Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30678

PixDiff-PIG: Palette-Informed Diffusion for Pixel Art Generation

Lang Li

Independent Researcher, Shanghai, China
lilang8936@gmail.com

Abstract. Pixel-style image generation remains an under-explored regime where conventional diffusion models, typically optimized for natural photographs, fail to preserve discrete palettes and crisp silhouettes. This paper introduces PixDiff-PIG, a palette-informed diffusion framework tailored for low-resolution sprite synthesis. The system integrates k-means palette quantization, OKLab color modeling, and language-vision guidance into a compact TinyUNet-based denoising diffusion probabilistic model (DDPM). Each sprite is compressed into a feature bundle comprising a quantized palette, palette index map, luminance channel, and BLIP-generated caption, enabling efficient training and interpretable conditioning. PixDiff-PIG operates directly in OKLab space and optionally employs CLIP-based guidance during sampling to align generations with textual prompts while respecting palette structure. Experiments on a curated 42k-sprite dataset demonstrate that PixDiff-PIG outperforms StyleGAN2 and a latent diffusion baseline, achieving lower FID and higher palette consistency and prompt alignment. User studies with experienced pixel artists confirm improved aesthetic quality and palette coherence. The proposed framework highlights how palette-aware diffusion can bridge handcrafted pixel-art workflows and modern generative models on commodity hardware.

Keywords: Palette-Informed Diffusion, Pixel Art Generation, TinyUNet, CLIP-Guided Sampling

1. Introduction

Diffusion models have established state-of-the-art performance on a wide range of image synthesis tasks [1-4]. However, most work focuses on continuous-tone photographic imagery, while pixel art remains a niche domain with distinct structural and aesthetic properties. Pixel sprites are characterized by limited palettes, hard edges, and small resolutions, and artists traditionally reason in terms of palettes, tiles, and silhouettes rather than global textures. Directly applying generic diffusion pipelines to such data often leads to blurred contours, palette drift, and loss of stylistic discipline.

This paper proposes PixDiff-PIG, a palette-informed diffusion framework specifically designed for pixel art generation. The core idea is to make palette structure a first-class citizen in the generative process. Instead of training on raw RGB images, we construct compact feature packs that encode a sprite's quantized palette, palette index map, luminance information, and an automatically

DOI: 10.54254/2755-2721/2026.TJ30678

generated caption. These features are then decoded into OKLab space, where a TinyUNet-based DDPM learns to predict additive noise. By operating in a perceptual color space and incorporating discrete palette structure, the model becomes more robust to the artifacts introduced by low resolution and hard quantization.

The goals of this aricle are threefold: (i) preserve palette structure and crisp silhouettes during training and sampling, (ii) support text-conditioned sprite generation through BLIP and CLIP without requiring large-scale text-image corpora, and (iii) maintain a compact architecture suitable for consumer GPUs. We further emphasize interpretability: explicit palette and luminance channels make it easier for artists to understand and manipulate intermediate representations.

Empirical results show that PixDiff-PIG improves over StyleGAN2 and latent diffusion baselines in quantitative metrics and human preference evaluations. The model achieves lower Fréchet Inception Distance (FID), higher Palette Consistency Score (PCS), and higher Prompt Alignment (PA), while being more sample efficient and faster at inference. Qualitative analysis reveals that PixDiff-PIG generates cleaner silhouettes, more coherent shading, and disciplined palette usage compared to competing methods. We also analyze limitations, such as dataset biases and resolution constraints, and outline directions for extending palette-aware diffusion to broader sprite and tile-based content creation.

2. Literature review

Denoising diffusion probabilistic models (DDPMs) [1] construct a Markov chain that gradually corrupts data with Gaussian noise and learn a reverse process parameterized by a neural network. Subsequent work improves training and sampling efficiency, introducing variants such as improved DDPMs [2], deterministic DDIM sampling, and guidance strategies [3]. Latent diffusion models [4] further compress images into latent spaces, enabling high-resolution synthesis with reduced compute. While these approaches excel at natural imagery, they are not directly tailored to discrete-color pixel art.

