Enhancing Creative Generation on Stable Diffusion-based Models

Jiyeon Han* [1], Dahee Kwon* [1], Gayoung Lee [2], Junho Kim† [2], and Jaesik Choi† [1,3]
[1]KAIST AI [2] NAVER AI Lab [3]INEEJI
{j.han, daheekwon, jaesik.choi}@kaist.ac.kr, {gayoung.lee, jhkim.ai}@navercorp.com

Figure 1. **Original** vs **C3 (Ours)**. Compared to the original diffusion models, Our C3 consistently generates more creative images with no added computational cost. Code is available at https://github.com/daheekwon/C3.

Abstract

Recent text-to-image generative models, particularly Stable Diffusion and its distilled variants, have achieved impressive fidelity and strong text-image alignment. However, their creative capability remains constrained, as including 'creative' in prompts seldom yields the desired results. This paper introduces C3 (Creative Concept Catalyst), a training-free approach designed to enhance creativity in Stable Diffusion-based models. C3 selectively amplifies features during the denoising process to foster more creative outputs. We offer practical guidelines for choosing amplification factors based on two main aspects of creativity. C3 is the first study to enhance creativity in diffusion models without extensive computational costs. We demonstrate its effectiveness across various Stable Diffusion-based models.

*Equally contributed.
†Corresponding authors.

1. Introduction

In recent years, the field of text-to-image generation has witnessed remarkable progress, marked by the development of models capable of producing high-fidelity images that align closely with user-specified text prompts, enabling applications across various fields, including art, entertainment, design, and research [27, 29]. Notably, Stable Diffusion [17, 20] and its distilled variants [12, 21] have emerged as powerful tools, delivering impressive image quality and semantic consistency. As the capabilities of these models continue to expand, so too does the interest in exploring the boundaries of their generative potential, particularly concerning the generation of creative and novel content. This practical pursuit of creativity in generative models has raised critical questions: to what extent can these models foster creativity, and how might we enhance their creative capabilities in an efficient, user-friendly manner?

Despite their prowess, Stable Diffusion-based generative

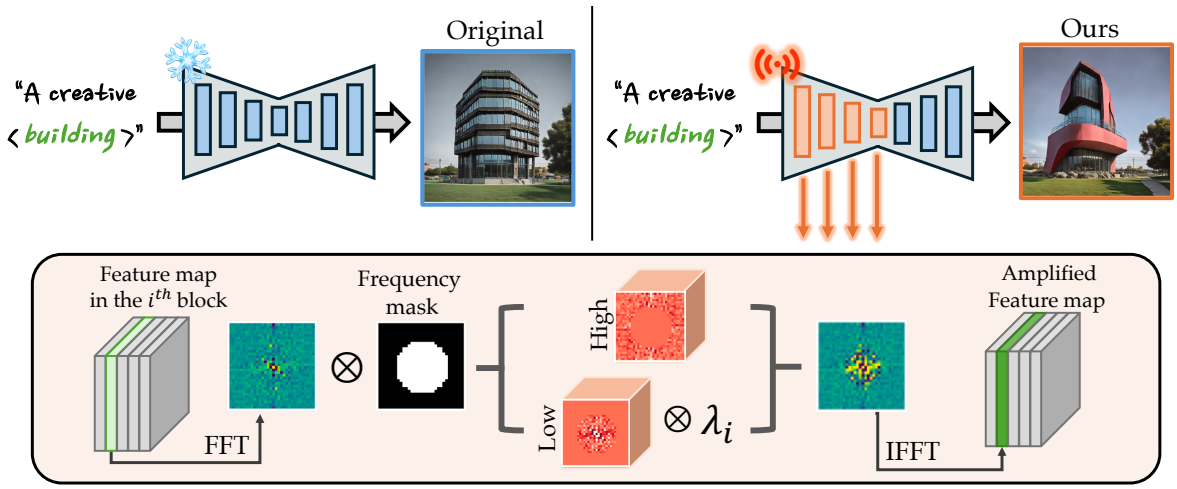

Figure 2. An overview of the proposed **C3** algorithm. We selectively amplify the low-frequency feature of the shallow blocks to enhance creative generations of the pretrained diffusion models.

models struggle to effectively produce creative images. As Figure 1 and our user study in Table 2 illustrate, adding the "creative" term to prompts often fails to yield satisfying creative variations, indicating limitations in the models' creative flexibility. While generating creative outputs is a crucial aspect of generative modeling, it remains relatively underexplored. Existing frameworks that aim to improve the novelty of images from diffusion-based models [13, 19, 25] typically rely on additional optimization steps or user-defined reference images, making them computationally expensive for scalable use.

