DOI: 10.1111/cgf.70384

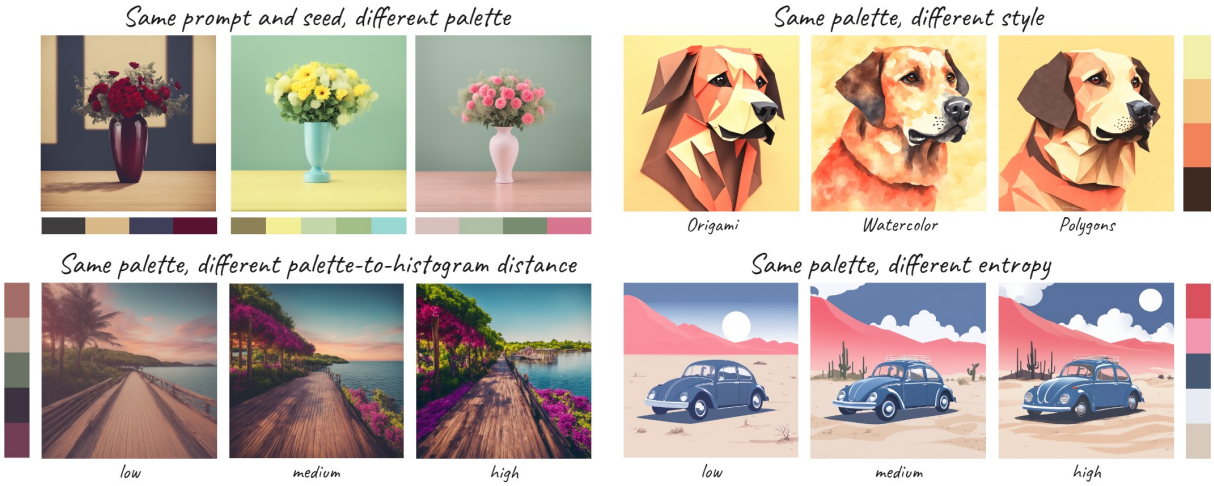
EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

Palette Aligned Image Diffusion

E. Aharoni[1] and N. Porat[1] and D. Lischinski[1] and A. Shamir[2]

[1]The Hebrew University of Jerusalem, Israel
[2]Reichman University, Israel



Figur e 1: *Our method enables precise control over color distribution, generating images, from richly colored to highly quantized, that are perceptually aligned with user-specified color palettes and consistent across different styles.*

Abstract

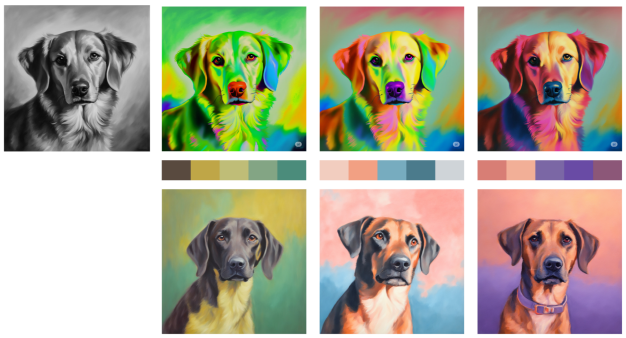
We introduce the Palette-Adapter*, a novel method for conditioning text-to-image diffusion models on a user-specified color palette. While palettes are a compact and intuitive tool widely used in creative workflows, they introduce significant ambiguity and instability when used for conditioning image generation. Our approach addresses this challenge by interpreting palettes as sparse histograms and introducing two scalar control parameters:* histogram entropy *and* palette-to-histogram distance*, which allow flexible control over the degree of palette adherence and color variation. We further introduce a* negative histogram *mechanism that allows users to suppress specific undesired hues, improving adherence to the intended palette under the standard classifier-free guidance mechanism. To ensure broad generalization across the color space, we train on a carefully curated dataset with balanced coverage of rare and common colors. Our method enables stable, semantically coherent generation across a wide range of palettes and prompts. We evaluate our method qualitatively, quantitatively, and through a human evaluation, and show that it consistently outperforms existing approaches in achieving both strong palette adherence and high image quality.*

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics; Neural networks;***

1. Intr oduction

Color plays a central role in visual media, shaping emotion, defining style, and directing attention. In both art and design, the ability to control color is essential: whether to evoke a mood, adhere to brand identity, or achieve aesthetic coherence. Recent advances in text-to-image diffusion models [RBL*22; RDN*22; SCS*22; PEL*23] have enabled the creation of highly diverse and photorealistic images from textual prompts. Meanwhile, mechanisms such as adapters [YZL*23; MWX*23] and ControlNet [ZRA23] have introduced fine-grained control over image structure and style. In this work, we focus specifically on controlling the *colors* that appear



Figur e 2: *Post-hoc recoloring of a grayscale image vs. palette-guided generation: we compare Gemini 2.5's palette-based grayscale recoloring (top row) and our model's palette-aligned generation (bottom row). Our model yields superior color alignment and adapts content, specifically the dog's breed, to the palette, creating semantically coherent images. Importantly, all images share the same prompt and seed. For Gemini, the grayscale image and palette were provided with the prompt: "Please recolor the grayscale image using the provided palette."*

in the generated image, by conditioning on a user-specified *color palette*. Unlike 2D maps used in adapters or ControlNet, a palette is a compact and simple 1D discrete representation that is intuitive and widely used in creative workflows (see Figure 1). A color palette is typically made up of a small fixed set of colors (usually 3 to 10) that guides artists and designers when creating a specific piece or visual identity. The practice of using color palettes dates back more than 40,000 years, when early human artists adopted a basic set of five pigments [Cha19]. Palettes offer a compact, expressive, and visually intuitive means of specifying the intended color theme of an image, without requiring detailed text prompts, reference images, or full color distributions. However, conditioning generative models on a palette alone is a highly ambiguous and ill-posed task. One challenge is *non-uniqueness*: different color distributions may share the same palette, and a single palette can correspond to a wide variety of plausible images. A second challenge is *instability*: small changes in the input palette can lead to large, unpredictable differences in the generated output. Moreover, a palette is typically intended as coarse guidance, as users may expect images to reflect the general hues and style of the palette, while still allowing for subtle shades, blends, and occasional deviations. This tension between expressive flexibility and visual consistency makes plain palette conditioning especially challenging, and therefore requires additional controls.

One might consider a two-stage approach for palette-guided generation: producing a grayscale image, followed by post-hoc recoloring to match the desired hues. However, colorizing a grayscale image using a palette is not straightforward. In addition, such decoupling breaks the alignment between color and content, as the recoloring step is typically unaware of the prompt semantics and may yield visually plausible but semantically inconsistent results (Figure 2, top row). Comparison with additional recoloring methods [QMW25; KC21] is provided in Figure 1 of the Supplementary Material.

Existing methods for joint conditioning attempt to overcome this limitation, but they often rely on strong additional inputs, such as a reference image [YZL*23], a spatial color layout [MWX*23; SHNY25], or an implicit color distribution derived from a reference image [LLG25]. While effective in some scenarios, these approaches do not support lightweight, intuitive control using a simple palette, and they often struggle to maintain prompt-color alignment without sacrificing content fidelity or diversity. Our method addresses these limitations by enabling fine-grained, stable color control using just a prompt and a palette (see Figure 1).