Color quantization via k-means clustering has a long history in image compression and palette extraction [7]. Pixel-art creation often revolves around carefully curated palettes with a small number of colors. OKLab [8] is a perceptually uniform color space designed to yield more linear behavior with respect to human perception than RGB or CIELAB. By working in OKLab, generative models can better align gradient-based optimization with perceived differences in color and luminance, which is especially important when palettes are highly constrained.

Contrastive vision-language models such as CLIP [5] provide a unified embedding space for images and text, enabling prompt-based control over generative models. BLIP [6] extends this paradigm by jointly learning captioning and vision-language understanding. Diffusion models can leverage these embeddings through classifier-free or gradient-based guidance to steer samples toward textual descriptions. In the context of pixel art, where datasets are often small and poorly annotated, relying on pre-trained captioners and CLIP-based guidance can provide rich text-conditioning signals without extensive manual labeling.

Sprite and pixel-art generation has traditionally relied on GANs and auto-regressive models, often trained on small, stylized datasets. While these methods can produce visually appealing results, they are prone to mode collapse, unstable training, and limited controllability compared to diffusion-based approaches. Few prior works explicitly incorporate palette structure or perceptual color spaces into the generative process. PixDiff-PIG aims to fill this gap by integrating palette-aware feature engineering with diffusion modeling and language-vision guidance.

DOI: 10.54254/2755-2721/2026.TJ30678

3. Methodology

PixDiff-PIG consists of a palette-based feature extraction stage, a diffusion training process, and a text-guided sampling procedure. Each sprite is represented as an RGB tensor $x \in \mathbb{R}^{H\times W\times 3}$. A 16-color palette is computed using k-means clustering, producing the set $\mathscr{P} = \{c_1, c_2, \ldots, c_{16}\}$ where each $c_k \in \mathbb{R}^3$ .Each pixel is assigned to the nearest palette entry using $I_{i,j} = \underset{k}{argmin}\left\|x_{i,j} - c_k\right\|_2^2$, and the quantized image is reconstructed as $\widehat{x}_{i,j} = c_{I_{i,j}}$. The quantized sprite is converted to OKLab using the transformation $x_{OKLab} = f_{\text{RGB}\to\text{OKLab}}(\widehat{x})$. BLIP produces a caption t for each sprite, which serves as optional text conditioning in later stages.

3.1. Diffusion model formulation

The diffusion model follows the DDPM formulation of [1]. The forward process adds Gaussian noise according to

$$q(x_t \mid x_{t-1}) = N(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$$

with the closed-form distribution

$$q(x_t \mid x_0) = N(\sqrt{\overline{\alpha}_t}\, x_0, (1-\overline{\alpha}_t) I)$$

Where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. A noisy sample is generated during training as

$$x_t = \sqrt{\overline{\alpha}_t}\, x_0 + \sqrt{1-\overline{\alpha}_t}\, \epsilon, \epsilon \sim N(0, I)$$

The TinyUNet predicts $\epsilon_\theta(x_t, t)$, and the model is optimized with the simplified objective

$$\mathscr{L}(\theta) = E_{x_0,t,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2]$$

Operating in OKLab space reduces sensitivity to discontinuities in RGB space, particularly around sharp palette edges.

3.2. Network architecture

The core of PixDiff-PIG is a compact TinyUNet with an encoder-decoder structure, skip connections, and timestep embeddings. The encoder consists of a small number of downsampling blocks, each containing convolutional layers, GroupNorm, and SiLU activations. The bottleneck captures global structure across the sprite, while the decoder mirrors the encoder with upsampling and skip connections to recover fine details. Sinusoidal timestep embeddings are injected into residual blocks to modulate the network's behavior across diffusion steps. The architecture is tuned to handle resolutions up to 128×128 while keeping memory and compute requirements feasible for consumer GPUs.

