https://doi.org/10.2352/CIC.2025.33.1.37

Color encoding in Latent Space of Stable Diffusion Models

Guillem Arias, Ariadna Solà, Martí Armengod, Maria Vanrell
Computer Vision Center / Universitat Autònoma de Barcelona

Abstract

Recent advances in diffusion-based generative models have achieved remarkable visual fidelity, yet a detailed understanding of how specific perceptual attributes—such as color and shape—are internally represented remains limited. This work explores how color is encoded in a generative model through a systematic analysis of the latent representations in Stable Diffusion. Through controlled synthetic datasets, principal component analysis (PCA) and similarity metrics, we reveal that color information is encoded along circular, opponent axes predominantly captured in latent channels $\vec{c}_3$ and $\vec{c}_4$, whereas intensity and shape are primarily represented in channels $\vec{c}_1$ and $\vec{c}_2$. Our findings indicate that the latent space of Stable Diffusion exhibits an interpretable structure aligned with a efficient coding representation. These insights provide a foundation for future work in model understanding, editing applications, and the design of more disentangled generative frameworks.

1. Introduction

In the last few years, artificial intelligence (AI) has significantly progressed in building generative models able to create realistic images [1, 2, 3, 4]. Among these, diffusion models have emerged as one of the most effective approaches, generating high-fidelity visuals by iteratively denoising random noise through learned distributions. A particularly influential architecture for this approach is Stable Diffusion [5], which combines a latent diffusion process with a powerful image encoder-decoder pipeline.

While much attention has been given to the quality and realism of the generated content, the mechanisms by which specific visual attributes —such as color— are preserved, transformed, or encoded during generation remain less explored [6]. N. Moroney in [7] was the first to explore different color prompts and perform some assessments. One of his conclusions is the strong preexisting color association for some objects that can provide some unexpected answers for some color prompts (see Figure 1). The inability of the model to attribute color to any object brought us to try to understand how color and shape is encoded by these models.

This question has become increasingly relevant given recent findings that color remains one of the least controllable attributes in diffusion pipelines. Rafegas and Vanrell analyze discriminative CNNs and report opponent-like hue organization with increasing color–shape entanglement across layers [8], while Arias et al. reveal color deficiencies and text-over-vision bias in CLIP [9]. In contrast, we study a *generative* latent-diffusion pipeline and directly probe the *VAE latent* that precedes diffusion, using controlled color/gray datasets, PCA, and channel ablations. This exposes a channel-wise mechanism—color along opponent directions mainly in $\vec{c}_3$/$\vec{c}_4$, intensity/shape in $\vec{c}_1$ (with low-frequency structure in $\vec{c}_2$)—that complements prior discriminative/VLM analyses. Recent works demonstrate that diffusion outputs often drift away from target color distributions and proposes explicit color alignment techniques to correct them [6]. Similarly, new inpainting methods reveal that hue inconsistencies between filled regions and their surrounding context are a persistent failure mode [10]. These results highlight that reliable color handling is still not solved at the algorithmic level, motivating a deeper analysis of how color is represented at the latent level of Stable Diffusion. Moreover, color has been shown to interact with perceptual illusions in diffusion models, where emergent biases mirror phenomena observed in human vision [11]. Understanding whether Stable Diffusion organizes chromatic information along opponent axes connects modern generative models with long-established theories of efficient coding in biological vision [12, 13].

In parallel, research has explored algorithmic approaches for controllable color editing. Methods such as content-aware color editing with auxiliary restoration tasks [14], latent-guided exemplar-based recolorization [15], and initialization strategies to better preserve global image properties such as background color consistency [16] have advanced recoloring capabilities. Other work focuses on inpainting directly in latent space [17]. However, these methods treat color primarily as an external editing objective. What remains missing is a systematic account of which latent channels encode intensity versus chromatic information, and how entangled these representations are with spatial structure. This is the gap we address.

In this paper we investigate how color information is encoded in Stable Diffusion models, with a particular focus on how the image encoder projects input images into the latent space before the diffusion process begins. We aim to understand what transformations occur in the latent representations as a first step to explore color and shape encoding.