We introduce the *Palette-Adapter*, a novel adapter for diffusion-based text-to-image models that enables precise and flexible control over color through palette-based conditioning. Although inspired by the IP-Adapter [YZL*23], our method departs from it in both functionality and design. Unlike IP-Adapter, which encodes spatial image features, Palette-Adapter conditions the model on a non-spatial color palette, applied at a single intermediate block in the denoising U-Net. To address the ambiguity of palette-based generation, we interpret the input palette as a degenerate histogram, one with a small number of uniformly weighted bins, and train the adapter using both the full image histogram and a derived palette histogram. In addition, we introduce two scalar conditions: the *entropy* of the target and its *distance* from the full image histogram. During inference, these parameters allow users to modulate the level of adherence to the palette. The entropy condition governs the overall color variability, while the distance condition permits controlled deviations beyond the original palette - enabling results that range from stylized, quantized compositions to rich and naturalistic renderings (see Figures 1 and 5). Although *Palette Alignment* is inherently ill-posed, we assume the image is more aligned with the palette when its colors are clustered around the palette colors. Entropy is used to distinguish the different use cases (e.g., higher for photorealistic images and lower for flat graphics). This approach is distinct from previous methods that allow palette color mixing, such as convex hulls, pigment modeling, or layer blending.

To further improve control during inference, particularly when using Classifier-Free Guidance (CFG) [HS22], we expand the concept of negative prompts with a *negative histogram*, a set of user-specified colors that should be suppressed in the generated image. Unlike standard CFG, which extrapolates away from an unconditioned result, our method allows extrapolation along a more semantically meaningful direction: from the negatively guided output toward the positively guided one. This formulation improves both stability and adherence to the desired palette, allowing users to reduce the presence of unwanted hues while retaining high-fidelity palette-aligned results (see Section 3.3).

Training the Palette-Adapter requires data with diverse and balanced color distributions. We found that the LAION-Art [LAI23] dataset, which we initially used for training, exhibits strong color biases: certain hues appear frequently, while others are relatively rare. As a result, many histogram bins are under-represented, limiting the model's ability to generalize to diverse palettes. To address this, we curate an additional set of images that emphasize low-probability colors and augment the training set accordingly. This targeted curation ensures better coverage of the color space and

improves the robustness of our model to a wider range of palettes (see Section 3.5, and Figure 10).

In summary, our contributions are as follows.

1. Palette-to-histogram conditioning: We cast palette conditioning as a special case of histogram conditioning, enabling more structured training and introducing entropy and distance as interpretable control parameters.
2. Enhanced guidance mechanisms: We introduce entropy and distance-based controls, to enable finer and more stable control over the generated color appearance, as well as positive and negative histogram conditioning.
3. Curated training set: A color-balanced dataset that improves generalization in rare and diverse palette configurations.

We train our palette adapter on Stable Diffusion XL [PEL*23], and evaluate it on a wide range of prompts and palettes. Our results include qualitative comparisons, quantitative metrics (Section 4.2), and a human evaluation (Section 4.3), demonstrating the effectiveness and flexibility of our approach. In particular, we show that the entropy and distance parameters provide intuitive handles for controlling stylistic abstraction and palette adherence with minimal user effort (Section 3.4). Our model and data will be made available upon publication.

2. Related Work

2.1. Color Palettes

Color palettes are a widespread tool for artists and designers. They are used to ensure visual harmony and consistency. Palettes can be composed of different types of colors based on their relationships on the color wheel, mood, or theme. The evolution of color theory, pigments, and digital media, further developed a set of rules for selecting color palettes (e.g., Contrastive, Analogous, Harmonic, Triadic, and more [Wik24]). Nowadays many artists use reference books [Wad32] or digital tools (e.g., Adobe Color [Ado24] and Coolors [Coo24]) to explore, create, and choose from millions of aesthetically pleasing, predefined palettes.

2.2. Color Transfer and Recoloring

Over the years, there has been much work on the transfer of color from one image to another, or on colorization of grayscale images. Early work included the pioneering approach of [RAGS01], which transfers colors by matching global color statistics, an approach that was later improved by several follow-up works based on decomposition or distribution transfer [PKD07; AASP17; KC21], harmonization [TEG25], score-distillation-based stylization [BS24], content-aware [DZX*24; QMW25; WXQ*22] or medium-specific such as photographs [CFL*15], pigments [TDLG18] and watercolor paintings [ASL17].

While recoloring techniques can produce pleasing results, they have several limitations:

1. These methods are not generative, and therefore address the recoloring of an already existing image.
2. The global transfer of color statistics may produce unexpected results.
3. Recoloring does not adapt the composition of the source image to align with the desired palette, potentially violating semantic color coherence.

2.3. Non-Textual Conditioning in Diffusion Models

Several methods were proposed to guide diffusion models with non-textual conditions. DALL-E 2 [RDN*22] makes the first attempt to support image prompts, where the diffusion model is conditioned on image embedding rather than on text embedding. IP-Adapter [YZL*23], ControlNet [ZRA23], and T2I-Adapter [MWX*23], have all demonstrated mechanisms to add conditioning to an existing model. IP-Adapter employs duplicates of the existing cross-attention text layers to allow additional independent conditional input. Maintaining the fidelity and text control during inference is typically done using Classifier-Free Guidance (CFG). Our Palette-Adapter is inspired by IP-Adapter, but uses a non-spatial palette condition, and applies it at a single intermediate U-Net block.

2.4. Color-Based Image Generation

Also related to our work is Composer [HCL*23], which conditions image generation on a smoothed CIELab histogram rather than palettes. However, their model is trained from scratch instead of extending an existing pre-trained generative model. Another closely related work is T2I-Adapter [MWX*23], which provides a low-resolution 2D image as a condition. Thus, this is effectively a super-resolution approach rather than palette guidance.

A recent approach by [LLG25] modifies the diffusion sampling process to incorporate the differentiable Sliced 1-Wasserstein distance between the color distribution of the generated image and the reference palette/histogram. However, their method requires a significant number of inference/decode steps (approx. $8\times$), and struggles with out-of-distribution latent perturbation, which we found to compromise image fidelity, especially when working with sparse palettes instead of full histograms (Figure 7, and Figure 4 in the Supplementary Material). Specifically, one can consider the degenerate case of an Earth Mover's Distance (EMD) of 0, where each pixel in the image is strictly taken from the provided palette with balanced distribution. This scenario would inevitably produce severely color-collapsed and heavily posterized images.

Another recent approach by [SHNY25] also provides generation based on 2D palette input. Their finetune approach is conditioned by concatenating a smoothed encoded 2D palette with the denoised latent. Their zero-shot approach directly aligns an unconditionally generated latent to the smoothed color condition during sampling by approximating a one-to-one matching between latent vectors and the conditioning colors, effectively permuting the conditioning colors to fit the latent structure. Similarly to [LLG25], their zero-shot method struggles with out-of-distribution latent perturbations as reported in the paper. Additionally, we found that their finetune approach (reported by design to accept a smoothed image to prevent artifacts), while generally providing good results, tends to produce washed-out images, at times with geometrical artifacts, and struggles with very limited palettes (see Figure 15).

To establish a practical and aesthetically pleasing target, we

calculate a "good" target EMD by inspecting images for which palettes were meticulously extracted by designers, as detailed in our quantitative evaluation. We conclude that an optimal palette-based generation approach must strike a balance between prompt and palette adherence while maintaining fidelity.

Crucially, neither of these prior works explicitly addresses the inherent ambiguity of palette conditioning or provides controls to mitigate it, such as our use of negative palettes and palette-to-histogram distance. Additionally, neither offers further refinement controls like our entropy-based method, which we found highly beneficial, as demonstrated later in this paper.

