Enhancing Creative Generation on Stable Diffusion-based Models

Supplementary Material

A. Implementation Details

Baseline Settings. We compared our **C3** method with the original Stable Diffusion-based models and ConceptLab [19]. Specifically, we evaluated four Stable Diffusion-based models: SDXL and its distilled variants—Turbo, Lightning 1-step, and Lightning 4-step. The results for each model are compared with the corresponding version enhanced by **C3**. Unless specified otherwise, we used the default settings of the original models, including the classifier-free guidance scale and negative prompts, to ensure a consistent baseline for evaluating the effectiveness of the **C3** method. For the ConceptLab, we adhered to the default settings outlined in the original paper. These settings include a batch size of 1, 2500 training steps, and "[object]" as the positive class. To generate 100 samples, we trained the embeddings using 10 different initial seeds, and for each trained embedding, we generated 10 new images during inference.

C3 Settings. In Table C, we provide the detailed parameter settings used in Section 4. These settings are designed to be broadly applicable, ensuring that configuring parameters within the provided range will likely produce satisfactory results for most prompts. The cut-off parameters indicate the extent to which specific frequencies are amplified. For the Turbo model, we set the cut-off threshold to 5 for every block. For the other models, the cut-off thresholds were set as [10, 5, 5, 5], corresponding to their respective blocks. The rationale for these settings is that the cut-off threshold should align with the resolution of the internal features, ensuring optimal handling of feature granularity at different blocks. For the amplification factor selection, we use the mean usability score $\text{Use}(\mathcal{I}) = \frac{1}{N}\sum_{i=1}^{N} Aesthetic(I_i) + \frac{1}{N}\sum_{i=1}^{N} CLIP(I_i, c)$ for $\mathcal{I} = \{I_i\}_1^N$ to provide statistically consistent amplification factors. In the experiments, we use the number of samples $N = 100$. To balance the scale of the aesthetic score and CLIP score, we min-max scale each score over the configurations $\{(l, \lambda_l^i)\}_{(l,i)}$. The usability bumper is a parameter designed to balance the usability and novelty of the images generated with **C3**. For the SDXL model, we set the usability bumper to 0.7, while for the other models, it was set to 0.8. For the sum constraint applied to scaling factors, which controls the degree of amplification across multiple blocks, we used the sum values of 0.6, 0.8, and 1. The specific value was chosen based on the given prompt and the model in use. Additionally, we provide detailed block-wise scaling factors for a more comprehensive understanding of the amplification strategy. In the next section, we conduct in-depth analyses of the effects of various hyperparameters on the results.

B. Ablation Study on the Hyperparameters

B.1. Analysis on Cutoff Threshold

In this subsection, we analyze the effect of various cutoff thresholds. The cutoff threshold c defines the extent to which frequency we would amplify. The low-frequency mask $M_L \in [0,1]^{n\times n}$ is then defined with the cutoff threshold c as follows.

$$M_L^{i,j} = \begin{cases} 1 & \text{if } r(i,j) < c \\ 0 & otherwise \end{cases} \tag{7}$$

Here, $M_L^{i,j}$ denotes the element of M_L located at the i^{th} row and j^{th} column, and $r(i,j) = \sqrt{(i-\frac{n}{2})^2 + (j-\frac{n}{2})^2}$. The resolution $n \times n$ varies across blocks and models. Therefore, the resolution should be considered when setting the cutoff threshold. In Figure K, we present the detailed amplification results for various cutoff thresholds for each block. In the first and second down blocks, a larger cutoff threshold facilitates more colorful variation. However, too large cutoff threshold introduces a tile pattern in the image that degrades quality. Conversely, a smaller cutoff threshold successfully prevents this tile pattern but, if set too low, can result in excessive information loss and over-smoothing of the object. By adjusting the cutoff threshold, one can find outcomes with a unique shape and color pattern. Compared to the shallow blocks, the amplification results on the third down block and the middle block indicate that these deeper blocks are less sensitive to the cutoff thresholds. Furthermore, we observe that a smaller cutoff threshold generally permits greater amplification, while a larger cutoff threshold tends to generate noise images with a smaller amplification factor.