To address this gap, we propose a simple yet effective training-free approach, C3 (Creative Concept Catalyst), specifically designed to elevate the creative capabilities of Stable Diffusion-based models. Inspired by pioneering work in feature manipulation [3, 6, 11, 23, 26], C3 promotes the generation of creative visual concepts by directly amplifying feature maps within the denoising process. In this way, we can bypass the need for reference images. As will be analyzed in this paper, we observe that different blocks contribute variably to creativity-related image generation, with shallower blocks playing a more significant role in producing novel visuals. Building on this insight, we amplify features at shallower blocks to promote more creative outputs. However, consistent amplification of intermediate features may cause unwanted noise. To address this, we shift the intermediate features into the Fourier domain and selectively manipulate the low-frequency components, effectively boosting creativity while minimizing noise. Also, we offer practical guidelines for selecting appropriate amplification factors tailored to two core aspects of creativity.

To the best of our knowledge, this paper is the first to propose a method for enhancing the creativity of diffusion models without additional optimizing steps. We empirically validate the efficacy of C3 across diverse objects and Stable Diffusion-based models, demonstrating its impact and utility in creative image generation.

2. Related Work

Research on achieving creative generations in generative models has been continuously advancing. Based on GANs, creative generations are encouraged by employing contrastive loss or diversity loss from existing categories or samples [5, 16, 22]. Recent advances in generative modeling have aimed to balance creativity with diversity in image generation, focusing on approaches that allow inspiration from existing concepts without direct replication. ProCreate [13], an energy-based approach, proposes guiding diffusion model outputs away from reference images in the latent space, thus improving diversity and concept fidelity in few-shot settings. This method prevents training data replication and has enhanced sample creativity across various artistic styles and categories. On the other hand, Inspiration Tree [25] introduces a structured decomposition of concepts, where a hierarchical tree structure captures different visual aspects of a given concept. Adding to this line of creative generative techniques, ConceptLab [19] leverages a Vision-Language Model (VLM) with diffusion priors to further push the boundaries of novel concept generation within broad categories. By iteratively applying constraints that differentiate generated concepts from existing category members, ConceptLab enhances the creation of unique, never-before-seen concepts, enabling hybridization and exploration within a given category. While these approaches represent advancements in generating creatively inspired outputs, they require burdensome additional training or optimization. An overview of Stable Diffusion-based

Figure 3. Block-wise feature amplification results. All frequency bands are amplified.

Figure 4. **(Top)** Uniform amplification across all frequency-band features in the first down block. **(Bottom)** Amplification of low-frequency features. Enhancing only low-frequency features helps eliminate noise and mosaic patterns.

models and feature map manipulation research is provided in Appendix G.

3. Methodology

3.1. Motivation

Stable Diffusion models [17, 20] are widely adopted for their efficient text-to-image generation capabilities, supported by openly accessible checkpoints. Their distilled variants, Turbo [21] and Lightning [12] are optimized specifically for faster sampling. They share a U-Net backbone with three down blocks, a middle block, and three up blocks to generate latent noise. For users aiming to create novel and creative images, a straightforward approach is to include the word "creative" in the prompt. However, as shown in Figure 1, this naive approach proves ineffective across all models.

Our objective in this paper is to address these limitations and enhance the creative generation capacity of Stable Diffusion-based models. Inspired by feature manipulation methods, we amplify feature maps given a creativity-specified text prompt. If not mentioned otherwise, we use "a creative [obj]" for the text prompt. As depicted in Figure 3, our empirical analysis demonstrates that each block contributes differently to creativity. Amplifying the first and second blocks primarily induces color and structure changes, while the third down block and middle block impact attributes associated with texture and shape. In contrast, manipulating up blocks mainly affects properties like noise, blur, and contrast. These findings guide our approach: by focusing on the down and middle blocks, we aim to enhance the generation of consistently creative images through targeted feature manipulation.

3.2. Creative Concept Catalyst (C3)

Here, we introduce a simple yet effective method, C3 (Creative Concept Catalyst), designed to enhance creative image generation without additional training steps. The method overview is depicted in Figure 2.