3. Method

Our goal is to enable palette-based control over the colors in text-to-image diffusion models, while maintaining their ability to generate high-quality images. The resulting image must be semantically aligned with a prompt, as well as perceptually aligned with a color palette. This is challenging since color palettes are inherently ambiguous which makes this an ill-posed task, as explained in Section 1. To achieve our goal, we propose a novel Palette-Adapter, inspired by the IP-Adapter [YZL*23] architecture, but featuring several key novel components that enable non-spatial palette conditioning:

1. Color palettes are treated as sparse color histograms, allowing flexible conditioning and quantitative color specification both during training and inference (Section 3.1).
2. The distribution and variety of colors in the output is managed using two controls. The *distance* condition controls the amount of deviations of colors outside the palette. The *entropy* controls the variation distribution of colors within the palette (Section 3.4).
3. Color alignment is improved by employing dual conditioning: using both positive and negative color conditions. This is implemented via Classifier-Free Guidance (CFG). Guiding away from unwanted hues improves adherence and reduces unrelated colors (Section 3.3). The simplicity and effectiveness of this approach is demonstrated in Figure 4.
4. Training uses a balanced conditioning inspired by [FVSC24; CNS24]. We use a multistage training strategy. Initially, all U-Net blocks are conditioned to identify the critical block for color control. Subsequently, the adapter is finetuned by conditioning only on these blocks. This approach helps prevent overfitting and structural artifacts (see Supplementary Material, Section 1.2).
5. Training uses a curated dataset that addresses existing color biases and ensures balanced representation of diverse colors. The dataset augments LAION-Art with images containing colors from under-represented bins (Section 3.5).

3.1. Encoding Color Conditions

How to compare the multiple colors of an image to a discrete input palette is ambiguous. In addition, there is a vast combinatorial space of possible palettes, and their variable length makes sequence-based encoding non-trivial during training. To address these challenges, we encode the palette color conditions as sparse histogram embeddings. This approach offers several advantages:

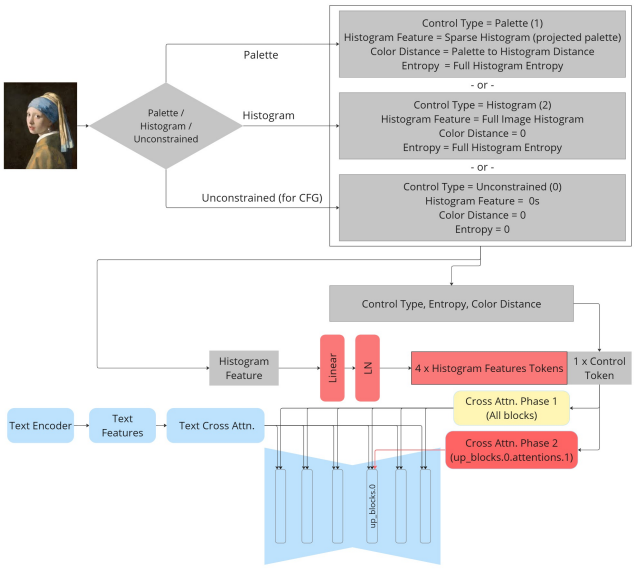


Figure 3: *Training the Palette-Adapter: During training, we randomly choose the control type (histogram, palette, or unconditioned), and compute the distance between the extracted palette/histogram and the full image histogram, as well as the full histogram's entropy. The histogram is projected and normalized into 4 tokens. The remaining features are added as a separate 5th token. All tokens are fed into the adapter's cross-attention layers. We employ a multistage training strategy as described in the Supplementary Material in Section 1.2: trainable layers are colored red, while frozen layers are blue. Layers participating only during the initial training phase are colored yellow.*

1. *Unified Representation:* both sparse palettes and full histograms can be supported as positive or negative conditions.
2. *Quantitative Control:* desired weights can be specified for the palette's colors, if desired.
3. *Training Stability:* both full and sparse histograms can be mixed to reduce sample complexity and dependency on a specific palette extraction method.
4. *Additional Controls:* additional conditioning controls can be defined, such as *entropy* and palette-to-histogram *distance*.

We chose to discretize the color space using a 3D histogram in HSV color space with (34, 12, 10) bins for the H, S, and V channels respectively. Such representation is inspired by the Munsell Color System, whose color atlases commonly feature 40 distinct hues to represent significant perceptual differences, 5–10 chroma levels, and 11 lightness levels. This choice of granularity allows us to capture subtle color variations reflecting the human capacity to discern numerous hues. Our scheme provides a balance between perceptual color resolution and computational tractability, remaining efficient for conditioning and data sampling.

The color histogram of an image is defined by projecting all pixel colors into the histogram bins, while a palette is repre-

sented by projecting its colors into the corresponding histogram bins with uniform probabilities. During training, we alternate between the palette and the histogram conditions with equal probability, employing a dedicated scalar to inform the model of *condition type* (Figure 3). Explicitly providing this condition type allows the model to learn and generalize the inherent differences between palettes and histograms, while also effectively capturing the color information within each bin.

3.2. Adapter Architecture and Training

We propose a multistage training of an IP-Adapter [YZL*23] as further explained in the Supplementary Material (Section 1.2). We use conditioning of a pre-trained frozen diffusion model by adding decoupled cross-attention layers to their denoising U-Net. However, while IP-Adapters are typically used with spatial image data, and used to target *all* U-Net layers, we modify them to accept non-spatial data, and target *specific* layers, making them suitable for our histogram-based color conditioning.

Our architecture, shown in Figure 3, randomly chooses whether to use a sparse color palette or a full histogram as the condition for each training image. The model projects the chosen condition into four tokens through a trainable projection layer, feeding them into the adapter's cross-attention layers. This enables color conditioning throughout the entire diffusion process, similarly to text conditioning. Crucially, this allows the model to both enforce the color conditions and generate compositions that are semantically aligned with them. For instance, a palette can affect the type of the flowers in Figure 1, or the breed of the dog in Figure 2.

To enhance control and alignment, we supplement the color condition with three additional scalar features:

1. *Augmentation type*: palette, histogram, or unconditioned;
2. Distance between the condition and the full histogram of the input image;
3. Entropy of the full image histogram.

Entropy is a global measure of the color distribution information in the image. Posterized or monochromatic images have low entropy, while vivid, colorful, and gradient-rich images have high entropy. We show later how entropy can be used to guide the generation process independent of the histogram condition.

During training, only the Palette-Adapter layers are updated while the pre-trained diffusion model and color condition extraction modules remain frozen. The model processes five attention tokens: four for the projected histogram and one for the additional scalar features. The procedure for identifying color-critical layers is presented in the Supplementary Material Section 1.2.

In our implementation, we adopt the following design choices, though alternatives are viable if consistently applied:

1. *Palette extraction*: For training, we use the median-cut implementation by [Pyl24] to extract color palettes from images in our dataset. We extract up to 8 colors to increase the diversity of the sampled colors (compared to just 3–5).
2. *Color distance*: To measure distances between colors, we use a clipped and sharpened variant of CIEDE2000 [SWD05], similar to [PW09].

Figure 4: *Negative Palettes: Images generated based on a color palette condition (top color bar) and a manually selected negative condition (bottom color bar). The pair on the left was generated with the prompt "An interior design showing a sofa, lamp and a wall with a colored wooden pattern.", while the pair on the right was generated with the prompt "A watercolor painting of one large single centered apple.". We use weighted negative palettes with weights of 0.07 for the sofa, and 0.3 for the apple. A negative palette attenuates or removes the specified color(s).*

3. *Palette to histogram distance*: To measure the distance between the extracted palette and the full histogram of the image, we use the Quadratic-Chi Histogram Distance [PW10]. This method takes cross-bin relationships into account, while reducing the effect of differences caused by bins with large values. For bin-to-bin similarity we use the above mentioned color distance metric.