To this end, we analyze the output of the Variational Autoencoder (VAE) component used in Stable Diffusion, examining its ability to retain and reconstruct accurate color information. We also study how modifications in latent space affect color reconstruction and whether color features are encoded in disentangled or entangled representations. Previous research suggest that color attributes are tightly constrained by the realistic image distribution the model is trained on [7], making it challenging to manipulate colors as flexible, independent attributes during generation or editing.

By illuminating the mechanisms of color encoding in Stable Diffusion, this work contributes to a deeper understanding of generative AI systems and offers insights for improving their performance in applications where accurate color reproduction is essential. Our contributions on exploring the 4 channels of the latent space can be summarized as follows:

- Color presents an opponent representation based on Black-White, Cyan-Magenta, and Orange-Blue channels.

(a) `"red banana"` (b) `"pink lemon"` (c) `"purple banana"` (d) `"purple lemon"`

Figure 1. *Images generated by Stable-Diffusion for the given prompts formed by unnatural color-object pairs that probably were not part of the training datasets.*

- Chromatic information appears to be mostly in channel 3 (Cyan-Magenta) and channel 4 (Orange-Blue), although channel 4 is strongly entangled with shape.
- Shape is mostly encoded in the Black-White channel 1, but channel 2 and 4 encode some shape too.
- In channel 4 shape is combined with orange-blue chromatic information, and channel 2 seems to be encoding more low-frequency shape information.

These results confirm again the adequacy of color-opponent representation as an efficient coding when training on natural images.

2. Stable Diffusion

In this section we summarize the basic concepts of the Stable Diffusion process for image generation (see figure 2):

Encoder: The encoder is a CNN-based component derived from a Variational Autoencoder (VAE). It maps high-dimensional image data into a lower-dimensional latent space, compressing semantic and spatial information into a compact representation. The encoder learns a distribution over the latent space, enabling structured sampling during generation.

Latent Space: A compact representation where spatial and semantic information is preserved, allowing efficient and scalable diffusion with reduced computational cost compared to pixel space.

Diffusion Process: A forward Markov chain that incrementally adds Gaussian noise to the latent, transforming structured data into noise, thus modeling the image distribution.

Inverse Diffusion: A learned denoising process (typically a U-Net) that removes noise step-by-step to reconstruct the image, predicting noise components at each timestep.

Conditioning and Attention: External inputs like text or labels are injected via cross-attention layers in the U-Net, guiding generation by modulating internal representations.

Decoder: A CNN-based decoder from the VAE transforms the denoised latent back into RGB image space, ensuring photorealistic reconstruction.

3. Latent space exploration

In contrast to previous studies that primarily focus on understanding the diffusion process itself [18, 19, 20] or the prompting process [21], here we explore how information is encoded in the latent space of diffusion models. We propose two complementary

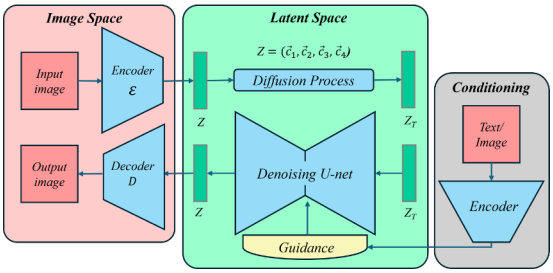

Figure 2. *Stable Diffusion Model components.*

experiments to analyze this relationship. The first experiment focuses on how color is entangled in the latent representation by using a curated set of color images sampling the color space. The second experiment investigates the influence of spatial structure by using gray-level images with varying spatial properties.

Experiment 1: Color

To analyze how color is represented in the latent space of a diffusion model, we conducted a controlled experiment using 300 uniformly colored images spanning a wide range of hues, saturations, and intensities in the HSV color space (12 hues, 6 saturations, 5 intensities). This representation was chosen for its perceptual relevance and for its ability to separate chromatic components (hue and saturation) from luminance (value), facilitating the isolation of color-related information.

Each image was passed through the model's image encoder, yielding a latent tensor of shape 4×64×64. Given that the input images were spatially uniform, we averaged activations across the spatial dimensions, obtaining a 4-dimensional latent vector per image. This allowed us to discard spatial variation and focus exclusively on how color is encoded in each latent channel.