At its core, C3 works by amplifying the internal feature maps in three down blocks and a single middle block within the U-Net. While uniformly amplifying all feature values has the potential to make the image more creative, it often introduces noise and a colorful tile pattern as side effects (See Figure 4-(Top)). We presume this mainly happens due to the amplification of high-frequency details. In image processing, it is well understood that low-frequency components relate to the main content or objects in an image, while high-frequency components capture finer details. Based on this insight, we selectively amplify the low-frequency components in the frequency domain.

Let x_l denote the output feature maps of the l-th block, and $f(x_l) = \text{FFT}(x_l)$ represent the feature maps transformed into the frequency domain. To isolate the low- and high-frequency components, we apply a binarized low-frequency mask M_L with a specified cut-off threshold. The cut-off threshold defines the range of low-frequency components: a higher threshold creates a broader boundary for low frequencies, allowing more extensive modifications in the image, even affecting some finer details. A more detailed analysis of the cut-off threshold, along with our empirical guidelines, is provided in Appendix B. We obtain the low-frequency component $f_L(x_l)$ by multiplying the low-frequency mask with the transformed features $f(x_l)$ element-wise.

$$f_L(x_l) = f(x_l) \odot M_L \tag{1}$$
$$f_H(x_l) = f(x_l) \odot (1 - M_L) \tag{2}$$

We amplify the obtained low-frequency components of the feature maps with an amplification factor λ_l^* while preserving the high-frequency components. This technique effectively produces a clear but more creatively enhanced object without introducing noise, as illustrated in Figure 4-(Bottom). The processed features x_l^*, transformed back into the spatial domain using the inverse Fourier transform, then serve as the input of the $(l+1)$-th block in the U-Net.

$$f^*(x_l) = \lambda_l^* \cdot f_L(x_l) + f_H(x_l) \tag{3}$$
$$x_l^* = \text{IFFT}(f^*(x_l)) \tag{4}$$

Figure 5. The image generated with the automatically selected amplification factors for each block and the combined amplification of all blocks.

3.3. Parameter Selection

When the amplification factor λ_l is too small, the image does not change, and when λ_l is too large, the image diverges to an unrecognizable noise. Further, the proper amplification factor λ_l varies for each block as the later blocks are less sensitive to the change. Here, we propose an automatic way to find a proper amplification factor for each block based on the core aspects of creativity.

Creativity is known to be recognized in two main aspects: usability and novelty [2, 14]. A creative sample should not only be novel within the population but also meet a certain standard of quality. In the image domain, usability may consider both whether the content of the image can be recognized as the target object and whether the image quality is satisfactory. For instance, while the images in the last column of Figure 3 can be identified as the target objects, the noise in the image diminishes their usability. An aesthetic score evaluates image quality based on human perception, potentially considering factors such as color harmony, global layout, and the rule of thirds. On the other hand, the CLIP score [7] calculates the similarity between the image embedding and the text embedding in the same space, evaluating how well the image aligns with the given text prompt. We then define the usability score of an image as follows,

$$\text{Use}(I) = Aesthetic(I) + CLIP(I, c) \tag{5}$$

for a generated image I and the text prompt c. We use a publicly available aesthetic score predictor[1] trained on a large-scale database for aesthetic visual analysis (AVA) [15].

Assessing the novelty of an image poses more significant challenges, as it requires consideration of all potential outcomes within the population. Rather than measuring novelty directly, we rely on the intuition that novelty increases monotonically with the amplification factor. We then search for the maximum λ_l under the usability constraint such that it maximizes novelty while maintaining an acceptable level of quality.

$$\lambda_l^* = \max \lambda_l^i \quad \text{for } \lambda_l^i \in \Lambda_l, \\ \text{s.t., Use}(I(\lambda_l^i)) \geq \epsilon \cdot \text{Use}(I(\lambda_l^0)). \tag{6}$$