See Section 4 in the Supplementary Material for color-distance-related formulas and implementation details.

3.3. Negative Palettes

To control the influence of color conditioning during generation, we employ Classifier-Free Guidance (CFG) [HS22]. CFG works by extrapolating the denoising direction away from an unconditional baseline toward the conditioned prediction. While traditional CFG uses unconditioned (zero) values for the unconditional denoising step, we found this approach insufficient for strict color alignment. Generated images might acceptably contain colors absent from the input palette, particularly when dealing with prototypical associations (e.g., yellow bananas).

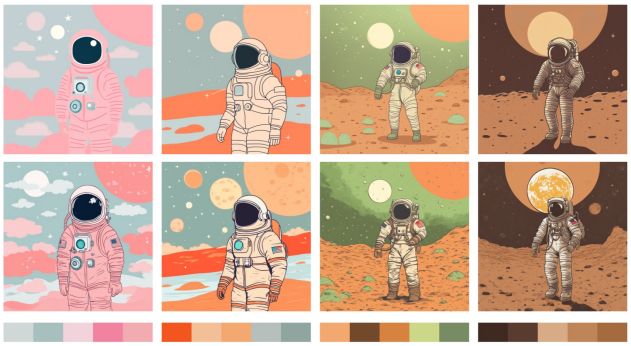
We introduce a novel *negative color condition* approach. Instead of using zero values for unconditional denoising, we allow a manually specified negative palette containing unwanted colors. The predicted noise $\hat{\epsilon}_\theta(x_t, c^{\pm}, t)$ is calculated using both positive (c^+) and negative (c^-) palette conditions:

$$\hat{\epsilon}_\theta(x_t, c^{\pm}, t) = w\epsilon_\theta(x_t, c^+, t) + (1-w)\epsilon_\theta(x_t, c^-, t) \quad (1)$$

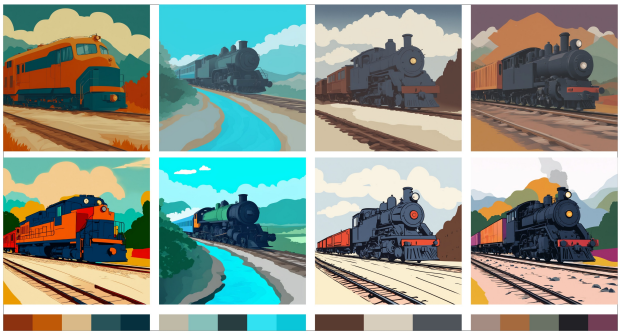
Figure 4 demonstrates how manually specifying a negative color palette can improve alignment with the positive condition. We also found it possible to apply weights to the negative palettes to control the level of attenuation.

3.4. Distance and Entropy Measures

During training, we use the *distance* between the training image and the input palette or histogram as an additional input. When using a histogram, this distance will be small, while the distance



Figur e 5: *Entropy conditioned images: Generated with low and high levels of relative entropy, and different palettes for the prompt "A poster of an astronaut on the moon. Flat colors. Flat vector sticker style", and the same seed. Note how as the relative-entropy grows in the second row, more color shades, texture, and details are added.*



Figur e 6: *Distance conditioned images: Generated with low and high levels of relative distance, and different palettes for the prompt "A train engine driving through a landscape, flat colors, vector graphic, poster art, 8k.", and the same seed. Note how as the relative-distance grows in the second row, more hue shades, contrast and details are added.*

from a palette would be large. At inference time, this allows for additional control over the distribution of colors outside the palette. Larger distances allow more colors outside the palette to appear in the resulting images.

The *entropy* of the color distribution (histogram) can be used as a continuous and effective way to control the level of detail and the color distribution of the generated image. The entropy of the histogram, computed using Shannon's formula [Sha48], effectively controls how much the colors are "binned". If H is the histogram and p_i is the probability of bin i:

$$Entropy(H) = -\sum_{i=1}^{\#bins} p_i \log_2 p_i \tag{2}$$

Thus, conditioning on low entropy will lead to a "binned" histogram (flat, monochrome, etc.), and conditioning on high entropy will lead to more distributed colors (gradients, details, etc.). Note that, different from the palette *distance* measure, entropy does not depend on the specific palette and only captures the nature of the distribution. Figure 5 illustrates images generated for different entropy values for different palettes. Since entropy is a positive measure, to lower the entropy from the unconditioned state, we use CFG with a positive entropy value for the null condition, and a lower value for the positive condition, effectively resulting in a negative "relative entropy" (RE) condition.

3.5. Curating the Dataset

Diffusion models are inherently data-driven, as their inference relies on sampling from a learned distribution. Prior work, e.g., [PP23], has demonstrated that dataset biases concerning gender, race, and age are reflected in the generated imagery. For colors, most datasets are skewed towards natural images, leading to similar color biases in models trained on them. In general, this is a desirable trait, but can become problematic for color-conditioned models, which require sufficiently diverse and balanced data for adequate representation of each color condition embedding (bin). To ensure adequate representation of rare colors, we curate a more balanced dataset as motivated by the bias shown in Figure 13. Specifically, our dataset consists of LAION-Art combined with LAION-2B, filtered for images with rare histogram entries. Further details are available in the Supplementary Material, Section 1.1.

4. Experiments

Our model is based on SDXL 1.0 [PEL*23] and is trained on a curated dataset of 2.4 million images, as detailed in Section 3.5 of the Supplementary Material. All images were resized to a resolution of 512×512 pixels for training. We adopted an approach similar to that of [YZL*23], incorporating drop probabilities as specified in Table 2 of the Supplementary Material to enable Classifier-Free Guidance (CFG), utilizing an empty string for text prompts and zero tensors for other conditioning.

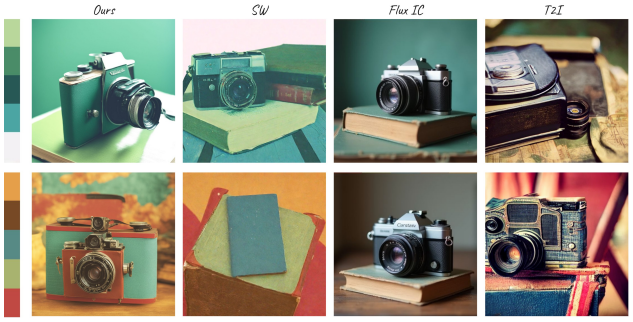
In the following sections, we present both qualitative and quantitative results, including a human evaluation.

4.1. Qualitati ve Results

The results generated by our method are demonstrated in figures throughout the paper. Unless explicitly stated, images in this paper were generated *without* using the negative palette, entropy, and distance controls. In particular, Figures 1, 5, and 6 demonstrate the core capabilities of our method: fine-grained control (entropy, distance, style), and in particular, how different palettes for the same prompt yield dramatically different results, closely matching the conditioning palettes. We show a comparison with other methods in Figure 7. Additional comparison results are found in Section 2 of the Supplementary Material, including Figure 4. For completeness, we also include comparisons with other methods that we found less suitable or without an available model, such as Composer in Figure 16, Adobe Firefly and DALL·E 3 in Figure 19.