We performed Principal Component Analysis (PCA) on the set of these 4-dimensional vectors to identify whether any principal components aligned with perceptual color attributes. The variance explained by each principal component is given by:

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.5463,\ \ 0.3172,\ \ 0.1348,\ \ 0.0018) \quad (1)$$

where, the first three components account for over 99% of the variance, indicating a compact, 3-dimensional encoding of color

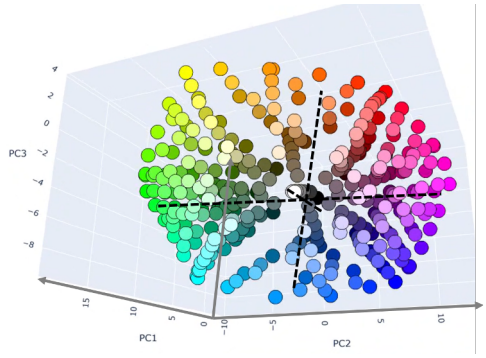

Figure 3. *Homogeneous color dataset on 3 eigen vector space.*

in the latent space. The eigen matrix, M_{eig}, is given by

$$\begin{pmatrix} \vec{e}_1 \\ \vec{e}_2 \\ \vec{e}_3 \\ \vec{e}_4 \end{pmatrix} = \begin{pmatrix} 0.572 & 0.8058 & -0.146 & -0.0466 \\ 0.1516 & -0.2781 & -0.7591 & -0.5687 \\ -0.024 & -0.0436 & -0.5917 & 0.8046 \\ -0.8058 & 0.521 & -0.2289 & -0.1641 \end{pmatrix} \quad (2)$$

we can see that PC1 is primarily associated with channels 1 and 2, while PC2 and PC3 draw most of their contribution from channels 3 and 4.

To visualize the structure of this eigen basis, we projected the latent vectors of the dataset images on the space defined by the first three principal components See figure 3). The resulting embedding confirms an organized configuration, where PC1 appears to encode intensity, with the central region corresponding to achromatic images (from black to white). More notably, as it can be seen in figure 3 the PC2 and PC3 plane exhibits a circular structure consistent with the hue wheel, with PC2 capturing a green–magenta axis and PC3 a blue–orange one. These preliminary results indicate that color information is distributed across distinct latent channels, with intensity and chromaticity encoded along separable, interpretable directions. However, as observed in subsequent experiments, these representations may still be entangled with spatial information when spatial patterns are present.

Experiment 2: Shape

Following a similar analysis, we now turn to explore the representation of shape, specifically in the absence of color. Based on earlier findings suggesting that chromatic information is concentrated in specific latent channels (channels 3 and 4), we hypothesize that shape information—especially when presented in grey-level images—could be encoded primarily in the remaining channels.

To test this hypothesis, we created a dataset composed of 80 synthetic grey-scale images depicting basic geometric shapes. These shape patterns varied systematically in frequency and orientation but with no color information. Unlike in the previous color experiment, where spatial information could be averaged out, shape representation is inherently spatial. Therefore, averaging across spatial dimensions was avoided, and PCA was not applied due to the high dimensionality of the spatial space.

Instead, we assessed the contribution of each latent channel to the reconstruction of spatial structure by selectively ablating channels and measuring the degradation in image quality. Given

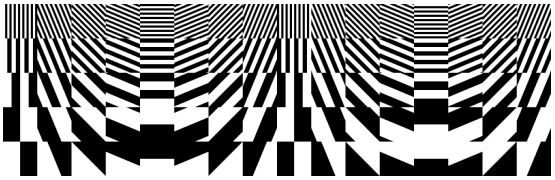

Figure 4. *Image dataset used in Exp.2 to explore shape representation.*

an input image X, we denote its encoded latent representation[1] as:

$$\mathscr{E}(X) = \vec{Z}_X = (\vec{c}_{1,X}, \vec{c}_{2,X}, \vec{c}_{3,X}, \vec{c}_{4,X}) \quad (3)$$