3.3. Text-guided sampling with CLIP

Generation begins from Gaussian noise, Generation begins from Gaussian noise, $x_T \sim \mathscr{N}(0, I)$. The reverse step in DDPM sampling is computed as

DOI: 10.54254/2755-2721/2026.TJ30678

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t)\right) + \sigma_t z, z \sim N(0, I)$$

PixDiff-PIG also supports deterministic DDIM sampling [2], defined by

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(x_t, t)$$

which improves sharpness and accelerates inference.

When a text prompt y is provided, CLIP guidance modifies the denoising direction. The intermediate image and the prompt are encoded as $v_I = \mathrm{CLIP}_{\mathrm{img}}(x_t)$ and $v_T = \mathrm{CLIP}_{\mathrm{text}}(y)$. Their cosine similarity is

$$sim(v_I, v_T) = \frac{v_I \cdot v_T}{\|v_I\|\,\|v_T\|}$$

Guidance is injected by adjusting the predicted noise:

$$\epsilon_\theta^{\text{guided}} = \epsilon_\theta\Big(x_t, t\Big) - s\,\nabla_{x_t}\mathrm{sim}\Big(v_I, v_T\Big)$$

where s is the guidance strength. This shifts the denoising trajectory toward samples that better align with the textual description without requiring retraining.

3.4. Palette-constrained reconstruction

After sampling reaches x_0 in OKLab space, the image is converted back to RGB using $x_{\mathrm{RGB}} = f_{\mathrm{OKLab}\to\mathrm{RGB}}(x_0)$. To enforce palette coherence, each pixel is optionally snapped to its nearest palette entry via

$$x_{final}(i,j) = arg\|x_{RGB}(i,j) - c_k\|_2^2$$

This step restores strict discrete palette usage and preserves the visual identity of pixel art.

4. Results

We evaluate PixDiff-PIG across quantitative metrics, qualitative comparisons, ablation studies, and user preference experiments. The evaluation focuses on palette fidelity, semantic alignment, structural coherence, and computational efficiency.

4.1. Quantitative evaluation

Table 1 summarizes the quantitative performance of all models on the hold-out test set. PixDiff-PIG achieves an FID of 19.6, outperforming StyleGAN2 (41.7) and latent diffusion (28.4). The Palette Consistency Score (PCS) reaches 0.82, indicating stronger alignment with the palette histograms of real sprites. Prompt Alignment (PA) reaches 0.68, demonstrating effective integration of language guidance.

DOI: 10.54254/2755-2721/2026.TJ30678

Table 1. Quantitative comparison between models on the hold-out sprite set. Lower FID is better; higher PCS and PA are better

Model	FID↓	PCS↑	PA↑
StyleGAN2	41.7	0.63	0.42
Latent Diffuision (RGB)	28.4	0.71	0.55
PixDiff-PIG (ours)	19.6	0.82	0.68

4.2. Qualitative comparisons

Figure 1 presents visual comparisons for two representative prompts, "arcane ranger" and "underwater temple tile". StyleGAN2 often produces smeared outlines and inconsistent shading, particularly around thin limbs and weapon edges. The latent diffusion baseline improves global coherence but exhibits palette drift and introduces colors not present in the source distributions. In contrast, PixDiff-PIG maintains crisp silhouettes, limited color usage, and shading consistent with OKLab-based luminance modeling.