4.2. Quantitati ve Evaluation

We compared our model against three palette-conditioned generative models: T2I-Adapter and FLUX IC Palette [Geo24] which are



(a) *prompt: "A vintage camera resting on an old book."*

(b) *prompt: "A poster design of an ancient statue in headphones and sunglasses on a cosmic background."*

Figur e 7: *Comparison of palette-based generation between our method, T2I, SW-Guidance, and FLUX IC. Our method produces high-fidelity results, adhering to both prompt and palette.*

training-dependent methods, and SW-Guidance [LLG25], which is a training-free method. These models directly address the problems of palette-conditioned generation without being confounded by unrelated style transfer techniques, ensuring a fair and focused comparison. Some examples can be seen in Figure 7. We also included the original SDXL 1.0 as a baseline.

Following [LLG25], we evaluated these methods on the COCO 2017 validation set, initially containing 5,000 images. To ensure that color conditioning is strictly driven by the palette provided, we filtered this set for images whose captions do not mention colors and used the first remaining 1,000 images. Since FLUX IC, and SW-Guidance support color palette conditioning (similar to our method), we used K-means (from Pylette) to obtain the five primary colors of each ground-truth image. For T2I Adapter, we initially generated 2D palettes by down-sampling to 8×8 images, and subsequently up-sampling (to 512×512) the ground-truth image using nearest-neighbor interpolation, to match the authors' instructions. Recognizing the potential for inconsistencies with the extracted 1D palettes (which could negatively affect the T2I scores), we have implemented an optimal-transport-based alignment between the downsampled image and the target palette using EMD. Eventually, each model received the same captions as text prompts, same seed, and the extracted palettes (1D or 2D) for generation, resulting in 1,000 generated images per model.

To comprehensively evaluate our model and the baselines, we employed three distinct metrics, each targeting a specific aspect of the generation process: color alignment to the reference palette, image quality, and semantic adherence to the text prompt. For color alignment, we used the EMD, which effectively quantifies the perceptual dissimilarity between the color distributions of the generated images and the provided palettes. To assess the visual quality of the generated images without relying on references, we utilized two trained variants of MusiQ [YKM*21], a state-of-the-art no-reference-image quality model developed by Google, to measure Koniq (Image Quality) and AVA (Image Aesthetics) metrics. Finally, to measure the semantic alignment between the generated images and their corresponding text prompts, we calculated the CLIP similarity score, commonly used to assess prompt alignment.

The results of the comparison are reported in Table 1. Although SW-Guidance provides the lowest EMD score (best), it achieves it by strongly optimizing against a sliced Wasserstein objective during inference. This strong emphasis on color distance appears to come at the cost of image quality, resulting in a lower EMD than observed in manually extracted palettes (see Table 1). SW-Guidance scores the lowest (worst) on both MusiQ metrics, AVA and Koniq, designed to measure aesthetic appeal and perceptual quality, respectively. This indicates that the generated images suffer from visual deficiencies, distortions, and a lack of aesthetic appeal (see Figures 3 and 4 in the Supplementary Material). In contrast, this work aimed to balance these competing factors, a goal our model effectively achieves. While not reaching top color accuracy, it delivers the highest perceptual quality, demonstrating successful color palette integration without sacrificing visual coherence.

Category	EMD $\mu \pm \sigma \downarrow$	Koniq $\mu \pm \sigma \uparrow$	AVA $\mu \pm \sigma \uparrow$	CLIP $\mu \pm \sigma \uparrow$
Our model	0.1414 ± 0.031	70.642 ± 5.789	5.619 ± 0.392	0.311 ± 0.028
SW-Guidance	0.063 ± 0.019	63.628 ± 8.251	4.997 ± 0.523	0.311 ± 0.032
FLUX IC	0.228 ± 0.071	66.429 ± 7.401	5.385 ± 0.615	0.309 ± 0.027
T2I (EMD)	0.1903 ± 0.060	68.008 ± 9.077	5.134 ± 0.601	0.308 ± 0.029
Baseline SDXL 1.0	0.257 ± 0.089	72.118 ± 5.080	5.647 ± 0.409	0.316 ± 0.027
Reference*	0.170 ± 0.089	-	-	-

Table 1: *Average distance of generated images from the conditioning palettes. *The reference distance is the average distance of a set of human-extracted palettes taken from [Pic24] and is used as a baseline, and is not computed based on COCO prompts and palettes as all others.*

4.3. Human Evaluation

We conducted a human evaluation to evaluate our model's generated images for perceptual color alignment with given palettes and for overall quality. Participants ranked the match between a palette and an image on a 1-5 Likert scale. A higher score indicated that all palette colors were present in the image and all image colors were close to a palette color. Participants also rated the overall quality of each image, considering the provided palette.

The study used 40 high-quality images and their *human-extracted* palettes, randomly selected from [Pic24], a set curated by designers. For each, we created a color-neutral descriptive prompt (e.g., "A photo of a bird of paradise flower."). Crucially, all compared methods generated images using the same seed for each prompt. Our method was used without our fine-grained controls (entropy, distance, or negative color prompting), which can further improve palette adherence.

Figure 8: *Ablation: Conditioning on Color-Critical U-Net Blocks. Pairs of images generated using identical prompts, palettes, and seeds to compare two conditioning strategies. Top: Images generated using our final model, where the Palette-Adapter conditions only a few, pre-identified color-critical blocks. Bottom: Images generated using our stage 1 only training, where all cross-attention layers are conditioned. This comparison demonstrates that our 2-stage selective conditioning strategy successfully reduces structural artifacts and prevents overfitting, while maintaining strong color alignment.*

Figure 9: *Ablation: Palette Precision. Pairs of images generated using identical prompts, but different palette precision. Each row from left-to-right showing the original palette, the histogram projected palette, and 2 renditions. The top row showing our selected HSV histogram totaling 4,080 bins, while the bottom row showing an RGB alternative totaling 512 bins. This figure demonstrates the importance of sufficient bin count for palette representation to prevent color skews.*

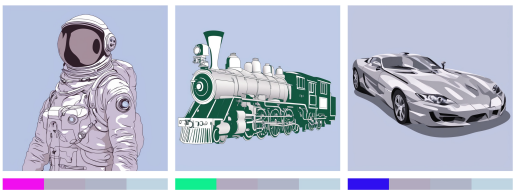


Figure 10: *Ablation: Failure in color conditioning with rare LAION-Art colors prior to sampling. This figure shows examples of failing color conditioning in earlier training experiments where the model trained on non-carefully curated images as described in Section 3.5. The leftmost palette color is taken from the 30 rare colors of the LAION-Art dataset used for training this variant.*

For each palette and prompt, participants evaluated images corresponding to five conditions: the original reference image, and the image generated using our method, FLUX IC Palette, T2I, and SW-Guidance. They rated color alignment and overall quality for each of the resulting 200 images (40 palettes, for 4 models and a reference). These tasks were divided into four forms, each containing 10 palette sections. Forty participants (aged 20-40, from diverse backgrounds, including designers and researchers) each completed one form.

The results are presented in Table 2. Our model's color alignment score (3.87) is very close to SW-Guidance's (3.97), while our mean quality score was notably higher (3.44 vs. 2.88), despite both models being based on the SDXL 1.0 architecture. Our model's palette alignment also ranks higher than FLUX IC, and T2I. The harmonic mean of both scores, along with the Pareto front (see Figure 2 in the Supplementary Material), shows that our model strikes the best balance between adherence to the palette and quality. As expected, FLUX IC Palette, utilizing a newer architecture and larger model, achieved higher overall quality, and all models trailed behind the high-quality real reference images.