To estimate the importance of a specific channel (e.g., channel 3), we reconstructed the image after zeroing out that channel and computing the similarity between the original and the reconstructed image, this is

$$\text{Similarity}(X, \mathscr{D}((\vec{0}, \vec{0}, \vec{c}_{3,X}, \vec{0}))) \quad (4)$$

we also compute a baseline reconstruction similarity using the full latent code,

$$\text{Similarity}(X, \mathscr{D}(\mathscr{E}(X))) = \text{Similarity}(X, \vec{Z}_X) \quad (5)$$

To quantify image similarity, we primarily use the Structural Similarity Index (SSIM) [22], a perceptually motivated metric that correlates well with human judgment of image quality. To confirm the results, we also computed the Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR). By comparing the reconstruction quality across different combinations of retained channels (individual, pairwise, and triplet combinations), we characterized the relative importance of each latent channel for shape representation.

4. Results

In next sections we report all the results of the above mentioned experiments.

Color Representation

In figure 3 we already outlined the color configuration in the latent space. To further investigate the role of the first principal component (PC1), we analyze its relationship with the average image intensity of the dataset. As shown in Figure 5, there is a clear dependency that presents a Pearson's correlation index of 0.72[23].

Analysis of the component loadings reveals that PC1 is primarily influenced by latent channels $\vec{c}_1$ and $\vec{c}_2$, indicating that intensity-related information is predominantly encoded in these two channels.

To further investigate the representation of chromatic information in the latent space, we focused on the encoding of hue along the second and third principal components (PC2 and PC3), which primarily correspond to channels 3 and 4. To assess the distribution of hue within this subspace, we analyzed the angular relationship between PC2 and PC3 and compared it to the hue angle of the original images.

As illustrated in Figure 5, the analysis reveals a close correspondence between the structure of the latent space and the hue

[1] We use vectorial representation for images in latent space.

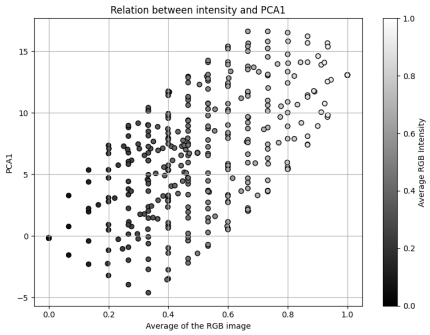

Figure 5. *Correlation between the first principal component (PC1) versus the mean RGB intensity of each input image.*

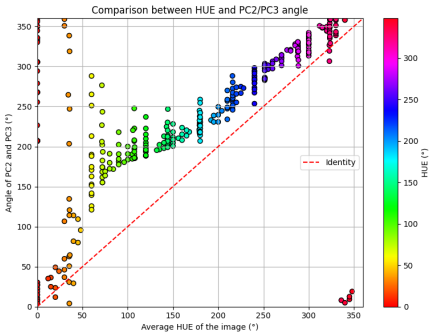

Figure 6. *Correlation plot between the hue of the original images and the angle formed by the second and third principal components in latent space.*

values, indicating that hue is encoded in a directionally organized manner within the PC2-PC3 plane. This result supports the interpretation that hue is embedded in a circular latent structure, wherein variations in color are systematically aligned with angular changes in the projection space, with a circular correlations index of -0.69 (Jammalanamdak-Sarma [12]).

To visualize the functional role of each latent channel to both image and PCA latent space, we built a color wheel stimulus with constant intensity and varying hue. This controlled setting isolates chromatic variation and allows a targeted examination of the latent encoding. As shown in Fig. 8, the first row displays decoded images obtained by projecting the original image X onto each of the first four principal components. Rows 2–5 illustrate reconstructions where only a single latent channel is active, enabling an assessment of the individual contributions of $\vec{c}_1$ through $\vec{c}_4$. The bottom-left cell ($D(\vec{0})$) represents the decoding of a zero vector.