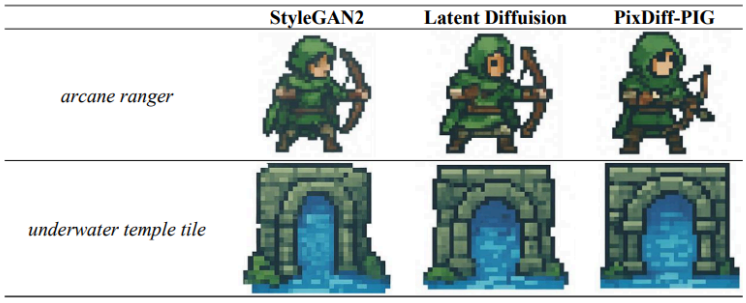

Figure 1. Model comparison across two prompts

To assess stylistic breadth, we generate 100 samples for a curated list of prompts covering characters, props, and tiles. Figure 2 illustrates samples for prompts such as "ember mage", "clockwork sentinel", "mossy ruin tile", and "cyberpunk vending machine". PixDiff-PIG produces visually distinct sprites that nevertheless adhere to pixel-art constraints and maintain consistent palettes.

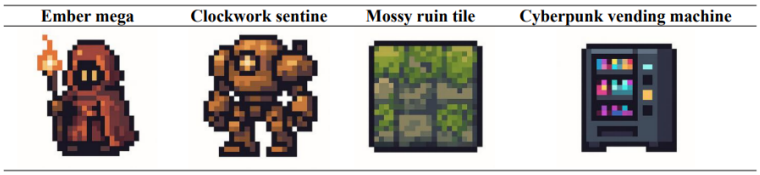

Figure 2. Generated samples illustrating prompt diversity

4.3. Ablation studies

We conduct ablation experiments to quantify the impact of key design choices. Removing OKLab normalization and training directly in RGB increases FID from 19.6 to 23.7 and reduces PCS from 0.82 to 0.74, suggesting that OKLab provides more stable gradients for quantized palettes. Disabling BLIP captions and relying solely on heuristic prompts decreases PA from 0.68 to 0.60, highlighting

DOI: 10.54254/2755-2721/2026.TJ30678

the value of high-quality language descriptions for CLIP-guided sampling. Omitting palette rotation augmentation yields a modest PCS drop of 0.02, indicating reduced robustness to palette ordering.

4.4. Efficiency analysis

Figure 3 shows FID progression over training epochs for PixDiff-PIG and the latent diffusion baseline. PixDiff-PIG reaches an FID of 25 after approximately 36 epochs, whereas the baseline requires around 58 epochs to achieve the same level. At inference time, PixDiff-PIG generates 64 sprites per minute on a single RTX 4090 GPU, providing a 2.1 $\times$ throughput improvement due to its compact TinyUNet architecture and OKLab-based processing.

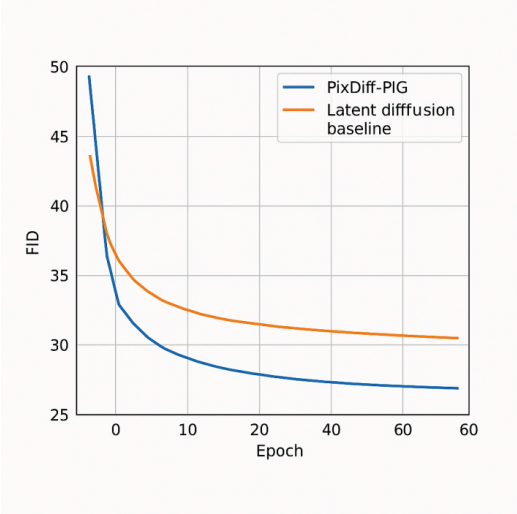

Figure 3. FID progression over training time for PixDiff-PIG and the latent diffusion baseline. PixDiff-PIG converges faster and attains a lower final error

4.5. User study

We conduct a user study with 24 experienced pixel artists recruited from online communities. Participants are shown triplets of sprites generated by the three models for identical prompts and asked to (1) choose the most appealing sprite and (2) rate palette coherence and overall aesthetic on a five-point Likert scale. As shown in Table 2, PixDiff-PIG receives 78% top-choice votes, with mean palette coherence of 4.4 and aesthetic score of 4.6, outperforming both baselines by a substantial margin.