Model	Palette Alignment ↑ $\mu \pm \sigma$	Quality ↑ $\mu \pm \sigma$	Harmonic Mean ↑ μ
Our Model	3.873 ± 1.06	3.447 ± 1.10	3.647
FLUX IC	2.969 ± 1.24	3.867 ± 1.10	3.359
T2I	3.072 ± 1.20	3.127 ± 1.22	3.099
SW-Guidance	3.978 ± 1.08	2.884 ± 1.23	3.344
Reference	3.934 ± 1.06	4.230 ± 0.938	4.076

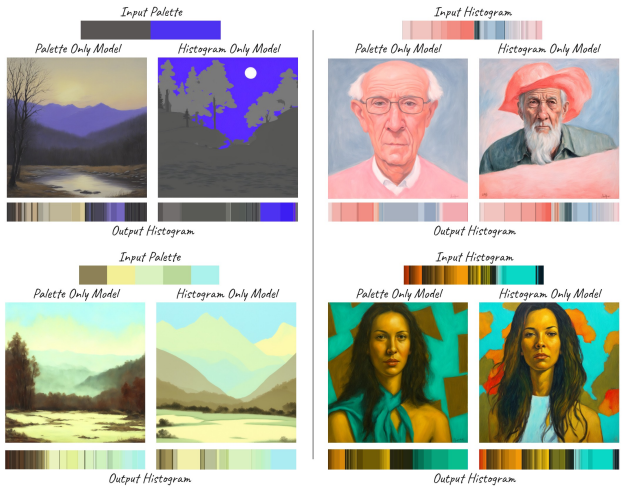
Table 2: *Human evaluation results: users rank palette-image matching, and overall quality (using Likert ranking between 1-5 where 5 is better), based on reference taken from [Pic24].*

4.4. Ablation

We conducted a series of ablation studies to validate key design decisions on data and architecture. We provide four examples below.

Identifying palette-adapter layers: As explained in Section 3.2, our training is based on a multistage approach to identify the vital color-control blocks. We found that this approach does not jeopardize color alignment (see Supplementary Material Section 1.2) while significantly reducing conditioning artifacts. We show qualitative results in Figure 8 where the top row, was generated by conditioning only color critical blocks, eliminates artifacts.

Palette precision: As explained in Section 3.1, our model is conditioned on a 3D HSV histogram shaped $34 \times 12 \times 10$, totaling 4,080 bins. The goal was to strike balance between color precision and total bin count, in order to avoid data and training sampling issues. In earlier iterations, we have explored models based on RGB histogram conditioning with a lower bin count ($8 \times 8 \times 8 = 512$). We visualize color skews as a result of projecting colors to 512 RGB bins compared with our selected HSV histogram totaling 4,080 bins. We note, however, that with lower bin counts the train-



Figur e 11: *Ablation: Palette-only and Histogram-only variants. This figure shows cases where model variants trained only on palettes, or only on full histograms, fail to generalize well for the other use case.*

ing converged faster (400K steps vs. 1M steps). As seen in Figure 9, the bottom row demonstrates how using too few bins leads to color skews, compared with keeping the colors similar when using a larger histogram.

Sampling images with rar e colors: As explained in Section 3.5 we curate a dataset that contains a more balanced representation of colors to ensure sufficient representation for rare bins during training. Earlier training iterations, which were based on the plain LAION-Art dataset, failed to generalize on rare colors. See Figure 10, where rare colors (leftmost on each palette) fail to condition prior to sampling. As demonstrated throughout this paper, and specifically in Figure 20, our current model conditions well on rare colors.

Mixing palettes and histograms: During training, we mix palette-based and full-image histogram representations to enable the model to accept both formats and generalize to weighted inputs. We trained palette-only and histogram-only models (1M steps with a batch size of 16). We observe that while the palette-only model typically performs well, training exclusively on palettes tends to ignore low-probability entries, occasionally skewing the distribution or introducing colors not present in the target condition. Meanwhile, training exclusively on histograms often results in color-collapsed images when provided with a palette input. Each pair of images was generated using the same prompt and input color condition. See Figure 11.

5. Summary and Discussion

Our method enables the generation of images conditioned on color palettes by incorporating a novel Palette-Adapter into an existing frozen text-to-image diffusion model. While color palettes offer a compact and standard means of communicating color schemes, they are inherently ambiguous and loosely defined. To address this challenge, we trained our adapter using a combination of color palettes and histograms, while introducing novel controls such as distance, entropy, and negative color conditioning. We demonstrated the effectiveness of our approach through a variety of experiments and a human evaluation study.

Figur e 12: *Relative Entropy: Artwork images generated based on a color condition and different levels of relative entropy guidance. Higher entropy introduces more details and texture.*

5.1. Limitations and Futur e Work

Our manual experiments reveal two limitations of the current approach. Firstly, we observe a trade-off between text and color alignment as indicated in Section 4.2 when comparing our method against the baseline SDXL 1.0. This means that not all prompts are as effective with every color palette, hindering the model's ability to consistently fulfill user requests in some cases. This could be compensated for by using a negative prompt, or different seeds.

Another limitation is that the scales of the relative entropy and histogram distance measures are somewhat overlapping. Future work could aim to develop a more user-friendly version for controlling the palette adherence.

We observe that some highly saturated colors still remain underrepresented in our training data, potentially leading to biases towards non-photorealistic outputs when included in the guidance palette. We plan to investigate the incorporation of synthetic data and histogram augmentation techniques (e.g. sharpening) to mitigate this.

While our method supports a negative color condition, this condition is currently specified manually and could potentially be provided automatically, if so instructed by the user (for example, by applying early decoding of the latent image during inference and inspecting significant color deviations).

We believe that our palette-aligned generation method provides a valuable tool for artists and designers. Therefore, we plan to release our code, model weights, and data, to facilitate both practical utilization and future research in this area.

Ackno wledgements

This work was partly supported by Israel Science Foundation Grants no. 2203/24 and 1427/25, and NSFC-ISF Research Grant no. 3077/23.

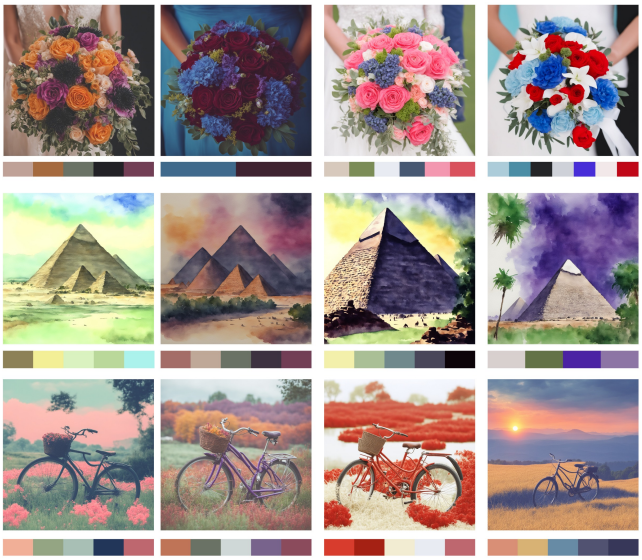
(a) Generated with different seeds for the prompt "A dress."

(b) Generated with different seeds for the prompt: "A waiting room with a sofa and a table."