The reconstruction from PC4 confirms its negligible contribution ($\lambda_4 = 0.018$), as no meaningful structure is observed. In contrast, PC1 captures intensity-related structure and is primarily composed of channels $\vec{c}_1$ and $\vec{c}_2$, as shown by their respective single-channel reconstructions. Channels $\vec{c}_3$ and $\vec{c}_4$ do not contribute to PC1, confirming their lack of involvement in intensity encoding. PC2 reveals an opponent axis corresponding to Magenta–Green, predominantly encoded in channel $\vec{c}_3$. PC3 encodes a second opponent axis, Blue-Orange, primarily represented in channel $\vec{c}_4$, with some contribution from $\vec{c}_3$. These axis would correlate with the peaks of the color distribution of the dataset

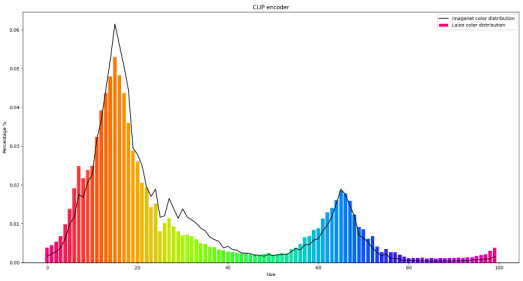

Figure 7. *Hue distribution of the Dataset Laion 2B used to train Stable Diffusion model.*

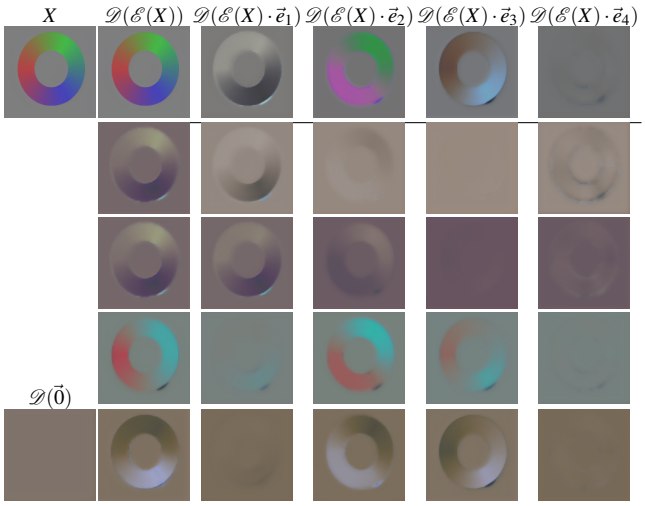

Figure 8. *Latent space color representation for the input image X, which varies uniformly in hue while maintaining constant intensity. The second column shows the full latent decoding $\mathscr{D}(\mathscr{E}(X))$. Columns 3 to 6 show reconstructions of X projected onto each of the first four principal components. Rows 2 to 5 (except column 1) display reconstructions using only a single active latent channel. For example, row 2, column 3 shows $\mathscr{D}(\vec{c}_1,\vec{0},\vec{0},\vec{0})$. The bottom left image shows the decoding of a zero vector $\vec{0}$, representing the average color in the training set.*

where the model was trained on, which is given in Figure7.

Overall, the analysis highlights a latent space where intensity is mainly encoded in channels $\vec{c}_1$ and $\vec{c}_2$, and chromaticity is captured across channels $\vec{c}_3$ and $\vec{c}_4$, reflecting a circular, opponent-based chromatic representation.

Shape Representation

In Table 1 we present the results of our similarity metrics calculated on the grey-scale shape dataset. The left side of the table displays SSIM, PSNR, and MSE values for all possible combinations of active latent channels, along with the percentage of structural information recovered relative to the minimum and maximum cases. The right side of the table organizes the same results based on the number of active channels per combination.

To define the reference points for recovery percentages, we treat the decoded output using all four latent channels as the maximum reference (100% SSIM recovery), and the output with all channels zeroed out as the minimum (0%). As expected, the