Here, $I(\lambda)$ represents the image generated with λ amount amplified feature and $\Lambda_l = \{\lambda_l^i | 1 = \lambda_l^0 < \lambda_l^i < \lambda_l^{i+1} < ... < \lambda_l^n = K_l\}$, where λ_l^0 represents no amplification and K_l denotes the maximum amplification. The usability bumper $0 \leq \epsilon \leq 1$ controls the trade-off between usability and novelty—the larger ϵ results in higher fidelity at the expense of novelty and vice versa. For the sake of computational efficiency, we empirically assigned different values for K_l to each block. Typically, we set $K_0 = K_1 = 2$ and $K_2 = K_3 = 10$. Amplification factors larger than these values likely result in noisy images. Nevertheless, it is also feasible to set $K_0 = K_1 = ... = K_L$. When amplifying multiple blocks simultaneously, the amplification factors need additional scaling to maintain the image quality. Our empirical findings indicate that the quality is maintained as long as the sum of the scaling factors is preserved. By default, maintaining the sum to 1 for Turbo and 0.6 for the rest of the models generally yields satisfactory results. The detailed analysis of the scaling factors is in Appendix B. The resulting images with feature amplification, using automatically determined amplification factors for each block, along with images amplified across all down and middle blocks, are presented in Figure 5.

4. Experimental Results

4.1. Settings

This section evaluates whether the proposed C3 method can generate sufficiently creative images across various Stable Diffusion-based models and object types. We selected four base models for this analysis: SDXL [17] and its distilled versions, Lightning [12] (1-step and 4-step) and Turbo [21]. We also included ConceptLab [19] as another baseline. Like our approach, ConceptLab generates creative images without requiring reference images. It uses Kandinsky [18] as the backbone architecture and incurs additional training costs. Detailed descriptions of the various hyperparameter settings are provided in Appendix A. Additionally, we provide an analysis of the step-wise effects of C3 at different stages in Appendix B, while the default setting applies C3 across all denoising steps.

4.2. Quantitative Results

In this section, we quantitatively analyze the performance of the proposed method across five distinct objects: chair, building, garment, car, and teddy bear. We use 100 uncurated images for each object. We categorize the evaluation metrics into three groups: novelty, diversity, and usability.

[1] https://github.com/discus0434/aesthetic-predictor-v2-5

Model	Method	Novelty		Diversity			Usability	
		FID* (↑)	Precision* (↓)	Recall (↑)	LPIPS (↑)	Vendi (↑)	CLIP (↑)	BLIP (↑)
Lightning (1-step)	Orig	123.95±44.94	0.89±0.08	0.69±0.24	0.26±0.10	5.31±1.83	**0.27±0.02**	**0.97±0.04**
	Ours	**163.01±58.01**	**0.65±0.20**	**0.79±0.11**	**0.34±0.07**	**6.24±2.15**	0.27±0.01	0.89±0.06
Turbo	Orig	146.43±54.48	0.87±0.06	0.27±0.17	0.22±0.06	3.54±1.31	**0.27±0.02**	**1.00±0.00**
	Ours	**164.07±55.47**	**0.51±0.21**	**0.68±0.18**	**0.36±0.09**	**4.93±1.76**	0.27±0.02	0.95±0.06
Lightning (4-step)	Orig	117.32±39.24	0.86±0.05	**0.92±0.05**	0.28±0.08	5.02±2.01	**0.27±0.02**	**0.99±0.01**
	Ours	**154.91±49.36**	**0.55±0.13**	0.83±0.09	**0.35±0.05**	**6.28±2.18**	0.26±0.02	0.92±0.06
SDXL	Orig	128.20±41.90	0.79±0.13	**0.96±0.01**	0.24±0.05	6.52±2.01	**0.27±0.02**	**0.95±0.04**
	Ours	**157.93±38.37**	**0.66±0.16**	0.93±0.04	**0.32±0.04**	**7.32±1.91**	0.27±0.02	0.86±0.07
Real-to-Ref	-	248.13±35.28	0.56±0.31	0.56±0.21	-	-	-	-
ConceptLab	-	251.27±61.29	0.65±0.20	0.64±0.18	0.37±0.02	8.41±1.26	0.25±0.02	0.32±0.30

Table 1. Quantitative results averaged over five objects. The Real-to-Ref row employs the prompt "a [ref-obj]" to generate reference fake samples. FID* and Precision* scores are interpreted in opposition to their conventional usage, as our method aims to generate novel samples distinct from the normal ones. **Bold** indicates the best for each metric.

FID computes the distributional similarity between real and fake datasets. Precision and recall measure the proportion of fake data within the real manifold and the proportion of real data within the fake manifold, respectively. In place of the real data, we use the images generated from SDXL with the text prompt "a [obj]".