B.2. Analysis on Usability Bumper

In this subsection, we analyze the effect of the usability bumper ϵ defined in Section 3.3. The usability bumper is used as a control parameter between usability and novelty. When ϵ is close to 1.0, the usability score is preserved similar to that of the

		Cut-off	Usability Bumper	Amplification Factors				Block-wise Scaling Factors
SDXL	chair	[10,5,5,5]	0.7	1.15	1.6	5	6	sum[0.3,0.3,0.1,0.1]=0.8
	teddy bear	[10,5,5,5]	0.7	1.2	1.8	4	2	sum[0.4,0.4,0.1,0.1]=1.0
	car	[10,5,5,5]	0.7	1.25	1.6	5	4	sum[0.3,0.3,0.1,0.1]=0.8
	building	[10,5,5,5]	0.7	1.25	1.8	5	4	sum[0.2,0.15,0.15,0.1]=0.6
	garment	[10,5,5,5]	0.7	1.2	1.8	5	2	sum[0.4,0.4,0.1,0.1]=1.0
Lightning (1-step)	chair	[10,5,5,5]	0.8	1.5	2.25	5	6	sum[0.2,0.2,0.1,0.1]=0.6
	teddy bear	[10,5,5,5]	0.8	1.5	2.75	6	7	sum[0.2,0.2,0.1,0.1]=0.6
	car	[10,5,5,5]	0.8	1.5	2.5	6	6	sum[0.2,0.2,0.1,0.1]=0.6
	building	[10,5,5,5]	0.8	1.6	2.5	7	8	sum[0.2,0.2,0.1,0.1]=0.6
	garment	[10,5,5,5]	0.8	1.3	1.9	6	7	sum[0.2,0.2,0.1,0.1]=0.6
	fish	[10,5,5,5]	0.8	1.4	2.75	5	5	sum[0.2,0.2,0.1,0.1]=0.6
Lightning (4-step)	chair	[10,5,5,5]	0.8	1.4	2	7	8	sum[0.2,0.15,0.15,0.1]=0.6
	teddy bear	[10,5,5,5]	0.8	1.4	2.25	6	9	sum[0.2,0.15,0.15,0.1]=0.6
	car	[10,5,5,5]	0.8	1.4	1.9	6	6	sum[0.2,0.15,0.15,0.1]=0.6
	building	[10,5,5,5]	0.8	1.3	1.9	8	7	sum[0.2,0.15,0.15,0.1]=0.6
	garment	[10,5,5,5]	0.8	1.25	1.8	5	4	sum[0.2,0.15,0.15,0.1]=0.6
Turbo	chair	[5,5,5,5]	0.8	1.25	1.5	9	10	sum[0.3,0.3,0.2,0.2]=1.0
	teddy bear	[5,5,5,5]	0.8	2	1.5	7	10	sum[0.4,0.4,0.1,0.1]=1.0
	car	[5,5,5,5]	0.8	1.75	2.5	8	10	sum[0.4,0.4,0.1,0.1]=1.0
	building	[5,5,5,5]	0.8	2.75	3.75	10	10	sum[0.2,0.15,0.15,0.1]=0.6
	garment	[5,5,5,5]	0.8	1.5	2.25	7	10	sum[0.4,0.4,0.1,0.1]=1.0
	sunglasses	[5,5,5,5]	0.8	3.75	5	7	10	sum[0.1,0.1,0.2,0.2]=0.6