Channels	SSIM (%)	PSNR	MSE↓
$[\vec{0},\vec{0},\vec{0},\vec{0}]$	**0.3522 (0%)**	6.05	0.2481
$[\vec{c}_1,\vec{c}_2,\vec{c}_3,\vec{c}_4]$	**0.4855 (100%)**	11.42	0.0721
$[\vec{0},\vec{0},\vec{c}_3,\vec{0}]$	0.3528 (4.34%)	12.70	0.0617
$[\vec{0},\vec{c}_2,\vec{0},\vec{0}]$	0.3581 (4.43%)	8.74	0.1338
$[\vec{0},\vec{0},\vec{0},\vec{c}_4]$	0.4263 (55.59%)	9.10	0.1232
$[\vec{c}_1,\vec{0},\vec{0},\vec{0}]$	0.4478 (71.72%)	9.24	0.1192
$[\vec{0},\vec{c}_2,\vec{c}_3,\vec{0}]$	0.3621 (7.43%)	8.49	0.1417
$[\vec{0},\vec{0},\vec{c}_3,\vec{c}_4]$	0.4285 (57.24%)	8.95	0.1273
$[\vec{c}_1,\vec{0},\vec{c}_3,\vec{0}]$	0.4427 (67.89%)	8.50	0.1413
$[\vec{0},\vec{c}_2,\vec{0},\vec{c}_4]$	0.4437 (68.64%)	10.66	0.0860
$[\vec{c}_1,\vec{0},\vec{0},\vec{c}_4]$	0.4716 (89.57%)	10.78	0.0835
$[\vec{c}_1,\vec{c}_2,\vec{0},\vec{0}]$	0.4721 (89.95%)	10.64	0.0864
$[\vec{0},\vec{c}_2,\vec{c}_3,\vec{c}_4]$	0.4469 (71.04%)	10.71	0.0850
$[\vec{c}_1,\vec{c}_2,\vec{c}_3,\vec{0}]$	0.4724 (90.17%)	10.44	0.0904
$[\vec{c}_1,\vec{0},\vec{c}_3,\vec{c}_4]$	0.4729 (90.55%)	10.68	0.0856
$[\vec{c}_1,\vec{c}_2,\vec{0},\vec{c}_4]$	0.4827 (97.9%)	11.40	0.0725

Table 1. Similarity metrics between original images and their reconstructions using all combinations of latent channels, sorted by SSIM.

Low Frequency		High Frequency	
Channels	SSIM (%)	Channels	SSIM (%)
$[\vec{0},\vec{0},\vec{0},\vec{0}]$	0.3669 (0.00)	$[\vec{0},\vec{c}_2,\vec{0},\vec{0}]$	0.2506 (0.00)
$[\vec{0},\vec{0},\vec{c}_3,\vec{0}]$	0.3702 (2.92)	$[\vec{0},\vec{c}_2,\vec{c}_3,\vec{0}]$	0.2605 (3.86)
$[\vec{0},\vec{c}_2,\vec{c}_3,\vec{0}]$	0.3985 (27.96)	$[\vec{0},\vec{0},\vec{c}_3,\vec{0}]$	0.2898 (15.28)
$[\vec{0},\vec{c}_2,\vec{0},\vec{0}]$	0.3991 (28.49)	$[\vec{0},\vec{0},\vec{0},\vec{0}]$	0.2975 (18.28)
$[\vec{0},\vec{0},\vec{c}_3,\vec{c}_4]$	0.4396 (64.34)	$[\vec{0},\vec{0},\vec{0},\vec{c}_4]$	0.3717 (47.21)
$[\vec{0},\vec{0},\vec{0},\vec{c}_4]$	0.4396 (64.34)	$[\vec{0},\vec{0},\vec{c}_3,\vec{c}_4]$	0.3856 (52.63)
$[\vec{c}_1,\vec{0},\vec{c}_3,\vec{0}]$	0.4514 (74.78)	$[\vec{c}_1,\vec{0},\vec{c}_3,\vec{0}]$	0.4095 (61.95)
$[\vec{0},\vec{c}_2,\vec{0},\vec{c}_4]$	0.4556 (78.50)	$[\vec{c}_1,\vec{0},\vec{0},\vec{0}]$	0.4098 (62.07)
$[\vec{c}_1,\vec{0},\vec{0},\vec{0}]$	0.4569 (79.65)	$[\vec{0},\vec{c}_2,\vec{0},\vec{c}_4]$	0.4129 (63.27)
$[\vec{0},\vec{c}_2,\vec{c}_3,\vec{c}_4]$	0.4574 (80.09)	$[\vec{0},\vec{c}_2,\vec{c}_3,\vec{c}_4]$	0.4225 (67.02)
$[\vec{c}_1,\vec{c}_2,\vec{c}_3,\vec{0}]$	0.4702 (91.42)	$[\vec{c}_1,\vec{0},\vec{0},\vec{c}_4]$	0.4645 (83.39)
$[\vec{c}_1,\vec{c}_2,\vec{0},\vec{0}]$	0.4705 (91.68)	$[\vec{c}_1,\vec{0},\vec{c}_3,\vec{c}_4]$	0.4726 (86.55)
$[\vec{c}_1,\vec{0},\vec{c}_3,\vec{c}_4]$	0.4729 (93.81)	$[\vec{c}_1,\vec{c}_2,\vec{0},\vec{0}]$	0.4770 (88.27)
$[\vec{c}_1,\vec{0},\vec{0},\vec{c}_4]$	0.4729 (93.81)	$[\vec{c}_1,\vec{c}_2,\vec{c}_3,\vec{0}]$	0.4804 (89.59)
$[\vec{c}_1,\vec{c}_2,\vec{0},\vec{c}_4]$	0.4785 (98.76)	$[\vec{c}_1,\vec{c}_2,\vec{0},\vec{c}_4]$	0.4983 (96.57)
$[\vec{c}_1,\vec{c}_2,\vec{c}_3,\vec{c}_4]$	0.4799 (100.0)	$[\vec{c}_1,\vec{c}_2,\vec{c}_3,\vec{c}_4]$	0.5071 (100.0)