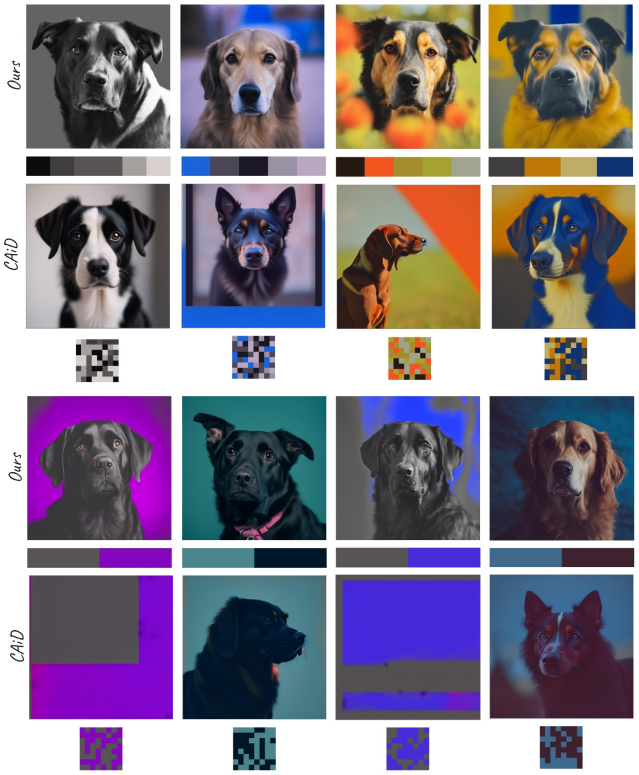
(c) Color distribution in the LAION-Art dataset by probability.

(d) The 30 rarest colors in the LAION-Art dataset.

Figur e 13: *The colors in random images generated by SDXL 1.0 using minimal color-neutral prompts (a, b) reflect well the color distribution in the LAION-Art dataset [LAI23] (c). The 100 most common colors in LAION-Art account for 87% of the total population, while the 30 rarest colors (d) account for only 0.008253%.*



Figur e 14: *A collection of images generated using similar prompts per row and different palettes.*

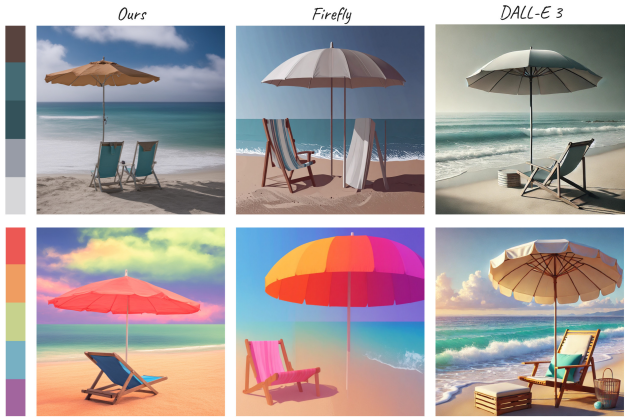


Figur e 15: *Comparison with Color Alignment in Diffusion: images generated using typical and limited palettes with the same seed and prompt "A high quality photo portrait of a single centered dog. 8K. DSLR."*

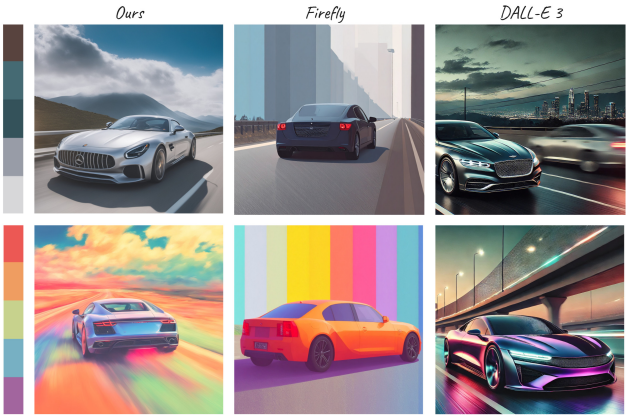
Figur e 16: *Post-hoc recoloring of a grayscale image using Composer vs. histogram-guided generation: we compare Composer's palette-based grayscale recoloring (top row) and our model's palette-aligned generation (bottom row), where the input is provided as a histogram, as our model supports both forms. Our model yields superior color alignment. Images generated using the prompt: "A vintage toy jeep with a soft top roof and open sides, placed on a neutral surface."*

Figur e 17: *Palette alignment with different styles: Generated using our method for different styles with the same seed (and similar prompts excluding the style) for photo, watercolor, children's book illustration, and wool felt.*

Figur e 18: *Portraits: Portraits generated using our methods with the same seed and prompt: "A photo portrait of an old man.".*



Prompt: "A beach chair under an umbrella on the beach."



Prompt: "An expensive car driving the highway."

Figur e 19: *Palette-based recoloring comparison: Comparison between our method, Firefly, and DALL-E 3. Prompt: "A vintage toy jeep with a soft top roof and open sides, placed on a neutral surface.". Our model generates better palette alignment without artifacts.*

Figur e 20: *Technical book illustrations generated with our model: Using the same seed and prompt per row with different palettes, demonstrating robustness to saturated and mild palettes.*

Refer ences

[AASP17] AKSOY, YAĞIZ, AYDIN, TUNÇ OZAN, SMOLIĆ, ALJOŠA, and POLLEFEYS, MARC. "Unmixing-Based Soft Color Segmentation for Image Manipulation". *ACM Trans. Graph.* 36.4 (July 2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3002176. URL: https://doi.org/10.1145/3072959.3002176 3.

[Ado24] ADOBE. *Adobe Color*. https://color.adobe.com/create/color-wheel. [Accessed 24-10-2024]. 2024 3.

[ASL17] AHARONI, ELAD, SHAMBIK, YAKOV, and LISCHINSKI, DANI. "Pigment-based recoloring of watercolor paintings". *Proc. NPAR 2017*. July 2017. DOI: 10.1145/3092919.3092926 3.

[BS24] BINNINGER, ALEXANDRE and SORKINE-HORNUNG, OLGA. "SD-πXL: Generating Low-Resolution Quantized Imagery via Score Distillation". *SIGGRAPH Asia 2024 Conference Papers*. ACM, Dec. 2024, 1–12. DOI: 10.1145/3680528.3687570. URL: http://dx.doi.org/10.1145/3680528.3687570 3.

[CFL*15] CHANG, HUIWEN, FRIED, OHAD, LIU, YIMING, et al. "Palette-based Photo Recoloring". *ACM TOG (Proc. SIGGRAPH)* 34.4 (July 2015) 3.

[Cha19] CHAUDHARY, RAKESH KUMAR. "Influence of colour on visual arts". *International Journal of Business Marketing and Management* 4.1 (2019), 4–9 2.

[CNS24] COHEN, NADAV Z., NIR, ORON, and SHAMIR, ARIEL. *Conditional Balance: Improving Multi-Conditioning Trade-Offs in Image Generation*. 2024. arXiv: 2412.19853 [cs.CV]. URL: https://arxiv.org/abs/2412.19853 4.

[Coo24] COOLORS.CO. *The Super Fast Color Palettes Generator!* https://coolors.co/. [Accessed 24-10-2024]. 2024 3.

[DZX*24] DU, ZHENG-JUN, ZHOU, JIA-WEI, XIA, ZI-XUN, et al. "Palette-Based Content-Aware Image Recoloring". *Computational Visual Media: 12th International Conference, CVM 2024, Wellington, New Zealand, April 10–12, 2024, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2024, 240–258. ISBN: 978-981-97-2091-0. DOI: 10.1007/978-981-97-2092-7_12. URL: https://doi.org/10.1007/978-981-97-2092-7_12 3.