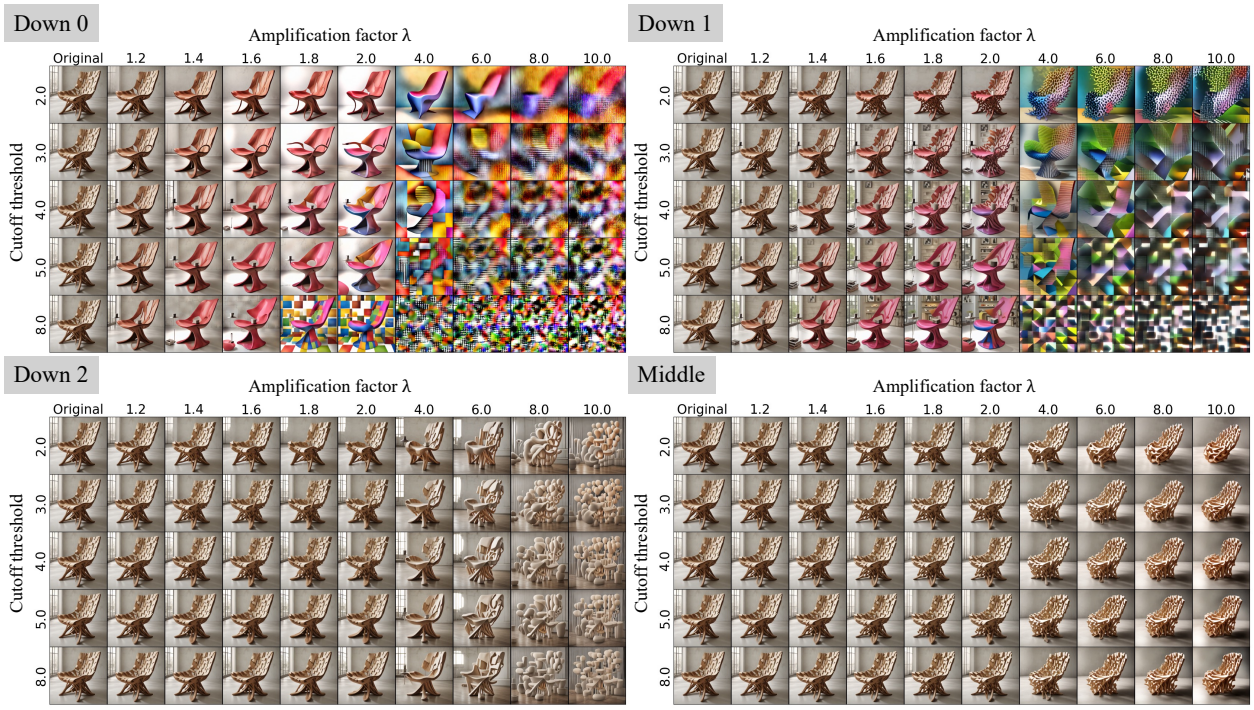
Table C. Detailed Settings for the used parameters in experiments. The numbers within the list represent the corresponding values applied to each block.

original image, albeit with a loss of novelty. Conversely, as ϵ decreases, it permits greater variation from the original image and enhances novelty, albeit at the expense of the usability score. Figure L shows examples with the use of the various usability bumper. Turbo with the prompt "a creative cup" is used for the generation. The images with the colored bounding boxes indicate the selected amplification factors with Equation (6) in Section 3.3. For each block, the amplification factors found with $\epsilon = 0.99$, marked with the blue bounding boxes, produce images that maintain high fidelity to the original image, with only slight changes in detail. Conversely, the amplification factors identified with $\epsilon = 0.4$, indicated by the green bounding boxes, generate images with significant variation from the original. Specifically, the shallower blocks exhibit artistic cup images with high color variation, while the deeper blocks show changes primarily in shape, albeit with compromised image fidelity. The amplification factors identified with $\epsilon = 0.8$ as used in the main experiment, marked with the orange bounding boxes, result in outcomes that fall between these two extremes. The following three images, indicated by colored bounding boxes, display the results of amplification across all four blocks. For the scaling factors, 0.3, 0.3, 0.2, 0.2 are used for each block, respectively. The results indicate that $\epsilon = 0.8$ produces a cup image that is both creative and feasible.