Table 2. Comparison of SSIM and relative improvement (%) for different channel combinations, separated by low and high frequency images.

complete reconstruction (all channels active) achieves the highest SSIM value (0.4855), while the zero-latent input yields the lowest (0.3522).

We begin by analyzing the individual contributions of each latent channel. Channel $\vec{c}_1$ is the most informative when used alone, recovering 71.72% of the total possible SSIM. This supports our hypothesis that structural (shape) information is strongly encoded in this channel. In contrast, channel $\vec{c}_3$, which was previously associated with color encoding, contributes the least to the reconstruction of grey-scale images. This is consistent with the findings in Section . Channel $\vec{c}_2$ performs only slightly better, suggesting limited involvement in shape encoding. Interestingly, channel $\vec{c}_4$ recovers over 55% of the structural information despite its established role in color representation, suggesting a dual encoding role.

When two channels are combined, the combination of $\vec{c}_3$ and $\vec{c}_2$ fall below 10%. However, the rest of pairs when $\vec{c}_1$ or $\vec{c}_4$ are present, the percentages grow significantly. Nonetheless, it is not a linear growing, since $[\vec{c}_1,\vec{c}_2,\vec{0},\vec{0}]$ achieves the maximum 89.95% SSIM recovery over the combination of $\vec{c}_1$ and $\vec{c}_4$.

Three-channel combinations reveal similar trends. Combinations lacking $\vec{c}_1$ show a significant reduction of similarity. For instance, $[\vec{0},\vec{c}_2,\vec{c}_3,\vec{c}_4]$ recovers only 71.04%, whereas $[\vec{c}_1,\vec{c}_2,\vec{0},\vec{c}_4]$ achieves 97.9%. These results consistently point to $\vec{c}_1$ as the principal carrier of shape-related information in the latent space.

Previous results, make us to conclude that channel $\vec{c}_2$ is the more unexplained. We try to further analyze it considering its behaviour under different frequency conditions. Table 2 show the results of analysising for low- and high-frequency images separately. In low-frequency cases, activations involving $\vec{c}_2$ contribute notably to structural recovery—up to 28.5% improvement when paired with $\vec{c}_3$. However, in high-frequency cases, the same combinations show minimal contribution (0–3.8%). This suggests that $\vec{c}_2$ may play a more active role in encoding low-frequency shape information, that would be more global shape information than image details.

