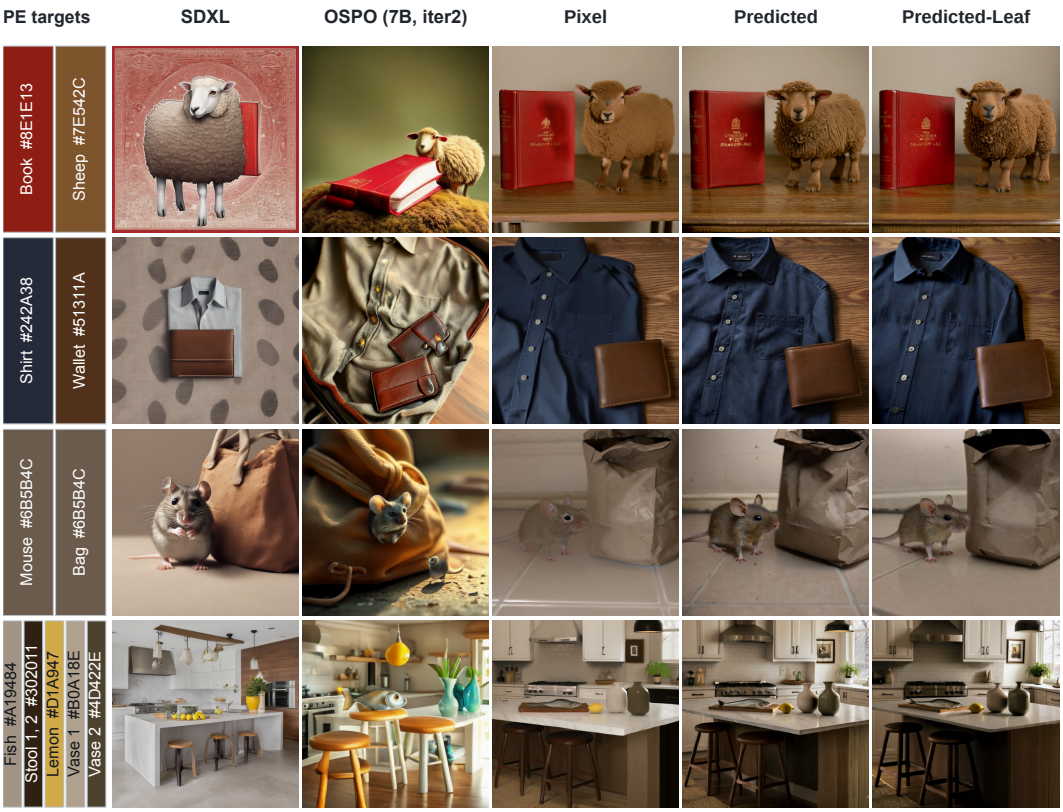

Figure 3: **Compositional generation on T2I-CompBench++.** Rows show color, texture, 2D spatial, and numeracy prompts (top to bottom). SDXL (Podell et al., 2024) and OSPO (Oh et al., 2026) receive the original text prompts; our three configurations share the same PE JSON. Color bands show our PE-declared targets. Prompts and selection protocol appear in Appendix E.4.

PIXEL than either alternative (Appendix C.8). Among these palette-trained models, PIXEL scores lowest without palettes but nearly matches PREDICTED's with-palette Aesthetic score (Table 17). Figure 4 shows changes in illumination, surface appearance, and color balance.

The numerical ranking reverses on matched prompts. We pair 2,205 NCU prompts with and without numerical annotations, retaining color words and using the same target HEX, scorer, and generation seed (Table 6). PIXEL leads with annotations and scores lowest without them. Its accuracy

Table 5: **Palette contribution and perturbation sensitivity** on PRISM. Mean loss averages ten perturbations with description color words removed (Table 25). Gains and losses use unrounded means; bold marks the highest scores and gain and the lowest loss.

	Palette contribution: Combined ↑			Perturbations: Aesthetic		
Configuration	Without	With	Gain ↑	Original ↑	Mean loss ↓	Cross-image ↑
PIXEL	6.51	7.15	**0.63**	6.63	1.80	4.31
PREDICTED	**6.93**	**7.22**	0.29	**6.68**	1.87	4.47
PREDICTED-LEAF	6.86	7.15	0.29	6.44	**1.54**	**4.71**

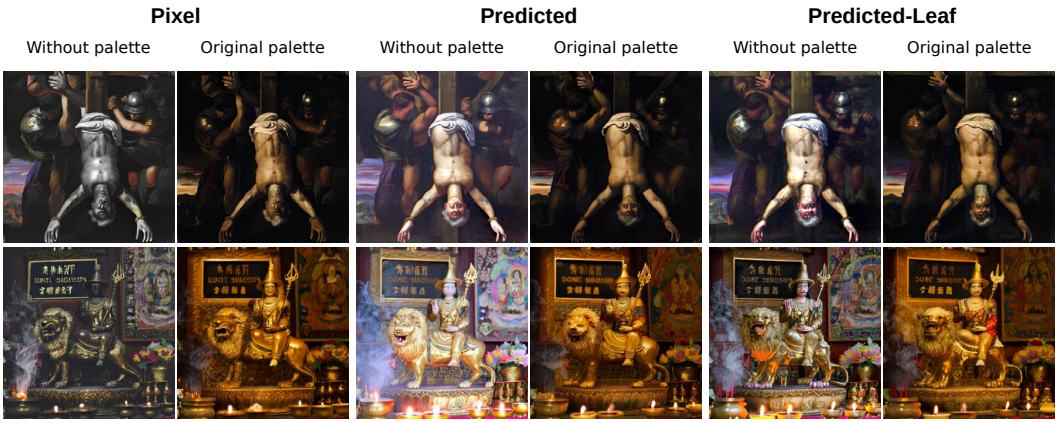

Figure 4: **Generation without and with the original palette.** Rows show a classical painting and a gilded statue; each configuration has adjacent without/with-palette outputs.

gain exceeds PREDICTED's by 18.37 percentage points (95% CI $[16.19, 20.59]$) and PREDICTED-LEAF's by 23.31 ($[21.04, 25.62]$). The ordering persists for 2,119 prompts differing only in palette fields (Appendix C.10). This measures recovery of a fixed numerical color, which the remaining description need not uniquely specify.

Color-word scores favor predicted annotations. On the same no-palette outputs, we compare the object's ISCC-NBS (Kelly & Judd, 1976) Level-2 class with classes assigned to the description's color phrase. PREDICTED and PREDICTED-LEAF exceed PIXEL by 8.25 $[5.94, 10.48]$ and 10.88 $[8.62, 13.11]$ points. Both advantages persist on 767 simple-word targets. These classes approximate phrase meanings through fixed naming anchors; they are not independently validated semantic labels (Appendix C.10). The NOCOLOR baseline scores 40.59%. Figure 8 gives examples; photographic color matching remains exploratory in the appendix.

7 QUALITY UNDER PALETTE PERTURBATIONS

The paired gains above concern the original palettes. We now test quality when those palettes change, using the same 700 PIXEL-format PRISM prompts with description color words removed. Inference retains each configuration's training palette scope: all-scope for PIXEL/PREDICTED and leaf-only for PREDICTED-LEAF (Table 5).

Changing the palette lowers quality. Replacing 20% of the palette entries with random colors lowers Aesthetic by 3.07 for PIXEL, 3.15 for PREDICTED, and 2.68 for PREDICTED-LEAF; substituting another image's palette lowers it by 2.32, 2.22, and 1.73 (Table 25; Figure 5). In every configuration, replacement causes the largest loss, followed by cross-image substitution, hue rotation, and lightness-order rearrangement. Rearrangement changes no HEX value, so the models respond to how the palette is serialized as well as to which colors it contains.

Table 6: **Color generation on matched prompts.** Numerical accuracy uses fixed PE-declared targets and the target-independent scorer; L2 evaluates color-word classes on outputs without numerical annotations. Accuracies are percentages; gain is with minus without (percentage points). Bold marks the largest value.

	Numerical acc@5			Color-word L2
Configuration	With ↑	Without ↑	Gain	Without ↑
PIXEL	**87.62**	3.27	**84.35**	30.98
PREDICTED	77.28	11.29	65.99	39.23
PREDICTED-LEAF	74.20	**13.15**	61.04	**41.86**

PREDICTED-LEAF has the smallest perturbation losses. PREDICTED's texture and appearance advantages do not come with smaller perturbation losses: its mean Aesthetic loss over the ten conditions is 1.87, against 1.80 for PIXEL. PREDICTED-LEAF has the smallest loss, 1.54, which is 0.33 below PREDICTED's (95% prompt-bootstrap interval [0.20, 0.46]; Table 26), and the highest mean perturbed Aesthetic score (4.91 vs. 4.84 and 4.82; original score minus mean loss in Table 5), while it has the lowest color fidelity in Table 2. This comparison establishes the quality sensitivity of the complete configurations under their training-consistent palette scopes; it does not isolate the effect of reduced training exposure or measure adherence to the perturbed colors.

8 DISCUSSION AND LIMITATIONS

Implications for annotation design. Strong supplied-palette accuracy alone does not establish reliable color generation from descriptions or stability under palette changes. The ranking reversal under exchanged annotation formats (Appendix D.4) further shows that annotation strategies must be assessed together with their intended inference format. Our results favor pixel-extracted palettes for precise numerical requests, while predicted annotations offer stronger generation from descriptions than pixel-extracted annotations. Evaluating both input conditions exposes this trade-off, which a supplied-palette benchmark alone would miss.

Palette removal alone does not imply deficient color generation: the description may not uniquely specify the withheld HEX value. The color-word evaluation tests agreement with the remaining description on the same generated images. Predicted annotations outperform pixel-extracted annotations under both no-palette criteria, so their advantage extends beyond matching an exact numerical target.

Limitations. The study uses one architecture and training subset, with one training run per configuration. The three palette-trained configurations jointly vary palette source, descriptions, layout, removal stage, and image exposure. Perturbation tests also use configuration-specific palette scopes, so their differences do not isolate the effect of training exposure. PRISM uses an automated judge; color-word targets use fixed naming anchors, and a description need not uniquely specify a withheld HEX value. Qualitative examples are selected, and the photographic-color test is exploratory. External comparisons use the published systems' respective settings. Protocol details and sensitivity analyses appear in Appendices C–D.