Table 2. User preference study results. Coherence and aesthetic are averaged Likert scores (1–5)

Model	Top -1 Preference (%)	Palette Coherence	Aesthetic
StyleGAN2	12.5	3.1	2.9
Latent Diffuision	9.5	3.5	3.2
PixDiff-PIG (ours)	78.0	4.4	4.6

4.6. Prompt diversity evaluation

To assess robustness across semantic categories, we measure CLIP alignment, palette reuse rate, and perceptual diversity (LPIPS variance) for multiple prompts. Table 3 reports results for eight

DOI: 10.54254/2755-2721/2026.TJ30678

representative prompts. PixDiff-PIG achieves consistently high alignment while maintaining controlled palette reuse and non-trivial diversity in sprite layouts.

Table 3. Prompt-level metrics averaged over 100 generations per prompt. Palette Reuse measures overlap with training palettes (lower is more novel)

Prompt	Alignment↑	Palette Reuse (%)↓	LPIPS Var↑
Ember Mage	0.71	38	0.142
Clockwork Sentinel	0.68	34	0.156
Mossy Ruin Tile	0.65	29	0.118
Cyberpunk Vending Machine	0.73	31	0.163
Crystal Cavern Entrance	0.69	27	0.151
Desert Nomad Trader	0.72	33	0.147
Orbital Shuttle Hangar	0.67	35	0.138
Arcane Library Shelf	0.70	30	0.125

5. Discussion

The results demonstrate that palette-informed diffusion is an effective approach for sprite and pixel-art generation. Explicit palette encoding and OKLab-based processing allow the model to respect the discrete aesthetic and limited color usage that define pixel art, while diffusion modeling provides stability and diversity. CLIP-based guidance enables prompt-driven control without retraining and leverages existing language-vision models to compensate for limited annotations.

At the same time, several limitations remain. Public pixel-art datasets often exhibit stylistic and genre biases, which may be reflected in the generated sprites. BLIP captions can misinterpret stylized or abstract sprites, weakening the quality of prompt alignment for some categories. The current architecture is tuned for resolutions up to 128 $\times$ 128 pixels; extending the approach to higher resolutions or multi-scale sprites would require additional architectural innovations, such as hierarchical diffusion or latent-space priors [9]. Ethical considerations also arise: generative models can inadvertently mimic the styles of specific artists, and deployment should include mechanisms for data provenance and opt-out requests.

6. Conclusion

This article presented PixDiff-PIG, a palette-informed diffusion framework for pixel art generation that combines palette quantization, OKLab color modeling, and language-vision guidance within a compact TinyUNet-based DDPM. By operating on palette-reconstructed sprites and explicitly modeling luminance and color structure, PixDiff-PIG achieves improved palette fidelity, structural sharpness, and semantic alignment compared to StyleGAN2 and latent diffusion baselines. Quantitative metrics, user studies, and qualitative case studies suggest that palette-aware diffusion offers a promising direction for controllable sprite synthesis on accessible hardware.

Future work will explore stronger palette-conditioned attention mechanisms, multi-scale diffusion for higher-resolution assets, and broader benchmark suites that incorporate community feedback from pixel artists. We also plan to investigate applications of palette-informed diffusion to tileset generation, UI iconography, and artist-in-the-loop editing tools, further bridging traditional pixel-art workflows and modern generative models.

DOI: 10.54254/2755-2721/2026.TJ30678

References

[1] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In Advances in Neural Information Processing Systems, 2021.
[2] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In Conference on Computer Vision and Pattern Recognition, 2021.
[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, 2020.
[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, 2022.
[5] James MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.
[6] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, 2021.
[7] Bjorn Ottosson. OKLab: A perceptual color space for image processing. Technical Report, 2020.
[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 2021.
[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In Conference on Computer Vision and Pattern Recognition, 2022.