Generally, a lower FID score indicates better performance, and a higher precision is preferred. However, in this study, we aim for outcomes that deviate from normal generations. Consequently, we interpret FID and precision in the opposite manner, where a higher FID is considered better, and a lower precision is favored. We establish reference values for FID and precision to prevent misleading results, which are computed on a separate set of samples generated with the prompt "a [ref-obj]". The reference objects are selected to belong to the same category as the target object while being significantly distinguishable from it. The used reference objects are sofa (↔ chair), monument (↔ building), scarf (↔ garment), bus (↔ car), and bunny doll (↔ teddy bear). Recall indicates how many modes are covered by the generated data, considering each real as each mode. LPIPS measures the perceptual distance between two images using image features extracted from a pre-trained backbone. In our analysis, we measure the average LPIPS between all pairs of generated images. The Vendi score measures the Shannon entropy of the eigenvalues of the similarity matrix between the generated images and can be interpreted as the number of effective modes. For usability, we use the CLIP score between the generated image and the text prompt "a creative [obj]" used for the generation. The BLIP score measures the portion of generated samples that receive a 'yes' response from the BLIP VQA model when asked, "Is this image [obj]?"

Table 1 summarizes the results. We average each score across five objects. The scores for each object are listed in Appendix C. For the novelty, both FID and Precision values are improved compared to the original outcomes. FID of ConceptLab is larger than ours, however, it exceeds the FID of the reference object, which means the result may be regarded as a different object. Regarding diversity, our metrics have shown improvements in general. The recall values for Lightning (4-step) and SDXL have experienced a slight decline but remain at satisfactory levels. Conversely, Turbo and Lightning have significantly enhanced recall scores, particularly the Turbo model, which has been notably affected by mode collapse. Considering the improvements of our method in terms of novelty and diversity, the loss in usability metrics is tolerable compared to ConceptLab.

4.3. Qualitative Results

We demonstrate the capacity of C3 to generate diverse creative concepts across multiple categories. Our results are compared with those of the original Stable Diffusion-based models and ConceptLab [19]. We generate 100 images for each case, selecting three carefully curated examples for visualization in Figure 6. Uncurated versions can be found in Appendix C. Each pair of images from the original models and C3 shares the same random seed, enabling a direct comparison of the changes introduced by the proposed method.

Original models have difficulty producing creative images even with prompts like "a creative obj". By incorporating our C3 method into these original models, we observe a marked enhancement in their creative generation capabilities. The images generated with our approach preserve the core semantic meaning of the original object while adding richer and more inventive elements. ConceptLab also generates distinct variations but often does so at the expense of the original object's semantic integrity. These results provide evidence of the considerable advancements achieved with the C3 method, underscoring its effectiveness in en-