9 CONCLUSION

We studied how color annotations in text-to-image pretraining affect numerical control and generation from language. Pixel-extracted palettes deliver the strongest numerical control among our configurations and exceed published NumColor scores in the three compared categories under their respective evaluation settings. Among the three palette-trained models, this advantage reverses when palettes are withheld from matched prompts, while predicted annotations support stronger color-word generation, texture binding, and composition. The pixel-annotated model also gains most in image quality from receiving palettes. Color-annotated pretraining thus enables precise color control, while our comparisons show that the annotation strategy with the strongest supplied-color fidelity also relies most on explicit color inputs.

REPRODUCIBILITY STATEMENT

Appendix A summarizes the structured annotations. Appendix B documents how the two captioners were fine-tuned, the paired audit of their outputs, and the construction of the PRISM prompt sets. Appendix C gives a model overview, configuration differences, and evaluation protocols; Appendices D and C.10 provide the complete palette perturbation and language-conditioned color results. The supplementary files provide the paired PRISM annotation formats, the earlier PIXEL-format set, and the palette-perturbation inputs, with integrity and construction checks. Per-prompt scores or counts and analysis scripts reproduce the paired comparisons for MOC accuracy, numerical-annotation removal, PRISM palette contribution, palette perturbations, and annotation-format swaps. Anonymous color-word and photographic-color records support score aggregation and paired uncertainty; the color-word anchor mapping is also supplied. Original color-evaluation implementations and the 500 paired captioner audit records are not included in the current package. The training images are drawn from the public BLIP3-o collection (Chen et al., 2025). The pretraining annotations, the trained weights (including the NOCOLOR baseline and the STAGED run of Appendix C.5), the two in-house captioners, and the prompt-expansion model are not released, so results that depend on them cannot be fully regenerated from the release. Changes in the commercial services used for annotation, prompt construction, and scoring may also affect reproduction.

ETHICS STATEMENT

No human study was conducted. The photographic color targets use the Object Color Statistics dataset's existing human-produced foreground masks (Appendix C.10). We use a subset of the public BLIP3-o collection (Chen et al., 2025). Captioner fine-tuning targets, evaluation-prompt construction, and PRISM scoring use commercial services under the providers' terms. Official GenColorBench prompts and Object Color Statistics photographs are not redistributed. Precise color control serves design, branding, and illustration; we identify no risks specific to color annotation beyond those of text-to-image generation.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, specifically to assist in writing and debugging the code for our experiments. Additionally, we used generative AI tools to suggest a structure for the paper and to edit the manuscript to improve readability, grammar, and clarity. We have reviewed all AI-assisted work. All AI-generated code was read, tested, and verified for correctness by the authors before use. All AI-suggested edits to the text were reviewed by the authors.

REFERENCES

Elad Aharoni, Noy Porat, Dani Lischinski, and Ariel Shamir. Palette aligned image diffusion. *Computer Graphics Forum*, pp. e70384, 2026.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. URL `https://cdn.openai.com/papers/dall-e-3.pdf`.

Muhammad Atif Butt, Kai Wang, Javier Vazquez-Corral, and Joost van de Weijer. ColorPeel: Color prompt learning with diffusion models via color and shape disentanglement. In *European Conference on Computer Vision (ECCV)*, pp. 456–472, 2024.

Muhammad Atif Butt, Alexandra Gomez-Villa, Tao Wu, Javier Vazquez-Corral, Joost van de Weijer, and Kai Wang. GenColorBench: A color evaluation benchmark for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36638–36648, 2026a.

Muhammad Atif Butt, Diego Hernandez, Alexandra Gomez-Villa, Kai Wang, Javier Vazquez-Corral, and Joost van de Weijer. NumColor: Precise numeric color control in text-to-image generation. *arXiv preprint arXiv:2603.13547*, 2026b.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models – architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-Reason-6M & PRISM-Bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. In *International Conference on Learning Representations (ICLR)*, pp. 63157–63186, 2026.

Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7545–7556, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Google DeepMind. Gemini 3.5 Flash-Lite. Model card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/`.

Xiefan Guo, Xinzhu Ma, Haiyu Zhang, and Di Huang. CTCal: Rethinking text-to-image diffusion models via cross-timestep self-calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43558–43567, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025a.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: Official project page, 2025b. URL `https://karine-h.github.io/T2I-CompBench-new/`. Benchmarking results table, accessed 2026-09-26.

Ideogram AI. Ideogram 4.0 technical details: Open model at the forefront of design. Technical blog, 2026. URL `https://ideogram.ai/blog/ideogram-4.0/`.

Kenneth L. Kelly and Deane B. Judd. Color: Universal language and dictionary of names. NBS Special Publication 440, National Bureau of Standards, 1976.

Héctor Laria, Alexandra Gomez-Villa, Jiang Qin, Muhammad Atif Butt, Bogdan Raducanu, Javier Vazquez-Corral, Joost van de Weijer, and Kai Wang. Leveraging semantic attribute binding for free-lunch color control in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7689–7698, 2026.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. ColorBench: Can VLMs see and understand the colorful world? A comprehensive benchmark for color perception, reasoning, and robustness. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Alexander Lobashev, Maria Larchenko, and Dmitry Guskov. Color conditional generation with sliced Wasserstein guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Yoonjin Oh, Yongjin Kim, Hyomin Kim, Donghwan Chi, and Sungwoong Kim. OSPO: Object-centric self-improving preference optimization for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7620–7631, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Qianru Qiu, Jiafeng Mao, and Xueting Wang. Exploring palette based color guidance in diffusion models. In *ACM International Conference on Multimedia (ACM MM)*, pp. 10287–10295, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Chenxi Ruan, Yihan Hou, Yu Xiao, Guosheng Hu, and Wei Zeng. ColorConceptBench: A benchmark for probabilistic color-concept understanding in text-to-image models. *arXiv preprint arXiv:2601.16836*, 2026.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

Shay Shomer-Chai, Wenxuan Peng, Bharath Hariharan, and Hadar Averbuch-Elor. Color Bind: Exploring color perception in text-to-image models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1916–1925, 2026.

Ka Chun Shum, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Color alignment in diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28446–28455, 2025.

Xingxi Yin, Zhi Li, Jingfeng Zhang, Chenglin Li, and Yin Zhang. ColorEdit: Training-free image-guided color editing with diffusion model. *arXiv preprint arXiv:2411.10232*, 2024.

Zixin Yin, Xili Dai, Ling-Hao Chen, Deyu Zhou, Jianan Wang, Duomin Wang, Gang Yu, Lionel Ni, Lei Zhang, and Heung-Yeung Shum. Training-free text-guided color editing with multi-modal diffusion transformer. In *International Conference on Learning Representations (ICLR)*, pp. 54707–54737, 2026.

A STRUCTURED COLOR ANNOTATIONS

Annotations are hierarchical JSON records combining a scene description with region-level information. Each region associates a semantic description and spatial location with an optional palette of sRGB HEX colors; parent–child relations represent objects and their parts. Global information describes the scene and its overall appearance.

All-scope inputs include global and region palettes. Leaf-only inputs retain palettes only on regions without children. Removing palettes preserves the descriptions, spatial layout, and hierarchy. The PIXEL and PREDICTED annotation regimes differ in the source of their palettes and color descriptions (Section 3.2); Figure 2(b) illustrates these differences on a shared region.

B CAPTIONERS, ANNOTATION STATISTICS, AND PRISM PROMPT SETS

B.1 CAPTIONER FINE-TUNING

Both captioners start from Qwen3.5-4B and are fine-tuned with the same recipe on records from 6M source images produced by our annotation pipeline. The PIXEL captioner's targets are the layout-only records. The PREDICTED captioner's targets were obtained by attaching rule-extracted palettes to the same records and prompting Gemini 3.5 Flash-Lite (Google DeepMind, 2026) with the image and record to (i) replace each element palette with the one to three colors of the object itself, (ii) rewrite the element description to name those colors, and (iii) leave the element inventory and boxes unchanged. Gemini outputs serve only as fine-tuning targets; all images in the BLIP3-o subset are annotated by the fine-tuned captioners. At annotation time both captioners receive the image alone.

B.2 PAIRED ANNOTATION STATISTICS

Table 7 compares the captioners' outputs on 500 held-out training images. Because the captioners predict element inventories and boxes independently, statistics are computed over all elements of each output. PREDICTED produces shorter element palettes and more descriptions containing color words, while global-palette length remains similar.

Table 7: **Annotation statistics on 500 held-out training images and on the 700 PRISM prompt pairs.** Color-word presence counts lexical matches (Appendix B.2).

Statistic	PIXEL	PREDICTED
Held-out training images: 500 pairs, two captioners		
Elements	6,262	6,229
Mean element-palette length	4.98	1.86
Mean global-palette length	14.64	14.56
Element descriptions with a color word (%)	59.6	82.8
PRISM pairs: 700 prompts, 19,452 aligned elements		
Mean element-palette length	5.00	1.73
Mean global-palette length	16.00	16.00
Element descriptions with a color word (%)	88.6	95.5
Element palette entries retained from the PIXEL format (%)	—	34.6

Lexical measurement. Color-word presence uses case-insensitive matching against a fixed color lexicon, excluding matches adjacent to word characters or hyphens. The statistic measures color vocabulary rather than color-naming accuracy; it does not disambiguate object nouns such as *orange* or test correspondence to palette entries.

B.3 PRISM PROMPT SETS

PRISM pairs. Each of the 700 PRISM prompts (Fang et al., 2026) (100 from each of seven tracks) was converted into a structured layout with the Reve 2.1 text-to-layout API and rewritten by GPT-5.6 into the two regimes' formats. The PIXEL-format version adds a 16-entry global palette and, for each element, five distinct colors covering the base, subject, and transition tones, leaving descriptions unchanged. The PREDICTED-format version keeps one to three subject colors from the PIXEL-format

version and rewrites each element description to name the retained colors in English. Element inventories, boxes, and other fields are held fixed within each pair, and every PREDICTED-format palette is a subset of its PIXEL-format counterpart. Table 7 summarizes the resulting annotations. The pairs are used by the annotation-format swap (Appendix D.4).