Moreover, combinations of $\vec{c}_2$ with $\vec{c}_4$ are particularly effective in low-frequency settings, recovering 78–80% SSIM, but are less impactful in high-frequency contexts (63–67%). Finally, combinations including both $\vec{c}_1$ and $\vec{c}_2$ consistently produce high recovery scores in both frequency bands, exceeding 88% in all cases. These findings suggest that while shape information is concentrated in $\vec{c}_1$, there can be some structures that can be distributed across other channels, particularly $\vec{c}_2$ and $\vec{c}_4$.

In summary, the results confirm a partial specialization of latent channels: $\vec{c}_1$ is primarily responsible for encoding shape; $\vec{c}_3$ is mostly color-related, but with some color information in $\vec{c}_4$. This last one, jointly with $\vec{c}_2$ appear to play more nuanced roles in shape representation aspects.

Conclusions

This work presents a detailed analysis of how color and shape are encoded in the latent space of Stable Diffusion models. Through controlled experiments on the latent representations of specific datasets. We extract the following key findings. Color presents an opponent representation in the latent space based on 3 main channels: Black-White, Magenta-Green and Blue-Orange. The 4 channels of the latent space exhibit a partially disentangled representation that can briefly be summarized as: $\vec{c}_1$ carries the bulk of intensity and shape information, $\vec{c}_2$ contributes to low-frequency structures and complements $\vec{c}_1$ with shape, $\vec{c}_3$ is a pure chromatic channel, and $\vec{c}_4$, while chromatic in nature, also exhibit varying degrees of spatial entanglement.

With this work we are just giving some preliminary steps towards the understanding of color and shape representation, further research is required to condition their generation. The results presented here demonstrate the feasibility of precise control over visual attributes such as hue, saturation, and shape by selectively manipulating latent channels. They also point to the limitations in achieving fully disentangled representations using standard training procedures on natural image data.

Acknowledgments

This work was partially supported by grants PID2021-128178OB-I00, PID2024-162555OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU, and by the Generalitat de Catalunya — Departament de Recerca i Universitats with reference 2021SGR01499 and CERCA Program.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

[2] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2016.

[3] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[4] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha G. Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. `arXiv: 2205.11487`, 2022.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[6] Ka Chun Shum, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Color alignment in diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[7] Nathan Moroney. Color terms and stable diffusion. *Color and Imaging Conference*, 32(1):84–84, 2024.

[8] Ivet Rafegas and Maria Vanrell. Color representation in cnns: Parallelisms with biological vision. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2697–2705, 2017.

[9] Guillem Arias, Ramon Baldrich, and Maria Vanrell. Color in visual-language models: Clip deficiencies. *Color and Imaging Conference, pp 101 - 106*, 2024.

[10] Yikai Wang, Chenjie Cao, Junqiu Yu, Ke Fan, Xiangyang Xue, and Yanwei Fu. Towards enhanced image inpainting: Mitigating unwanted object insertion and preserving color consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[11] Alexandra Gomez-Villa, Kai Wang, C. Alejandro Párraga, Bartłomiej Twardowski, Jesús Malo, Javier Vazquez-Corral, and Joost van de Weijer. The art of deception: Color visual illusions and diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18642–18652, 2025.

[12] S. Rao Jammalamadaka and Y. Sarma. A correlation coefficient for angular variables. In *Statistical Theory and Data Analysis II*, pages 349–364. Elsevier, 1988.

[13] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[14] Yixuan Ren, Jing Shi, Zhifei Zhang, Yifei Fan, Zhe Lin, Bo He, and Abhinav Shrivastava. Content-aware image color editing with auxiliary color restoration tasks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[15] Wenjie Yang, Ning Xu, and Yifei Fan. Latent-guided exemplar-based image re-colorization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[16] Jeffrey Zhang, Shao-Yu Chang, Kedan Li, and David Forsyth. Preserving image properties through initializations in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[17] Ciprian Corneanu, Raghudeep Gadde, and Aleix M. Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[18] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15098 of *Lecture Notes in Computer Science*, pages 172–188. Springer, 2024.

[19] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, and Wei Liu. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024.

[20] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023.

[21] Deepak Sridhar and Nuno Vasconcelos. Prompt sliders for fine-grained control, editing and erasing of concepts in diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2024.

[22] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[23] Karl Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.