[FVSC24] FRENKEL, YARDEN, VINKER, YAEL, SHAMIR, ARIEL, and COHEN-OR, DANIEL. *Implicit Style-Content Separation using B-LoRA*. 2024. arXiv: 2403.14572 [cs.CV]. URL: https://arxiv.org/abs/2403.14572 4.

[Geo24] GEORGEQI. *Color-Palette-Flux_dev Model (Commit 09dd190)*. Version 09dd19066ef30df4f1a2e817f85c4777b949b15e. Model card and files for a FLUX.1-dev adaptation focused on color palettes. Based on the FLUX.1 architecture. May 2024. URL: https://huggingface.co/GeorgeQi/Color-Palette-Flux_dev/tree/09dd19066ef30df4f1a2e817f85c4777b949b15e 6.

[HCL*23] HUANG, LIANGHUA, CHEN, DI, LIU, YU, et al. "Composer: Creative and Controllable Image Synthesis with Composable Conditions". *arXiv* abs/2302.09778 (2023) 3.

[HS22] HO, JONATHAN and SALIMANS, TIM. "Classifier-Free Diffusion Guidance". *arXiv* abs/2207.12598 (2022) 2, 5.

[KC21] KIM, SUZI and CHOI, SUNGHEE. "Dynamic closest color warping to sort and compare palettes". *ACM Trans. Graph.* 40.4 (July 2021). ISSN: 0730-0301. DOI: 10.1145/3450626.3459776. URL: https://doi.org/10.1145/3450626.3459776 2, 3.

[LAI23] LAION. *laion/laion-art · Datasets at Hugging Face*. Aug. 2023. URL: https://huggingface.co/datasets/laion/laion-art 2, 10.

[LLG25] LOBASHEV, ALEXANDER, LARCHENKO, MARIA, and GUSKOV, DMITRY. *Color Conditional Generation with Sliced Wasserstein Guidance*. 2025. arXiv: 2503.19034 [cs.CV]. URL: https://arxiv.org/abs/2503.19034 2, 3, 7.

[MWX*23] MOU, CHONG, WANG, XINTAO, XIE, LIANGBIN, et al. "T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models". *arXiv* abs/2302.08453 (2023) 1–3.

[PEL*23] PODELL, DUSTIN, ENGLISH, ZION, LACEY, KYLE, et al. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis". *arXiv* abs/2307.01952 (2023) 1, 3, 6.

[Pic24] PICMONKEY. *100 Color Combinations for Designs | Examples & Inspiration | PicMonkey — picmonkey.com*. https://www.picmonkey.com/blog/color-combinations-graphic-design. [Accessed 12-11-2024]. 2024 7, 8.

[PKD07] PITIÉ, FRANÇOIS, KOKARAM, ANIL C., and DAHYOT, ROZENN. "Automated colour grading using colour distribution transfer". *Computer Vision and Image Understanding* 107.1 (2007). Special issue on color image processing, 123–137. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2006.11.011. URL: https://www.sciencedirect.com/science/article/pii/S1077314206002189 3.

[PP23] PERERA, MALSHA V. and PATEL, VISHAL M. "Analyzing Bias in Diffusion-based Face Generation Models". *arXiv* abs/2305.06402 (2023) 6.

[PW09] PELE, OFIR and WERMAN, MICHAEL. "Fast and robust Earth Mover's Distances". *Proc. ICCV*. 2009, 460–467. DOI: 10.1109/ICCV.2009.5459199 5.

[PW10] PELE, OFIR and WERMAN, MICHAEL. "The Quadratic-Chi Histogram Distance Family". *Proc. ECCV*. Vol. 6312. July 2010, 749–762. ISBN: 978-3-642-15551-2. DOI: 10.1007/978-3-642-15552-9_54 5.

[Pyl24] PYLETTE. *A Python color extraction library*. https://qtiptip.github.io/Pylette/. [Accessed 09-11-2024]. 2024 5.

[QMW25] QIU, QIANRU, MAO, JIAFENG, and WANG, XUETING. *Exploring Palette based Color Guidance in Diffusion Models*. 2025. arXiv: 2508.08754 [cs.GR]. URL: https://arxiv.org/abs/2508.08754 2, 3.

[RAGS01] REINHARD, ERIK, ASHIKHMIN, MICHAEL, GOOCH, BRUCE, and SHIRLEY, PETER. "Color Transfer between Images". *IEEE Computer Graphics and Applications* 21 (Oct. 2001), 34–41. DOI: 10.1109/38.946629 3.

[RBL*22] ROMBACH, ROBIN, BLATTMANN, ANDREAS, LORENZ, DOMINIK, et al. "High-Resolution Image Synthesis with Latent Diffusion Models". *arXiv* abs/2112.10752 (2022) 1.

[RDN*22] RAMESH, ADITYA, DHARIWAL, PRAFULLA, NICHOL, ALEX, et al. "Hierarchical Text-Conditional Image Generation with CLIP Latents". *arXiv* abs/2204.06125 (2022) 1, 3.

[SCS*22] SAHARIA, CHITWAN, CHAN, WILLIAM, SAXENA, SAURABH, et al. "Photorealistic text-to-image diffusion models with deep language understanding". *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494 1.

[Sha48] SHANNON, CLAUDE E. "A Mathematical Theory of Communication". *Bell System Technical Journal* 27.3 (1948), 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x 6.

[SHNY25] SHUM, KA CHUN, HUA, BINH-SON, NGUYEN, DUC THANH, and YEUNG, SAI-KIT. *Color Alignment in Diffusion*. 2025. arXiv: 2503.06746 [cs.CV]. URL: https://arxiv.org/abs/2503.06746 2, 3.

[SWD05] SHARMA, GAURAV, WU, WENCHENG, and DALAL, EDUL N. "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations". *Color Research & Application* 30.1 (2005), 21–30 5.

[TDLG18] TAN, JIANCHAO, DIVERDI, STEPHEN, LU, JINGWAN, and GINGOLD, YOTAM. *Pigmento: Pigment-Based Image Analysis and Editing*. 2018. arXiv: 1707.08323 [cs.GR]. URL: https://arxiv.org/abs/1707.08323 3.

[TEG25] TAN, JIANCHAO, ECHEVARRIA, JOSE, and GINGOLD, YOTAM. "Palette-Based Color Harmonization". *IEEE Transactions on Visualization and Computer Graphics* 31.10 (2025), 7809–7819. DOI: 10.1109/TVCG.2025.3546210 3.

[Wad32] WADA, SANZO. *A Dictionary of Color Combinations*. Japan: Seigensha, 1932 3.

[Wik24] WIKIPEDIA, THE FREE ENCYCLOPEDIA. *Color Scheme*. http://en.wikipedia.org/w/index.php?title=Color%20scheme&oldid=1247518638. [Accessed 24-10-2024]. 2024 3.

[WXQ*22] WANG, YI, XIA, MENGHAN, QI, LU, et al. *PalGAN: Image Colorization with Palette Generative Adversarial Networks*. 2022. arXiv: 2210.11204 [cs.CV]. URL: https://arxiv.org/abs/2210.11204 3.

[YKM*21] YANG, FENG, KE, JUNJIE, MILANFAR, PEYMAN, et al. "MUSIQ: Multi-scale Image Quality Transformer". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 7.

[YZL*23] YE, HU, ZHANG, JUN, LIU, SIBO, et al. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models". *arXiv* abs/2308.06721 (2023) 1–6.

[ZRA23] ZHANG, LVMIN, RAO, ANYI, and AGRAWALA, MANEESH. "Adding Conditional Control to Text-to-Image Diffusion Models". *arXiv* abs/2302.05543 (2023) 1, 3.