PRISM (PIXEL format). The palette contribution, palette perturbation, and general generation experiments use an earlier structured conversion of the same 700 prompts in the PIXEL format: a 16-entry global palette and three to five colors per element (14,319 elements). Its element inventories differ from those of the pairs, so scores on the two sets are not comparable. Both sets contain palettes written by a language model rather than extracted from pixels. Here, PIXEL-format denotes a match to the training annotation format, not to its pixel-extracted source (Section 8).

C MODEL OVERVIEW, EVALUATION PROTOCOLS, AND DETAILED RESULTS

This appendix summarizes the shared model and training configurations, followed by evaluation protocols and detailed results. Appendix D describes palette perturbations and annotation-regime swaps; Appendix E documents qualitative sample selection.

C.1 MODEL AND TRAINING

We use an in-house single-stream diffusion transformer with 24 layers and joint attention over text and image tokens (Peebles & Xie, 2023). A frozen text encoder provides conditioning, and generation operates in the latent space of a frozen image autoencoder using flow matching (Lipman et al., 2023; Liu et al., 2023). The three palette-trained configurations and NOCOLOR share the architecture, training recipe, and BLIP3-o subset. The palette-trained configurations differ in annotation regime, palette scope, and removal stage (Table 1); NOCOLOR omits color fields from the training inputs. Each configuration is one training run.

Training exposure. Training uses a fixed token budget per step, so shorter annotations allow more images per step. Estimates from 852 paired held-out annotations indicate about 5% and 9% more images per step for PREDICTED and PREDICTED-LEAF than PIXEL. Their 40k checkpoints therefore correspond approximately to 42k and 44k PIXEL steps. The 30k–50k sweeps retain PIXEL's numerical-color advantage and lower no-palette PRISM score (Tables 8, 9, and 19). These are exposure estimates, not matched-exposure training runs; the remaining probes use one checkpoint per configuration.

C.2 PALETTE SCOPE AND DROPOUT

Scope projection precedes stochastic dropout: PIXEL and PREDICTED retain global and region palettes, while PREDICTED-LEAF retains only leaf-region palettes. All three use the same stochastic rates: a clean-anchor probability of 0.1, sample-level and per-element palette dropout of 0.5, and whole-caption dropout of 0.1. Description, style, and area-conditioned dropout are also shared. The clean anchor bypasses stochastic field dropout without changing the configuration's palette scope; whole-caption dropout applies independently.

PIXEL and PREDICTED mask dropped palette tokens after text encoding; PREDICTED-LEAF removes selected palette fields before encoding. The STAGED configuration and NOCOLOR also use text-level palette removal. No-palette evaluation preserves descriptions, including color words, while removing palettes.

C.3 SHARED COLOR CONVERSION

Color errors use CIEDE2000 (Sharma et al., 2005) in CIELAB. RGB values are normalized to $[0, 1]$, linearized with the standard inverse sRGB transfer function, and converted through XYZ with D65 white $(0.95047, 1, 1.08883)$. The following sections specify extraction, scoring targets, and failure denominators for each protocol.

C.4 TARGET-INDEPENDENT EVALUATION: NCU AND MOC

Tables 8–10 report 30k, 40k, and 50k EMA results, pooling four generated images per prompt for each frozen prompt set. The 40k rows supply Table 2. NCU uses protocol B: generation receives the PE JSON and scoring uses its declared object color. These results use the target-independent extraction below, distinct from the official scorer in Section C.5.

NCU: dominant-color accuracy. The official extractor uses the target to select pixels and permits color-neighborhood fallback. Our complementary protocol extracts the largest CIELAB K-means cluster without consulting the target, then compares its center with the exact target HEX. Each checkpoint/configuration has 2,255 prompts and 9,020 images. GroundingDINO (threshold 0.35) and SAM2 localize the object. Negative masks exceeding 95% of its area are rejected; subtraction is undone if less than 30% remains, followed by two-pixel erosion. K-means uses $K = 5$, at most 20,000 fitting pixels, and random state 0. There is no VQA gate or color fallback. Acc@τ requires a valid prediction with $\Delta E_{00} \leq \tau$; failures remain in the denominator, while mean and median errors use valid predictions only.

Table 8: **NCU color fidelity across EMA checkpoints.** Rates and accuracies are percentages; color errors are CIEDE2000. Bold marks the best metric within each step.

Step	Configuration	Mask ↑	Acc@5 ↑	Acc@10 ↑	Acc@15 ↑	Mean ↓	Median ↓
30k	NOCOLOR	99.89	13.66	45.76	69.56	13.363	10.684
	PIXEL	99.91	**87.57**	**92.42**	**94.02**	**4.067**	**1.592**
	PREDICTED	**99.98**	76.67	86.92	91.47	5.212	2.069
	PREDICTED-LEAF	99.88	72.87	86.74	91.61	5.589	2.667
40k	NOCOLOR	99.96	14.78	46.94	70.55	13.134	10.508
	PIXEL	99.92	**86.95**	**91.87**	**93.59**	**4.186**	**1.476**
	PREDICTED	99.96	77.85	87.30	91.46	4.988	1.848
	PREDICTED-LEAF	**99.97**	73.16	85.89	90.76	5.681	2.589
50k	NOCOLOR	99.94	13.46	45.14	70.44	13.313	10.783
	PIXEL	99.92	**87.93**	**92.88**	**94.16**	**3.965**	**1.410**
	PREDICTED	**99.98**	77.66	87.04	91.33	5.014	1.769
	PREDICTED-LEAF	**99.98**	72.26	84.77	89.94	5.930	2.544

MOC: localization and accuracy. Each checkpoint/configuration uses 50 sampling steps and four 512×512 images per prompt, covering 1,298 prompts, 5,192 images, and 12,080 target objects. GroundingDINO (threshold 0.35) and SAM2 produce masks, matched to the frozen PE element boxes at IoU ≥ 0.30 with one-to-one Hungarian assignment for repeated classes. CIELAB K-means ($K = 5$) extracts object palettes; a second Hungarian assignment minimizes CIEDE2000 between extracted and target colors. An object passes only if every target entry is matched within $\Delta E_{00} \leq 5$; an image passes only if every expected object passes. Localization, extraction, and matching failures remain in the fixed denominators of 12,080 objects and 5,192 images.

Table 9: **MOC localization coverage and color accuracy across EMA checkpoints.** Rates and accuracies are percentages. Bold marks the best metric within each step.

Step	Configuration	BBox match ↑	Evaluated	Object acc@5 ↑	Image acc@5 ↑
30k	NoColor	96.16	11,616	26.60	6.95
	Pixel	96.25	11,627	**85.93**	**72.46**
	Predicted	96.90	11,706	76.59	57.53
	Predicted-Leaf	**97.04**	11,723	72.73	51.69
40k	NoColor	96.62	11,672	23.68	5.72
	Pixel	96.59	11,668	**86.34**	**72.84**
	Predicted	97.01	11,719	77.14	57.53
	Predicted-Leaf	**97.19**	11,741	71.95	50.69
50k	NoColor	96.78	11,691	22.62	5.08
	Pixel	96.78	11,691	**86.18**	**72.69**
	Predicted	97.14	11,734	76.73	57.36
	Predicted-Leaf	**97.16**	11,737	71.47	50.40

At 40k, Pixel exceeds Predicted by 15.31 percentage points in image accuracy (95% CI [13.52, 17.05]) and 9.20 points in object accuracy ([8.22, 10.16]). We use 10,000 paired bootstrap resamples of prompts within the four MOC subsets, retaining each prompt's four generated images and all target objects. Each replicate divides pooled pass counts by pooled denominators. These marginal intervals describe prompt variability for fixed model outputs, not training-run variability. Per-prompt counts and the analysis script are provided in the supplementary files.

MOC: continuous errors. Matched ΔE_{00} averages the Hungarian color-pair costs within each localized element, then weights elements equally. Pair-weighted error pools all matched pairs. Palette EMD uses the same CIEDE2000 costs, uniform target weights, and generated K-means cluster pixel shares. These metrics exclude localization failures and must be read with Table 9.

Table 10: **MOC color errors across EMA checkpoints.** All metrics are lower-is-better. Bold marks the best metric within each step.

Step	Configuration	Matched mean	Matched median	Pair-weighted	Palette EMD
30k	NoColor	7.423	6.866	8.127	16.454
	Pixel	**2.517**	**1.998**	**3.170**	**8.458**
	Predicted	3.271	2.701	3.826	12.987
	Predicted-Leaf	3.652	3.001	4.203	12.610
40k	NoColor	7.811	7.328	8.536	17.152
	Pixel	**2.512**	**1.994**	**3.125**	**8.627**
	Predicted	3.196	2.581	3.720	13.091
	Predicted-Leaf	3.686	2.995	4.212	13.209
50k	NoColor	8.048	7.543	8.759	17.456
	Pixel	**2.494**	**1.954**	**3.102**	**8.625**
	Predicted	3.232	2.603	3.763	13.210
	Predicted-Leaf	3.708	2.998	4.235	13.627

C.5 Official NCU evaluation and prompt-conversion audit

Inputs and targets. Protocol A generates from the PE JSON and scores against the original benchmark HEX/RGB. B reuses those images but scores against the object colors declared in the PE JSON. C replaces only the target object's first palette entry with its benchmark color, generates from this color-injected JSON, and scores against benchmark targets. All remaining palette entries, descriptions, and other fields stay unchanged.

Table 11: **Official NCU results under protocols A, B, and C.** Top: pooled accuracy and mean color error. Bottom: sub-test accuracy. Bold marks the best configuration within each protocol.