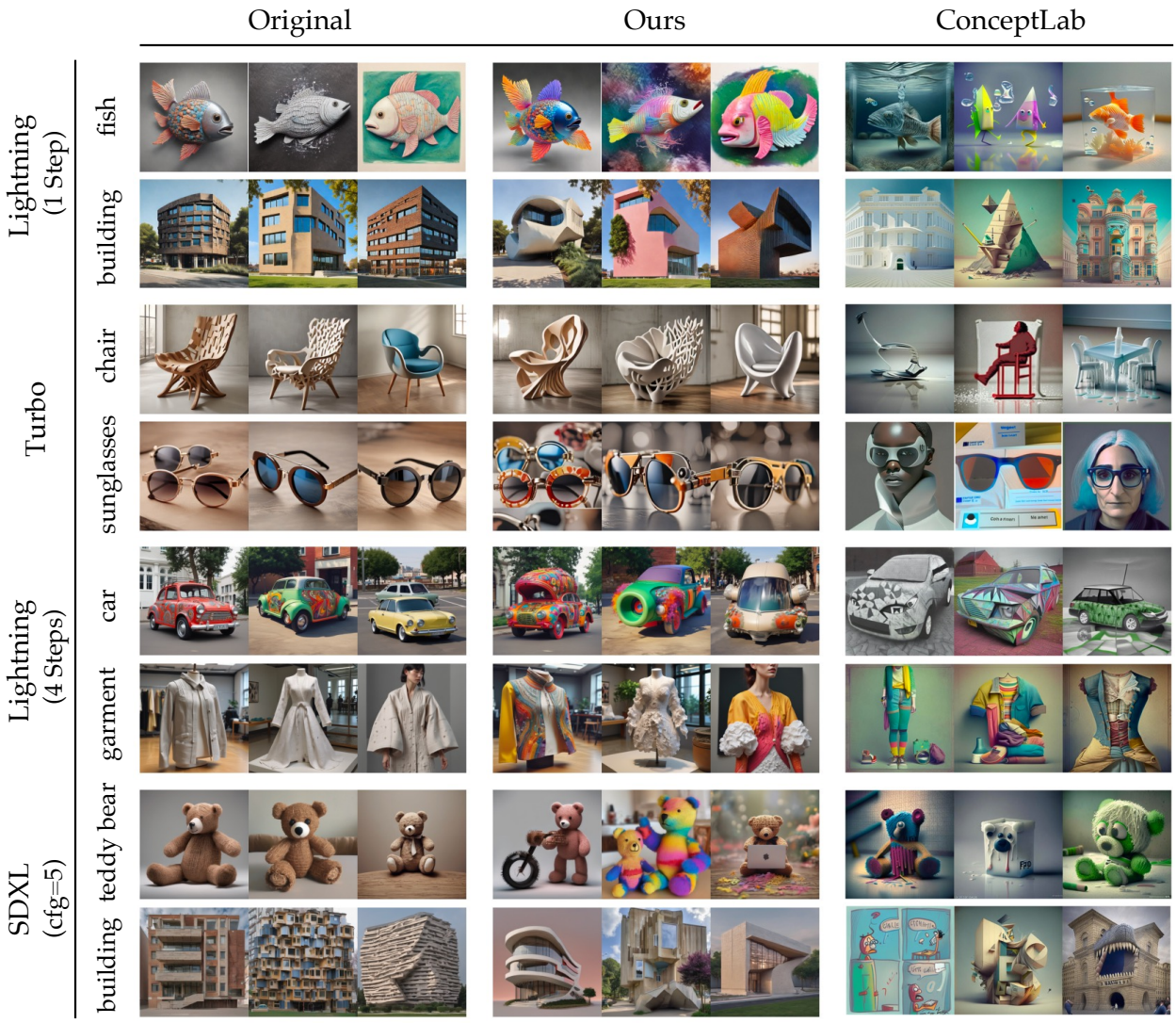

Figure 6. Qualitative results of **C3**, in comparisons of original generations and ConceptLab. For each object, three images are curated from 100 generations from each model/method. For ConceptLab, we train 10 different individual concepts and generate using 10 random seeds for each concept.

hancing the creative generation capabilities of the model.

4.4. User Study

We conducted a user study to evaluate the creativity of the generated images based on human perception. Participants were asked to respond to two questions for each instance, assessing both usability and novelty. The user study was conducted for the same five objects detailed in Section 4.2. The results are summarized in Table 2. Compared to the baseline models, our approach demonstrates improvements in novelty for both models. Notably, the novelty score surpasses that of ConceptLab. Although there are decreases in usability scores, these losses are smaller than the increases in novelty scores. Furthermore, the usability scores of our method are significantly higher compared to those of ConceptLab. More detailed results and experimental settings can be found in Appendix C.

5. Discussions

5.1. How C3 Enhances Creativity

"Creative" Ablation. We hypothesize that C3 only amplifies creativity when the prompt explicitly includes "creative." Enlarging the latent boosts cross-attention values associated with the creative object, enhancing its generation. To verify this, we ablate the "creative" in prompt and find that its removal results in ordinary images without noticeable artifacts (see Figure 7). The FID score even decreases,

Model	Method	Usability	Novelty
Lightning (1-step)	Orig	**4.62**	2.65
	Ours	4.19 (▼0.43)	**4.12** (▲1.47)
Turbo	Orig	**4.49**	3.08
	Ours	4.14 (▼0.35)	**3.79** (▲0.71)
ConceptLab	-	2.97	3.65

Table 2. User study results averaged over five objects. 'Usability' evaluates whether the image accurately represents the specified [obj], while 'Novelty' assesses the image's uniqueness. Responses are collected on a 5-point Likert scale.

indicating that C3 does not generate unexpected images without "creative" in prompt.

Figure 7. (Left) Images generated with and without "creative" in prompts. (Right) Change in FID scores after applying C3.