	A: PE JSON / benchmark targets		B: PE JSON / PE-declared targets		C: color-injected JSON / benchmark targets	
Training configuration	Acc. ↑	Mean ΔE↓	Acc. ↑	Mean ΔE↓	Acc. ↑	Mean ΔE↓
NoColor	27.92	7.876	31.78	6.532	27.99	8.167
Pixel	52.38	4.813	**92.47**	**1.283**	**86.97**	**2.053**
Predicted	**55.24**	**4.359**	90.24	1.504	85.50	2.333
Staged	52.85	4.434	84.92	1.816	78.03	2.906
Predicted-Leaf	53.74	4.401	81.98	1.911	77.45	2.875

Training configuration	Protocol	HEX-CSS	RGB-L1	RGB-L2	RGB-L3
NoColor	A	54.78	17.00	14.01	18.26
Pixel	A	**89.26**	43.23	29.91	39.22
Predicted	A	86.89	**47.38**	34.66	**45.07**
Staged	A	84.04	45.92	33.62	41.52
Predicted-Leaf	A	82.48	46.92	**34.83**	44.52
NoColor	B	58.59	18.92	19.05	22.11
Pixel	B	**95.22**	**91.38**	**92.76**	**90.00**
Predicted	B	93.15	90.00	91.16	86.67
Staged	B	91.52	83.38	82.97	80.74
Predicted-Leaf	B	89.67	79.46	79.31	77.81
NoColor	C	54.85	16.69	14.61	18.07
Pixel	C	**93.07**	**85.08**	82.89	**85.30**
Predicted	C	91.52	82.46	**82.93**	83.15
Staged	C	86.67	76.31	72.03	75.37
Predicted-Leaf	C	84.67	75.69	74.57	73.56

All five configurations use 90k EMA, 2,255 prompts, seeds 0–3, Euler with 100 steps, guidance 4.0, timeshift 4.6, and 512×512 output. The subsets contain 675 HEX-CSS, 325 RGB-L1, 580 RGB-L2, and 675 RGB-L3 prompts; RGB-CSS is excluded. Each protocol covers 9,020 images. The additional Staged configuration uses Pixel annotations through 60k, then Predicted annotations through 90k, with leaf-only palettes in both stages. Its distinct recipe uses $p_{\text{all}} = 0.75$, per-entry deletion 0.6 (retaining at least one), and palette-order shuffling 1.0; other drops follow the base recipe. These rates do not apply to Predicted-Leaf.

Scoring and failures. The original NCU evaluator uses GroundingDINO, SAM2, negative-label mask processing, target-guided OneHue/PCA extraction, and color-neighborhood fallback, without VQA. Its joint criterion requires $\Delta C < 5$, $\Delta E_{00} < 5$, and hue MAE $< 30°$; hue error is zero if either chroma is below 5, and CSS targets permit nearest ISCC-L3 fallback. Segmentation failures count as incorrect:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\text{correct}_i], \qquad N = 9{,}020. \tag{1}$$

Mean CIEDE2000 uses successfully segmented and extracted samples. Since the target also affects pixel selection, A versus B measures target-definition sensitivity on fixed images; it does not independently separate PE and generation errors. C additionally changes the generation input. Tables 11 and 12 give aggregate, sub-test, and object-category results.

PE color-rewriting audit. For 283 fruit/vegetable prompts, Table 13 compares the benchmark HEX with the PE-declared dominant whole-object color. Each prompt contributes once, without generated images or seed replication. Increased orange declarations for oranges and yellow declarations for bananas show that PE can replace unusual requested colors with canonical ones, losing instruction fidelity. Structural interpretation errors also occur: one mango is described as a color-card region

Table 12: **Official NCU results by object category under protocols A, B, and C.** Paired columns report accuracy (%) and mean CIEDE2000. Bold marks the best value across models for each category and protocol.

	NoColor		Pixel		Predicted		Staged		Predicted-Leaf	
Object category	Acc. ↑	ΔE_{00} ↓	Acc. ↑	ΔE_{00} ↓	Acc. ↑	ΔE_{00} ↓	Acc. ↑	ΔE_{00} ↓	Acc. ↑	ΔE_{00} ↓
A: PE JSON / benchmark targets										
Fruit/vegetables	19.79	12.710	34.89	9.188	**35.42**	**8.876**	31.27	9.319	32.60	9.226
Tools/misc.	27.69	7.106	51.96	4.539	**57.99**	3.395	56.40	**3.259**	57.49	3.325
Vehicles	30.71	6.751	60.14	4.115	**62.68**	**3.567**	60.60	3.783	**62.68**	3.635
Animals	22.58	8.352	45.25	5.568	**49.22**	**4.237**	44.67	4.713	46.41	4.688
Clothes/accessories	28.30	7.147	60.99	3.460	**62.91**	**3.209**	61.20	3.309	60.99	3.260
Furniture/household	33.69	7.081	50.89	4.527	52.84	4.692	52.48	**4.321**	**52.90**	4.381
Sports/toys	29.23	7.238	59.85	3.455	**63.03**	**3.342**	59.28	3.491	59.69	3.402
B: PE JSON / PE-declared targets										
Fruit/vegetables	30.39	6.275	**80.65**	**1.212**	75.44	1.639	68.37	2.116	65.46	2.087
Tools/misc.	31.47	6.890	**94.11**	1.462	94.04	**1.095**	92.30	1.327	89.03	1.461
Vehicles	32.43	6.607	**83.97**	**2.110**	82.52	2.115	75.27	2.668	71.65	2.738
Animals	29.75	6.814	**93.22**	**1.148**	89.05	1.582	78.00	2.164	75.68	2.228
Clothes/accessories	30.36	6.178	**99.04**	**0.488**	98.76	0.551	96.43	0.805	93.54	0.939
Furniture/household	34.75	6.475	**95.80**	**1.380**	92.43	2.078	86.64	2.098	83.92	2.291
Sports/toys	32.17	6.539	**96.17**	**1.314**	94.46	1.562	90.39	1.852	87.54	1.875
C: color-injected JSON / benchmark targets										
Fruit/vegetables	19.17	13.920	**67.49**	**5.712**	59.19	6.968	48.59	7.974	50.27	7.537
Tools/misc.	27.25	7.431	91.21	1.515	**94.19**	**1.007**	83.72	1.818	83.94	1.653
Vehicles	31.25	6.724	**94.11**	**1.218**	89.76	2.002	81.97	2.517	82.25	2.491
Animals	22.77	8.240	**84.69**	**2.096**	72.19	3.019	67.25	3.238	64.34	3.477
Clothes/accessories	28.98	7.148	89.08	1.368	**93.68**	**1.050**	88.32	1.632	87.57	1.638
Furniture/household	34.28	6.979	91.73	**1.653**	**92.14**	1.847	84.69	2.302	81.80	2.587
Sports/toys	28.58	7.826	86.64	1.533	**88.52**	**1.516**	82.90	1.931	83.96	1.731

with dark text. The audit therefore does not attribute every rewrite or A–C score difference to canonical-color bias.

Table 13: **PE color rewriting.** (a) Fruit/vegetable prompts. (b) Color-category prevalence before and after rewriting. (c) Example substitutions; ΔE_{00} compares declared colors.

(a) Aggregate color preservation: Fruit_Veg only ($n = 283$)	
Exact HEX retained	8.13%
Paired colors with $\Delta E_{00} \leq 5$	38.16%
Mean ΔE_{00}	11.99
Median ΔE_{00}	7.40
ISCC-NBS L1 category changed	33.92%
Yellow/orange prevalence: benchmark → PE	14.84% → 33.57%

(b) Object-specific color-category prevalence

Object	**Color category**	**Benchmark**	**PE**
Fruit			
orange	orange	2/41	24/41
banana	yellow	3/54	15/54
mango	yellow or orange	5/42	17/42
Non-fruit			
bear	brown / black / olive	11/43	16/43
duck	brown / black / olive	12/39	15/39
sports ball	black / white	12/60	15/60
elephant	gray	4/39	7/39

(c) Illustrative paired substitutions

Object	**Benchmark HEX**	**PE HEX**	**ΔE_{00}**
Fruit			
orange	`#652277`	`#F07404`	57.4
orange	`#00FFFF`	`#F98B00`	54.5
banana	`#3B74C0`	`#B99936`	53.1
banana	`#D51C3C`	`#F0B10A`	49.3
apple	`#F8F8FF`	`#C61414`	52.6
papaya	`#FFF5EE`	`#F98000`	33.6
Non-fruit			
bear	`#23EAA5`	`#504E35`	51.7
duck	`#43BDB8`	`#37342D`	52.3
sports ball	`#DEB887`	`#0E0D0B`	67.2
jeans	`#DC7D34`	`#13253B`	54.9
taxi	`#72672C`	`#B09343`	19.1
elephant	`#E7E1E9`	`#B09A8D`	20.9

The paired non-fruit examples and category counts in Table 13 show similar substitutions for bears, ducks, sports balls, and elephants. These descriptive shifts are smaller than for the reported fruit concepts; the examples do not estimate population-wide rewriting frequency or establish statistical significance.

C.6 PUBLISHED NCU REFERENCE RESULTS

Table 14 reproduces NumColor's reported scores for seven base models and five NumColor-augmented variants (Butt et al., 2026b), alongside our protocols A and C (Section C.5). Published systems use their own inputs, training resources, and evaluation settings and serve as external references. The cited columns retain the source's CSS3/X11 and ISCC-NBS level-1/level-3 labels; RGB-L2 and pooled accuracy are unreported.

Table 14: **External numerical-color accuracy (%, ↑).** Bold marks the best value within each of our protocol groups. Baselines: NumColor (Butt et al., 2026b); dashes denote unreported values.

Model	Protocol	HEX-CSS	RGB-L1	RGB-L2	RGB-L3	Pooled
Protocol A						
NOCOLOR	A	54.78	17.00	14.01	18.26	27.92
PIXEL	A	**89.26**	43.23	29.91	39.22	52.38
PREDICTED	A	86.89	**47.38**	34.66	**45.07**	**55.24**
STAGED	A	84.04	45.92	33.62	41.52	52.85
PREDICTED-LEAF	A	82.48	46.92	**34.83**	44.52	53.74
Protocol C						
NOCOLOR	C	54.85	16.69	14.61	18.07	27.99
PIXEL	C	**93.07**	**85.08**	82.89	**85.30**	**86.97**
PREDICTED	C	91.52	82.46	**82.93**	83.15	85.50
STAGED	C	86.67	76.31	72.03	75.37	78.03
PREDICTED-LEAF	C	84.67	75.69	74.57	73.56	77.45
Published methods						
Model	Input	CSS3/X11	ISCC-L1	—	ISCC-L3	Pooled
FLUX	prose	5.9	13.8	—	7.7	—
Sana	prose	9.5	25.1	—	12.9	—
SD3.5	prose	5.8	14.9	—	7.5	—
PixArt-α	prose	4.2	9.5	—	5.4	—
SD3	prose	4.6	11.8	—	5.9	—
PixArt-Σ	prose	4.1	10.2	—	5.1	—
CogView4	prose	4.3	11.0	—	5.3	—
FLUX + NumColor	parsed	52.0	55.7	—	48.0	—
SD3 + NumColor	parsed	47.2	49.2	—	39.5	—
SD3.5 + NumColor	parsed	46.9	51.7	—	43.4	—
PixArt-α + NumColor	parsed	38.5	44.9	—	36.2	—
PixArt-Σ + NumColor	parsed	43.9	48.6	—	40.8	—

C.7 T2I-CompBench++ protocol and complete results

We evaluate PIXEL, PREDICTED, PREDICTED-LEAF, and STAGED at 90k EMA on 2,396 prompts, generating one 512×512 image per prompt with seed 42, 50 Euler steps, guidance 4.0, and timeshift 4.6. All four models receive the same all-scope PE JSON (protocol A), including the leaf-only-trained PREDICTED-LEAF and STAGED. Scoring uses the official T2I-CompBench++ implementation at commit `4aa4042`. The eight-category mean in Table 15 is an unweighted descriptive average of metrics with different scales.

Published references use the benchmark authors' project-page table for SDXL, PixArt-α-ft, and DALL-E 3 (Huang et al., 2025b), and the respective papers for OSPO (7B) (Oh et al., 2026) and CTCal (SD3) (Guo et al., 2026). CTCal uses ShareGPT4V-CoT for non-spatial relations, so that entry is marked separately and its eight-category mean is omitted. These references use different inputs and model configurations.

Table 15: **T2I-CompBench++ results and published references.** Scores ↑; bold marks the best value among our four configurations. †Different non-spatial evaluator.

				Spatial			Non-		8-cat.
Method	Color	Shape	Texture	2D	3D	Numeracy	spatial	Complex	mean
Published reference results									
SDXL	0.588	0.469	0.530	0.213	0.357	0.499	0.312	0.324	0.411
PixArt-α-ft	0.669	0.493	0.648	0.206	0.390	0.506	0.320	0.343	0.447
DALL-E 3	0.779	0.621	0.704	0.287	0.374	0.588	0.300	0.377	0.504
OSPO (7B)	0.857	0.639	0.773	0.356	0.393	0.591	0.316	0.415	0.542
CTCal (SD3)	0.844	0.597	0.758	0.348	0.412	0.629	0.787†	0.381	—
Our configurations; PE JSON inputs (protocol A)									
PIXEL	0.800	0.584	0.693	0.432	0.381	0.715	0.313	0.389	0.538
PREDICTED	**0.825**	**0.615**	**0.736**	0.454	0.396	0.747	0.313	**0.405**	**0.561**
PREDICTED-LEAF	0.817	0.606	0.724	0.455	**0.402**	**0.763**	0.313	0.400	0.560
STAGED	0.814	0.595	0.726	**0.461**	0.389	0.755	**0.314**	0.399	0.557

Color-task A/B/C protocols. For the 300 Color prompts, A scores PE-JSON generations against benchmark targets; B scores the same images against the PE-declared colors; C generates new images from the full color-injected JSON and scores against benchmark targets. Unlike NCU C, CompBench C can revise the complete JSON. All three protocols use the generation settings above and the official BLIP-VQA scorer. Each model/protocol cell contains 300 valid images; where multiple candidates exist for PIXEL, we use the first without score-based selection. B and C apply only to Color; the eight-category evaluation remains protocol A.

Table 16: **T2I-CompBench++ Color scores under protocols A, B, and C.** Bold marks the best value per protocol.

Configuration	A ↑	B ↑	C ↑
PIXEL	0.800	0.830	0.771
PREDICTED	**0.825**	**0.849**	**0.815**
PREDICTED-LEAF	0.817	0.842	0.808

PREDICTED has the highest Color score under all three protocols, followed by PREDICTED-LEAF and PIXEL. B exceeds A for every configuration, while C is below A; the latter comparison changes the full generation prompt.

C.8 PAIRED EVALUATION WITH AND WITHOUT PALETTES

The paired palette comparison tests palette presence within each model on the 700 PRISM (PIXEL format) prompts (Table 17). With-palette inputs use all-scope palettes for PIXEL/PREDICTED and leaf-only palettes for PREDICTED-LEAF. Paired inputs remove all palette fields, preserving descriptions, color words, element order, and other fields. Both conditions use 40k EMA, one 512×512 image per prompt at seed 0, 100 Euler steps, guidance 4.0, and timeshift 4.6. The without-palette outputs are shared with the general generation evaluation (Section C.9).
The official PRISM evaluator, `gpt-4.1-2025-04-14`, scores Alignment and Aesthetic; Combined is their arithmetic mean. All 700 prompts have valid scores in both conditions. With-minus-without gains are computed from unrounded scores.

Table 17: **PRISM (PIXEL format) scores without and with palettes.** Δ is the with-minus-without difference. Bold marks the highest score or gain in each column.

	Aesthetic ↑			Alignment ↑			Combined ↑		
Regime	Without	With	Δ	Without	With	Δ	Without	With	Δ
PIXEL	5.900	6.683	**+0.783**	7.123	7.607	**+0.484**	6.511	7.145	**+0.634**
PREDICTED	**6.300**	**6.687**	+0.387	7.557	**7.753**	+0.196	**6.929**	**7.220**	+0.291
PREDICTED-LEAF	6.163	6.611	+0.449	**7.560**	7.690	+0.130	6.861	7.151	+0.289

Paired uncertainty. Table 18 gives percentile 95% intervals from 10,000 paired bootstrap resamples of the 700 prompts, retaining all six model–condition scores together. PIXEL's Combined gain exceeds PREDICTED's by 0.342 [0.217, 0.469] and PREDICTED-LEAF's by 0.344 [0.221, 0.469]. A bootstrap stratified by the seven PRISM tracks gives the same conclusions. Intervals are marginal, not multiplicity-adjusted, and conditional on the frozen checkpoints, generated images, and judge scores. Per-prompt scores and the analysis script are included in the supplementary files.

Table 18: **Paired Combined-score gains from palettes.** Intervals are 95% prompt-bootstrap CIs. Bold marks the largest mean gain.

Regime	Mean gain	95% CI
PIXEL	**0.634**	[0.541, 0.731]
PREDICTED	0.291	[0.204, 0.380]
PREDICTED-LEAF	0.289	[0.206, 0.374]

C.9 GENERATION WITHOUT PALETTES

Overall checkpoint results. The PIXEL, PREDICTED, and PREDICTED-LEAF no-palette PRISM runs use the input removal, sampler, and evaluator in Section C.8. Table 19 reports results at 30k, 40k, and 50k EMA, with 700 prompts per checkpoint, alongside the NOCOLOR baseline. Both predicted configurations exceed PIXEL in Combined score at all three checkpoints. The 40k rows also supply the without-palette scores in Table 5.

Table 19: **PRISM (PIXEL format) scores without palettes across EMA checkpoints.** Higher scores are better. Bold marks the best metric within each step.

Step	Configuration	Alignment ↑	Aesthetic ↑	Combined ↑
30k	NOCOLOR	7.013	5.627	6.320
	PIXEL	7.150	5.843	6.496
	PREDICTED	7.401	5.917	6.659
	PREDICTED-LEAF	**7.487**	**5.924**	**6.706**
40k	NOCOLOR	7.050	5.804	6.427
	PIXEL	7.123	5.900	6.511
	PREDICTED	7.557	**6.300**	**6.929**
	PREDICTED-LEAF	**7.560**	6.163	6.861
50k	NOCOLOR	7.130	5.996	6.563
	PIXEL	7.167	6.130	6.649
	PREDICTED	7.614	**6.396**	7.005
	PREDICTED-LEAF	**7.661**	6.356	**7.009**

Task-level checkpoint results. Table 20 separates the Combined scores into seven tasks, each containing 100 prompts. At every checkpoint, text rendering gives both predicted configurations their largest task-level gain over PIXEL.

Table 20: **PRISM (PIXEL format) Combined scores by task without palettes.** Higher scores are better. Bold marks the best metric within each step.

Step	Configuration	Imagination	Entity	Text rendering	Style	Affection	Composition	Long text
30k	NOCOLOR	7.105	6.040	2.465	7.100	**7.745**	8.075	5.710
	PIXEL	7.630	6.100	2.675	7.190	7.715	8.080	6.085
	PREDICTED	7.395	6.200	3.510	**7.510**	7.655	**8.100**	**6.245**
	PREDICTED-LEAF	**7.635**	**6.455**	**3.580**	7.345	7.680	8.095	6.150
40k	NOCOLOR	7.085	6.050	2.645	7.365	7.780	8.235	5.830
	PIXEL	7.325	6.355	2.820	7.100	7.835	8.075	6.070
	PREDICTED	**7.625**	6.475	**3.990**	7.490	**8.000**	**8.525**	6.395
	PREDICTED-LEAF	7.460	**6.570**	3.910	**7.515**	7.780	8.340	**6.455**
50k	NOCOLOR	7.420	6.230	2.810	7.425	7.870	8.310	5.875
	PIXEL	**7.465**	6.360	3.240	7.260	7.870	8.140	6.205
	PREDICTED	7.460	**6.755**	4.510	**7.510**	**8.000**	8.455	6.345
	PREDICTED-LEAF	7.450	6.710	**4.600**	7.440	7.935	**8.570**	**6.355**

C.10 COLOR GENERATION FROM LANGUAGE

Color-word generation evaluates descriptive color terms; an exploratory photographic-color test compares generated object colors with reference photographs. Both use 40k EMA, seed 0, one 512×512 image per prompt, and 50 denoising steps for all four configurations.