Justification for C3. To further support our hypothesis, we examine whether amplifying first-block features increases the second-block cross-attention map of "[obj]" as "[obj]" implies "creative [obj]" due to the self-attention in the text encoder. The t-test results confirm a significant increase for the cross-attention map of "[obj]." in both Turbo and Lightning model with p-value $< e^{-30}$. C3 amplifies only low-frequency features, which clearly enhance creativity in contents, as creativity driven by high-frequency features appears as colorful mosaics or noise (see Figure 4 in Sec. 3). Furthermore, we automatically select the amplification factor based on the usability score to maintain generation quality.

5.2. Types of Creativity

We investigated which aspects of creativity are enhanced in images generated by the proposed method. This identification was performed using a multi-modal LLM (GPT 4o [1]) to extract responses. The statistical findings, displayed in Figure 8, reveal that images generated by our method, depicted in vivid colors, achieve significantly higher values than those from the original model, shown in muted colors, highlighting an overall increase in creativity. Furthermore, we observe that different creative aspects are emphasized for each object type. For example, in the case of a chair, creativity is enhanced primarily in the shape aspect, garments in texture, and teddy bears in color. This observation suggests that our model effectively identifies and emphasizes suitable creative attributes for various objects, enhancing the generated images accordingly.

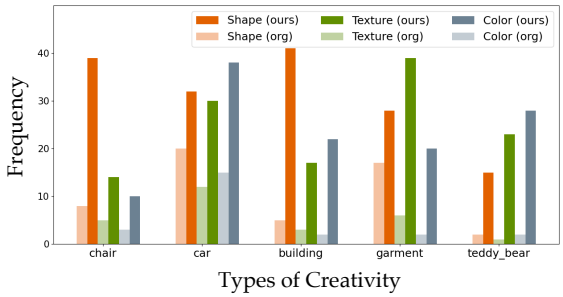

Figure 8. Types of creativity classified by GPT 4o for Lightning (1-step) generations. Responses are multiple-choice, among 'Shape', 'Texture', and 'Color'. 50 images are used for each method.

5.3. Use Cases

Integration with ControlNet [31]. To demonstrate the versatility of C3, we integrate it with ControlNet, a widely recognized plugin adapter. As shown in Figure 9, this combination enables effortless generation of creative samples while adhering to input constraints.

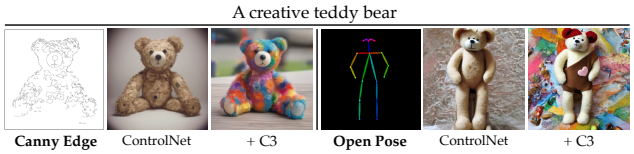

Figure 9. Combining C3 with ControlNet on SDXL.

Integration with SDXL Hyperparameters. We compare the results of our method with the readily available hyperparameters of the SDXL model, specifically in terms of classifier-free guidance (CFG) scale [8] and negative prompts. CFG is a hyperparameter that regulates the blend between the outputs of the given prompt and the negative prompt. Higher CFG values are known to enhance text-image alignment. However, we have observed that, in some cases, there are only marginal improvements in creativity, even with large CFG values. As illustrated in Figure 10, increasing the CFG from the default setting of 5 to 30 or using a negative prompt such as "normal teddy bear" leads to minimal changes in creativity. In contrast, our method enhances the default results without utilizing CFG or negative prompts. Notably, our method can also be employed alongside CFG and negative prompts, yielding more creative samples across all four settings.

Figure 10. Comparisons between pre-defined Hyperparameter controls of SDXL and **C3**. **(Default)** CFG=5 / Negative prompt="". **(CFG)** CFG=30. **(Neg)** Negative prompt="normal teddy bear".

Extension to Alternative Prompts. Our method is designed to function with textual descriptions that include *"creative"*. To assess the versatility of our approach, we tested whether similar effects could be achieved using alternative templates beyond "creative." In Figure 11, we constructed prompts using a total of four similar adjectives, {*creative, rare, innovative, ingenious*} and analyzed the outcomes when integrated with C3. We observed that across all templates, distinct and creative characteristics were expressed just as effectively as with "creative." This suggests that C3 consistently enhances creativity across a range of creativity-associated prompts.

Figure 11. Example outcomes using alternative prompts in place of "creative."