Color-word generation: data and targets. From 2,246 paired PE prompts and targets in the four GenColorBench NCU subsets (Butt et al., 2026a), we retain the 2,205 prompts with a recognized natural-language color phrase (Table 21). The target element's first recognized phrase supplies the permitted ISCC-NBS classes at Levels 1–3: we project its fixed RGB naming anchors into ISCC classes and take their union. These operational targets approximate color-word meanings rather than

independently validated semantic labels. The numerical target is the saved PE-declared HEX selected from the element's palette. Generation inputs retain descriptive color words but remove all palette fields and HEX/RGB/HSL specifications.

Table 21: **Prompt counts and exclusions for color-word generation.**

NCU subset	Paired PE/target	No color phrase	Retained
HEX-CSS	669	1	668
RGB-L1	325	11	314
RGB-L2	580	10	570
RGB-L3	672	19	653
Total	2,246	41	2,205

Color-word generation: scoring. GroundingDINO and SAM2 localize and segment the target. We subtract negative-part masks, erode by two pixels, and select the largest CIELAB K-means cluster ($K = 5$) independently of the target color. The primary metric is ISCC-NBS Level-2 accuracy against the phrase's permitted classes. Table 22 also reports Level-1/Level-3 accuracy and CIEDE2000 against the saved PE HEX. All accuracies include detection and extraction failures in the denominator of 2,205; mean color error uses valid extractions only.

Table 22: **Color-word generation results.** Rates and accuracies are percentages; bold marks the best value per column.

	Detection	L1	L2	L3	Acc. at $\Delta E_{00} \leq \tau \uparrow$			Mean
Regime	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\tau = 5$	10	15	$\Delta E_{00} \downarrow$
NoColor	99.77	57.10	40.59	**18.14**	12.38	43.99	68.71	13.53
Pixel	99.82	43.36	30.98	5.53	3.27	16.01	32.56	22.41
Predicted	99.59	55.92	39.23	14.33	11.29	44.04	70.02	13.10
Predicted-Leaf	**99.86**	**59.05**	**41.86**	16.96	**13.15**	**47.98**	**71.79**	**12.82**

Color-word generation: uncertainty and sensitivity. We use 10,000 paired prompt-bootstrap replicates within the four NCU subsets, retaining failures. The Predicted-minus-Pixel and Predicted-Leaf-minus-Pixel Level-2 differences are 8.25 $[5.94, 10.48]$ and 10.88 $[8.62, 13.11]$ percentage points (marginal 95% intervals). A sensitivity analysis retains the 767 targets whose selected phrase is exactly red, orange, yellow, green, blue, purple, pink, brown, black, white, or gray/grey, without selecting on model scores. Accuracies are 33.77%, 46.68%, and 52.02% for Pixel, Predicted, and Predicted-Leaf; the latter two exceed Pixel by 12.91 $[8.87, 16.69]$ and 18.25 $[14.47, 22.03]$ points. Thus compound color names do not solely explain the ordering. This subset retains the original anchor mapping and does not ensure each description mentions only one color. Intervals condition on the fixed targets, checkpoint, and generated samples; they do not measure target validity or training variability.

Matched numerical-annotation removal. We pair these no-palette outputs with the original NCU outputs for the same 2,205 prompts, using only seed 0 in both conditions. Checkpoints and sampling settings match, and every model receives the same all-scope annotations when present, including PREDICTED-LEAF. Both conditions use the same object, PE-declared HEX, and target-independent extraction protocol; accuracy requires $\Delta E \leq 5$, with failures retained. Thus Table 6 measures recovery of the same numerical target, rather than comparing numerical and ISCC accuracies. It uses a smaller prompt set and one seed, whereas Table 2 pools four samples per NCU prompt.

Annotation removal also clears numerical color specifications from text and normalizes whitespace. For 2,119 pairs, the serialized input differs only by removal of palette fields; we retain the full 2,205 pairs and use this input-defined subset as a sensitivity check. Table 23 gives marginal 95% intervals from 10,000 paired bootstrap draws within the four NCU subsets, keeping all six model–condition outcomes together. PIXEL's gain exceeds PREDICTED's by 18.37 $[16.19, 20.59]$ points and PREDICTED-LEAF's by 23.31 $[21.04, 25.62]$ points.

Table 23: **Paired numerical-accuracy gains from annotations.** Gains are percentage points; CIs are conditional on the frozen checkpoints, images, and extracted colors.

Configuration	Gain at $\Delta E \leq 5$	95% CI
PIXEL	84.35	[82.81, 85.90]
PREDICTED	65.99	[63.85, 68.07]
PREDICTED-LEAF	61.04	[58.87, 63.22]

On the 2,119 strict pairs, the gains are 84.52, 65.88, and 60.83 points for PIXEL, PREDICTED, and PREDICTED-LEAF. The same with/without ordering holds at thresholds 10 and 15 on all 2,205 prompts. These point-estimate sensitivities were defined before computing the paired results. The experiment measures reliance on numerical annotations for matching a fixed HEX; it does not test whether the remaining language uniquely specifies that HEX, or how models resolve conflicting language and palettes. Anonymous paired errors, validity flags, and analysis code are included in the supplementary files.

Exploratory photographic color matching: data and targets. The Object Color Statistics dataset provides photographs, object labels, and manual foreground masks.[1] Matching 44 candidate concepts through a fixed alias mapping yields 276 images across 28 concepts, with unequal concept sizes. We erode each foreground mask by one pixel and extract the largest CIELAB K-means cluster ($K = 5$) as the target color. Color extraction succeeds for all 276 references; this does not establish agreement between the descriptions and reference colors.

Photographic color matching: generation and scoring. Each photograph is converted to a structured prompt; palette fields are removed while descriptions, including any color words, are retained. GroundingDINO localizes the generated object using its source label, and SAM2 supplies its mask. The same target-independent K-means procedure extracts the generated dominant color, which is compared with the photograph's target using CIEDE2000. Detection and extraction failures count as accuracy failures; mean color error uses valid extractions only.

Photographic color matching: aggregation. Macro accuracy averages concept-wise accuracies equally across the 28 concepts; micro accuracy pools all 276 instances. Macro mean error averages concepts with valid extractions (27 for PREDICTED-LEAF, 28 for the others); micro mean error pools valid samples. On the shared 27 concepts, mean errors are 19.45, 25.06, 19.71, and 19.38 for NOCOLOR, PIXEL, PREDICTED, and PREDICTED-LEAF. Accuracy retains all 28 concepts (Table 24).

Photographic color matching: uncertainty and scope. With 10,000 paired bootstrap replicates resampling photographs within each fixed concept, the macro acc@10 gains over PIXEL are 13.52 $[6.23, 20.81]$ points for PREDICTED and 16.91 $[10.30, 23.64]$ for PREDICTED-LEAF. Four concepts have one photograph and contribute no within-concept resampling variation. Resampling whole concepts instead gives intervals of $[-1.11, 27.38]$ and $[5.65, 28.49]$, respectively; the first includes zero. These conditional estimates depend on the concept set and frozen outputs, and do not establish new-concept generalization.

An audit of the archived prompt text flags 21 cases in which a source-query object is omitted or replaced in the generation description, such as a road target paired with a water-surface description. The full results retain these records. Excluding the 21 flagged cases leaves 255 photographs across all 28 concepts, with macro acc@10 of 36.91%, 24.14%, 34.31%, and 39.96% for NOCOLOR, PIXEL, PREDICTED, and PREDICTED-LEAF. Both predicted configurations still exceed PIXEL, but this sensitivity check does not validate the remaining descriptions, reference colors, or masks. We therefore treat this assay as exploratory rather than direct evidence of color-word adherence.

Table 24: **Exploratory photographic color matching (all 276 records).** Macro averages concepts; micro averages instances. Rates and accuracies are percentages; bold marks the best value within each aggregation. Description–target mismatches limit interpretation (see text).

		Detection	Acc. at $\Delta E_{00} \leq \tau \uparrow$			Mean
Aggregation	Regime	$\uparrow$	$\tau = 5$	10	15	$\Delta E_{00} \downarrow$
Macro	NOCOLOR	**98.47**	**15.08**	34.33	46.63	22.14
	PIXEL	98.06	9.32	22.37	34.40	27.08
	PREDICTED	**98.47**	14.69	35.89	49.60	20.58
	PREDICTED-LEAF	93.73	14.67	**39.28**	**50.55**	**19.38**
Micro	NOCOLOR	98.55	**15.94**	35.14	50.36	19.34
	PIXEL	**98.91**	4.35	15.94	26.45	30.17
	PREDICTED	98.55	14.49	36.96	**53.62**	**19.03**
	PREDICTED-LEAF	97.46	11.59	**37.68**	52.90	19.12

[1] https://neidatacommons.nei.nih.gov/perception-cognition-and-action/object-color-statistics

D PALETTE PERTURBATIONS AND ANNOTATION-REGIME SWAPS

D.1 EVALUATION INPUTS AND SCORING

We use the 700 PRISM prompts (Fang et al., 2026) in PIXEL format (Appendix B.3) with corrected structured descriptions, 40k EMA checkpoints, 50 sampling steps, and seed 0, generating one image per prompt per condition. All conditions, including the original-palette baseline, remove color words from generation-input descriptions to avoid explicit text–palette conflicts in the model input. Palette scope follows training: PIXEL and PREDICTED receive all-scope palettes, whereas PREDICTED-LEAF receives leaf-only palettes. Each configuration uses the same scope for its original-palette and perturbed conditions. All 11 conditions cover all 700 prompts for each configuration, with no missing scores. The official PRISM evaluator uses `gpt-4.1-2025-04-14` for Aesthetic and Alignment; higher scores are better. Alignment uses the original official English caption, including its color words, as a fixed reference across conditions; it measures caption–image agreement rather than adherence to the injected palette.

Color-term removal. We use a case-insensitive lexicon of 426 terms, combining 51 manually supplied terms with GenColorBench color names. Matching prioritizes longer terms and includes up to two preceding color modifiers and an optional `colored`/`coloured` suffix. Removal applies to scene, style, background, and region descriptions; labels, rendered text, spatial structure, and palettes are preserved.

D.2 PERTURBATION CONSTRUCTION

The original-palette baseline is compared with ten perturbation conditions. Strengths are defined separately for each intervention.

Hue rotation. We convert sRGB colors to CIELAB/LCh under D65 and rotate hue by $60°$, $120°$, or $180°$, preserving lightness and initially preserving chroma. For out-of-gamut colors, binary search reduces chroma to the largest in-gamut value at fixed lightness and rotated hue; results are rounded to eight-bit sRGB.

Random replacement. For each palette of length N, we replace the first $\lceil pN \rceil$ entries in their original order, with $p \in \{0.2, 0.6, 1.0\}$, by colors sampled uniformly from 24-bit sRGB. A deterministic random sequence for each prompt and palette is shared across strengths, making replacements nested.

Lightness-order rearrangement. Within each global or element palette, we preserve the HEX multiset and rearrange entries toward descending CIELAB lightness. Let M be the length of a minimum adjacent-swap sequence to this target order. At $r \in \{0.2, 0.6, 1.0\}$, we execute the first $\lceil rM \rceil$ swaps. The percentage therefore measures progress toward the target ordering; color values remain unchanged.

Cross-image substitution. A fixed one-to-one mapping assigns each prompt another prompt as its palette donor, without self-pairs. We copy the donor's global palette and assign its complete element palettes cyclically in JSON order, without semantic matching. The jth palette-bearing target element receives donor element palette $j \bmod M$, with zero-based indexing and M donor element palettes. A missing global palette falls back to the donor's first palette; missing element palettes fall back to its global palette. For PREDICTED-LEAF, assignment operates on leaf palettes; other element fields remain present.

D.3 RESULTS AND AGGREGATION

Let $A_{m,0}$ denote the baseline Aesthetic mean and $A_{m,c}$ the mean for regime m under non-baseline condition c. We report the signed change $\delta A_{m,c} = A_{m,c} - A_{m,0}$ and summarize degradation by

$$D_m = \frac{1}{10} \sum_{c=1}^{10} \big(A_{m,0} - A_{m,c} \big). \tag{2}$$

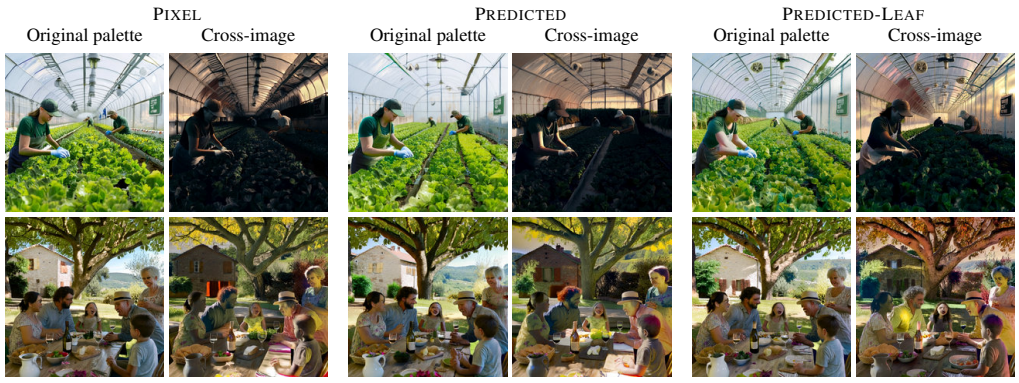

Figure 5: **Cross-image palette replacement.** Original and substituted palettes for a greenhouse (top) and family gathering (bottom). Each pair holds the model and generation seed fixed; palette scope follows each model's training configuration. These are selected examples from the aggregate perturbation evaluation.

The ten conditions receive equal weight; three belong to each perturbation family and one is cross-image. The resulting D_m values are 1.797, 1.866, and 1.537 for PIXEL, PREDICTED, and PREDICTED-LEAF, respectively; smaller values indicate less degradation. Table 25 reports the complete condition-wise results. Displayed scores and changes are rounded independently, so subtracting printed means can differ from the reported change by 0.001.

Table 25: **Condition-wise palette perturbation results.** Aes.: Aesthetic; Align.: Alignment; δA: change from the original-palette score; L-order: lightness-order rearrangement. Bold marks the highest value per metric within each condition.

	PIXEL			PREDICTED			PREDICTED-LEAF		
Condition	Aes.↑	Align.↑	δA ↑	Aes.↑	Align.↑	δA ↑	Aes.↑	Align.↑	δA ↑
Original palette	6.633	7.663	0.000	**6.681**	**7.667**	0.000	6.443	7.664	0.000
Hue 60°	**5.294**	6.530	-1.339	5.226	6.561	-1.456	5.203	**6.663**	**-1.240**
Hue 120°	**5.046**	6.254	-1.587	4.990	6.296	-1.691	4.999	**6.443**	**-1.444**
Hue 180°	4.941	6.261	-1.691	5.001	6.347	-1.680	**5.156**	**6.534**	**-1.287**
Replacement 20%	3.567	5.540	-3.066	3.536	5.513	-3.146	**3.767**	**5.779**	**-2.676**
Replacement 60%	3.364	5.221	-3.269	3.294	5.096	-3.387	**3.477**	**5.361**	**-2.966**
Replacement 100%	3.343	5.143	-3.290	3.366	5.147	-3.316	**3.434**	**5.276**	**-3.009**
L-order 20%	6.351	7.436	-0.281	**6.381**	7.499	-0.300	6.337	**7.524**	**-0.106**
L-order 60%	**6.126**	7.307	-0.507	5.883	**7.374**	-0.799	6.034	**7.374**	**-0.409**
L-order 100%	**6.011**	7.254	-0.621	**6.011**	7.360	-0.670	5.934	**7.369**	**-0.509**
Cross-image	4.313	5.883	-2.320	4.466	5.960	-2.216	**4.713**	**6.176**	**-1.730**

Paired uncertainty. For each prompt, we subtract PREDICTED-LEAF's baseline-relative Aesthetic loss from PREDICTED's, then average across conditions within each family or across all ten conditions. Table 26 reports the mean difference ΔD. Percentile 95% confidence intervals use 10,000 bootstrap resamples of 700 prompts, retaining all conditions and models for each sampled prompt together. Intervals are marginal, not multiplicity-adjusted, and conditional on the fixed checkpoints, images, and judge scores.

Averaging absolute Aesthetic scores over the ten perturbations, PREDICTED-LEAF exceeds PREDICTED by 0.090 [0.042, 0.138] and PIXEL by 0.070 [0.016, 0.125]. Thus, its smaller loss accompanies a small absolute quality advantage. A bootstrap stratified by PRISM track supports both comparisons and the loss differences. The supplementary files provide the actual perturbation inputs, per-prompt scores, and verification and analysis scripts.

Table 26: **Paired Aesthetic-loss differences under palette perturbation.** Positive ΔD indicates a smaller mean loss for PREDICTED-LEAF than PREDICTED.

Family	Mean ΔD	95% CI
All	0.329	[0.199, 0.460]
Hue	0.285	[0.145, 0.430]
Replacement	0.400	[0.253, 0.545]
Lightness order	0.249	[0.110, 0.390]
Cross-image	0.486	[0.303, 0.674]

D.4 ANNOTATION-REGIME SWAP

We evaluate a 2×2 crossing of training model (PIXEL or PREDICTED) and inference annotation format (PIXEL or PREDICTED). Both inputs cover the 700 PRISM pairs (Appendix B.3), whose PIXEL-format and PREDICTED-format versions share element inventories and boxes. Descriptions and palettes are exchanged together. Every condition uses the 40k EMA checkpoint, Euler sampling with 100 steps, guidance 4.0, timeshift 4.6, and one image per prompt at seed 0. The same official PRISM evaluator returns Alignment and Aesthetic; Combined is their arithmetic mean. Each model scores higher on all three metrics with inputs in its own training format (Table 27). These PRISM pairs use a different structured conversion from the other PRISM experiments on the same 700 prompts, so their absolute scores are not directly comparable.

Paired model comparisons. The PIXEL-minus-PREDICTED Combined difference is $+0.441$ (95% CI $[0.322, 0.559]$) with PIXEL-format inputs and -1.199 ($[-1.315, -1.082]$) with PREDICTED-format inputs. Their difference is 1.639 ($[1.471, 1.805]$), supporting the model-ranking reversal. Marginal intervals use 10,000 paired bootstrap resamples, drawing 100 prompts within each of the seven tracks and retaining all four model–format scores for each sampled prompt. An unstratified paired bootstrap gives the same conclusion. These intervals describe prompt variability for fixed models, generated images, and judge scores. Because the formats change descriptions and palettes together, the reversal does not isolate either field's effect. Per-prompt scores and the analysis script are included in the supplementary files.

Table 27: **PRISM-pair scores under annotation-regime swaps.** Columns compare input annotation formats; bold marks the better input format for each model and metric.

		Inference annotation format	
Model	Metric	PIXEL format	PREDICTED format
PIXEL	Alignment ↑	**6.943**	6.356
	Aesthetic ↑	**5.153**	4.613
	Combined ↑	**6.048**	5.484
PREDICTED	Alignment ↑	6.530	**7.386**
	Aesthetic ↑	4.684	**5.980**
	Combined ↑	5.607	**6.683**

E QUALITATIVE EXAMPLES AND SELECTION DETAILS

The examples illustrate selected behaviors rather than their frequency in the benchmarks. Images are displayed without retouching.

E.1 REGION-LEVEL COLOR EXAMPLES

Figure 6 compares PIXEL, PREDICTED, and PREDICTED-LEAF on four GenColorBench MOC cases at 40k EMA, fixing the prompt and generation seed across models. Per-object scores follow

Appendix C.4. The teddy-bear, tool, and garment cases include nine outputs and 27 target-object scores, all satisfying $\Delta E \leq 5$, while showing differences in fur, reflections, and surface texture.