5.4. C3 on Non-Stable-Diffusion Models

We applied C3 to Kandinsky 3.0 and HunYuan-DiT, in addition to SDXL variants, as presented in Figure 12. Kandinsky 3.0 is based on a U-Net structure as SDXL, while HunYuan-DiT is based on a stacked transformer structure. The block-wise analysis for each model is provided in Appendix F. While the results open the possibility of expanding C3 to non-SDXL-based models, the specific architecture and components of the model may affect the application of C3 to other models, and more comprehensive analyses might be needed, especially to understand transformer-based models, which we leave as future work.

Figure 12. Application of C3 on Kandinsky 3.0 and HunYuan-DiT models.

5.5. Comparison with FreeU

Unlike FreeU [23], which enhances image fidelity by modifying low-frequency components in skip connections and boosts backbone features without frequency-based control—using uniform parameters s and b across blocks—C3 specifically targets low-frequency backbone features in the down and middle blocks, applying block-specific parameters. This distinction allows C3 to better steer creativity. Figure 13 shows that FreeU produces less clean and creative results than C3, regardless of parameter values. Especially when $b > 1$, noise persists regardless of s.

Figure 13. Results with prompt "A creative chair" in SDXL-Turbo.

6. Conclusion

In this paper, we propose a simple yet effective method to enhance the creative outputs of pre-trained Stable Diffusion-based models. C3 boosts creativity by amplifying internal features with auto-selected amplification factors, preserving quality without extensive fine-tuning or extra optimization. There are, however, limitations to our method. Our method heavily relies on the generation capabilities of the pre-trained models. Acting as a catalyst, our method may fail to generate a creative sample if the model itself has a limited concept of creativity for the target object. Moreover, the effectiveness of C3 across different model architectures and components requires comprehensive analysis, which is reserved for future work. Despite these limitations, we believe that our work, as the first training-free method for enhancing creative generations, can significantly contribute to the creative AI research community and inspire users, such as product designers, through the improved outcomes produced by our method.

Acknowledgment

This work was partly supported by the KAIST-NAVER Hypercreative AI Center, the Korean Institute of Information & Communications Technology Planning & Evaluation, and the Korean Ministry of Science and ICT under grant agreement No. RS-2019-II190075 (Artificial Intelligence Graduate School Program (KAIST)), No. RS-2022-II220984 (Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation), No.RS-2022-II220184 (Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics), No. RS-2024-00457882 (AI Research Hub Project) and Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD230017TD)

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint*, 2023. 7, 10

[2] Frank Barron. The disposition toward originality. *The Journal of Abnormal and Social Psychology*, 51(3):478, 1955. 4

[3] David Bau, Jun Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019. 2, 12

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023. 12

[5] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks generating "art" by learning about styles and deviating from style norms. In *ICCC*, 2017. 2, 12

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 2, 12

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528, 2021. 4

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 7

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 11

[10] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint*, 2024. 12

[11] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 12

[12] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint*, 2024. 1, 3, 4, 11

[13] Jack Lu. Procreate, don't reproduce! propulsive energy diffusion for creative generation. In *ECCV*, 2024. 2, 12

[14] Deborah Mateja and Armin Heinzl. Towards machine learning as an enabler of computational creativity. *IEEE Transactions on Artificial Intelligence*, 2(6):460–475, 2021. 4

[15] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415, 2012. 4

[16] AH Nobari, MF Rashad, and F Ahmed. Creativegan: Editing generative adversarial networks for creative design synthesis. In *IDETC-CIE*. American Society of Mechanical Engineers (ASME), 2021. 2, 12

[17] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 1, 3, 4, 11

[18] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In *EMNLP (Demos)*, 2023. 4

[19] Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics*, 43(3):1–14, 2024. 2, 4, 5, 1, 12

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3, 11

[21] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, pages 87–103. Springer, 2025. 1, 3, 4, 11

[22] Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. Design: Design inspiration from generative networks. In *ECCV workshops*, pages 37–44, 2018. 2, 12

[23] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, pages 4733–4743, 2024. 2, 8, 12

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 11

[25] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM TOG*, 42(6):1–13, 2023. 2, 12

[26] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint*, 2023. 2, 12

[27] Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Diffusion-based visual art creation: A survey and new perspectives. *arXiv preprint arXiv:2408.12128*, 2024. 1

[28] Yilun Xu, Gabriele Corso, Tommi Jaakkola, Arash Vahdat, and Karsten Kreis. Disco-diff: enhancing continuous diffusion models with discrete latents. In *ICML*, pages 54933–54961, 2024. 12

[29] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2024. 1

[30] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model, 2023. 5

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 7