Figure 1a shows the PIXEL outputs for the teddy-bear and tool cases alongside their supplied object palettes; Figure 6 gives the corresponding three-model comparison.

The teddy-bear and tool cases use seed 0; the garment case uses prompt 230, seed 1. The snowboard/skis case (prompt 181, seed 4) illustrates a color-fidelity failure: both PREDICTED-LEAF color errors exceed the threshold.

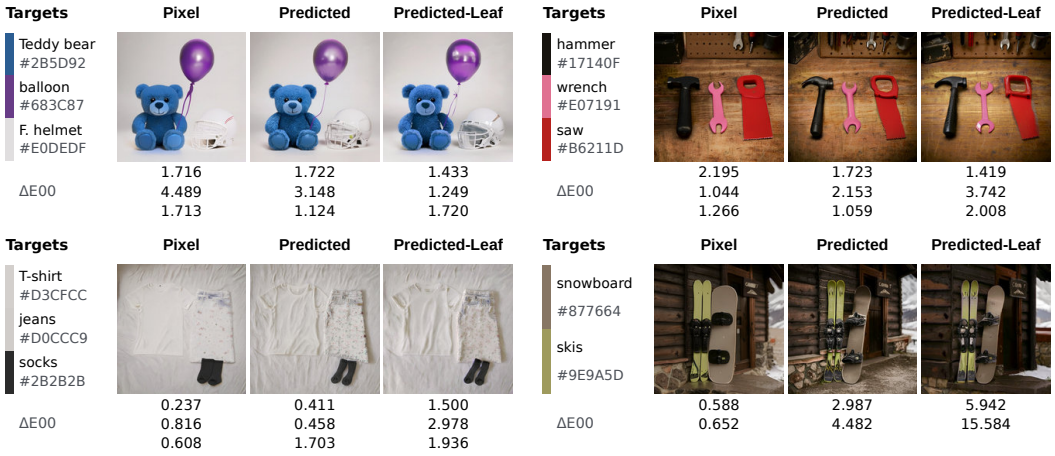

Figure 6: **Color fidelity and material appearance in matched MOC examples.** Left: object targets and HEX values. Below each output: per-object ΔE in target order.

E.2 SELECTION OF PALETTE-CONDITIONING EXAMPLES

Figure 4 shows the classical painting and gilded statue from the paired palette comparison, chosen for positive Aesthetic gains across all three configurations. Figure 5 shows the family gathering and greenhouse from the palette perturbation experiment, chosen for visible degradation and lower cross-image Aesthetic scores. Within-model settings are matched. The palette-removal and perturbation protocols are in Appendices C.8 and D, respectively.

E.3 PRISM (PIXEL FORMAT) PAIRED PALETTE EXAMPLES

Figure 7 compares each palette-trained model with and without palettes on four imagination prompts, holding its weights, descriptions, generation seed, and sampling settings fixed. Table 17 reports the aggregate paired result over all 700 prompts; Table 19 includes NOCOLOR as a no-palette baseline.

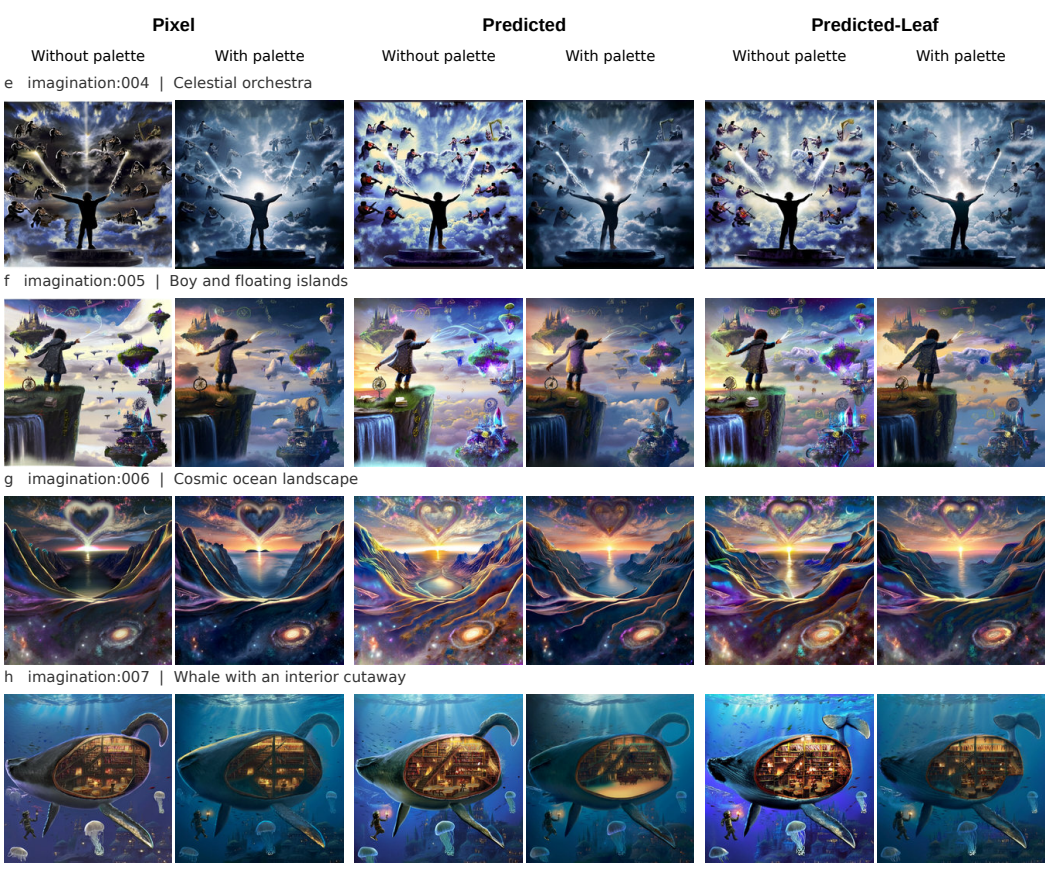

Figure 7: **Without/with-palette output pairs on PRISM (PIXEL format).** Imagination items 004–007.

E.4 T2I-COMPBENCH++ EXAMPLE PROTOCOLS

Cases and selection. We compare SDXL, OSPO, PIXEL, PREDICTED, and PREDICTED-LEAF on 32 T2I-CompBench++ prompts, four from each of its eight categories (160 images). In a fixed candidate order, we retained the first four prompts per category whose PE conversion preserved the requested objects, attributes, counts, and relations under manual inspection. Selection preceded external-baseline generation and is conditional on conversion fidelity. Figure 3 shows the first retained color, texture, 2D spatial, and numeracy cases. Their prompts, from top to bottom, are "a red book and a brown sheep"; "a fabric shirt and a leather wallet"; "a mouse near a bag"; and "a fish, two stools, a lemon, and two vases were on the kitchen island."

Inputs and sampling. SDXL and OSPO receive the original English benchmark prompt. Our three models receive the same all-scope PE JSON used for the 90k evaluation (protocol A), retaining palettes and descriptive color words. This includes all-scope inference for the leaf-only-trained PREDICTED-LEAF. Our outputs use 90k EMA, 512×512 resolution, Euler sampling with 50 steps, guidance 4.0, and timeshift 4.6. All displayed images use seed 42 and the first generated sample for each prompt.
SDXL uses `stabilityai/stable-diffusion-xl-base-1.0` without a refiner, at 1024×1024, with 50 Euler steps, CFG 5, and an empty negative prompt. OSPO uses `KU-AGI/OSPO-Janus-Pro-7B-iter2` through its official Janus inference wrapper, at 384×384, with 576 image tokens, temperature 1.0, and CFG 5. Table 15 reports published baseline scores rather than scores from these qualitative samples.

Interpretation and display. The examples illustrate color binding, material appearance, spatial layout, and counting requests under different input pathways: PE supplies explicit HEX values and layout detail, while model scales, native resolutions, and sampling procedures also differ. Images are displayed at the same size.

E.5 COLOR GENERATION WITHOUT EXPLICIT PALETTES

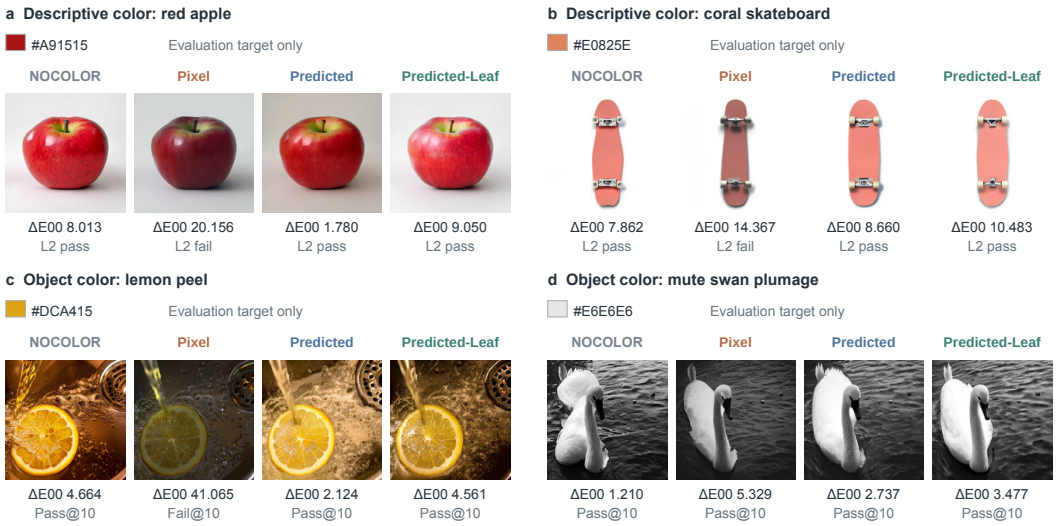

Figure 8: **Color generation without palette input.** (a,b) Color-word generation. (c,d) Photographic color matching. Swatches show evaluation targets, not input palettes; Pass@10/Fail@10 mark whether the instance counts toward Acc@10 ($\Delta E \leq 10$).