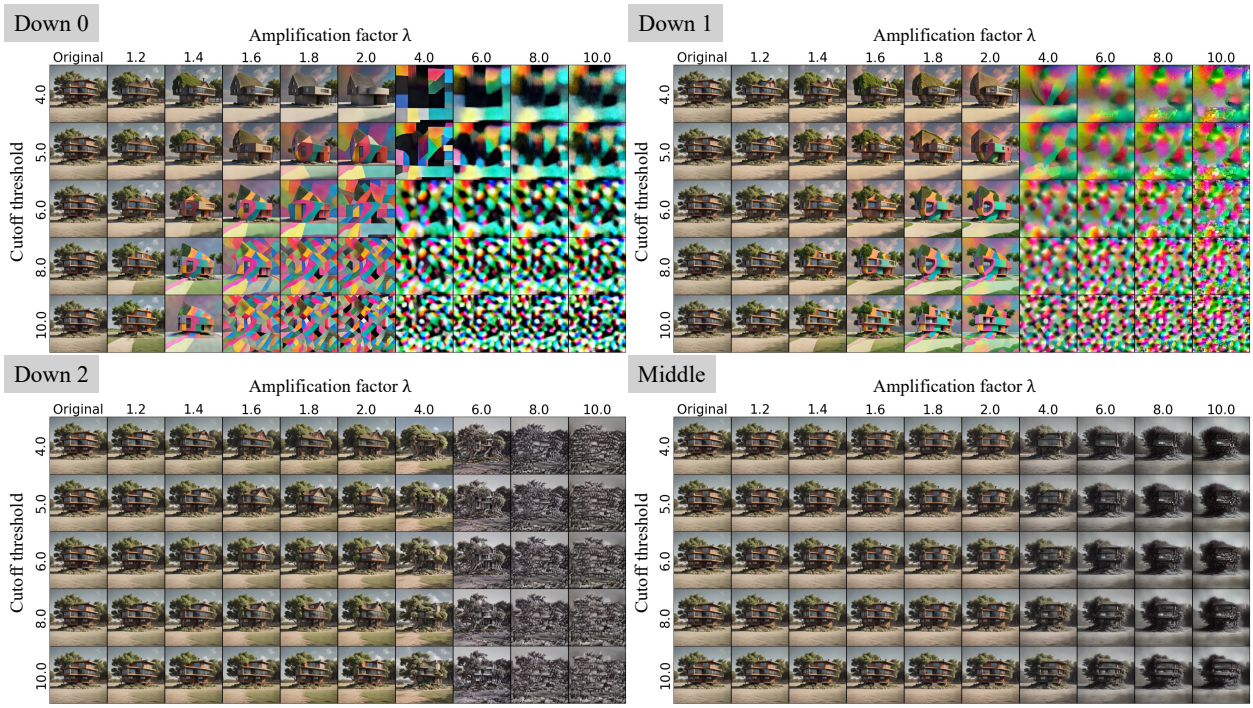
B.3. Analysis on Scaling Factors

We introduce an automatic strategy to determine the optimal amplification factor λ_l^* in Section 3.3. This approach strikes a balance between usability and novelty, generating semantically meaningful yet creative features that lead to creative images.

Applying our method across multiple blocks simultaneously enables the generation of more flexible and creative images. However, when the changes in multiple blocks are simply accumulated, the resulting features may exceed the allowable range of the pre-trained Stable Diffusion-based model, leading to broken or degraded images. To address this, we apply additional scaling factors s_l during multi-block applications of **C3** to preserve image quality. Then, we can formulate the **C3** method applied across multi-blocks as follows:



(a) SDXL-Turbo, prompt="a creative chair"



(b) SDXL-Lightning (1-step), prompt="a creative house"

Figure K. Amplification results for various cutoff thresholds.

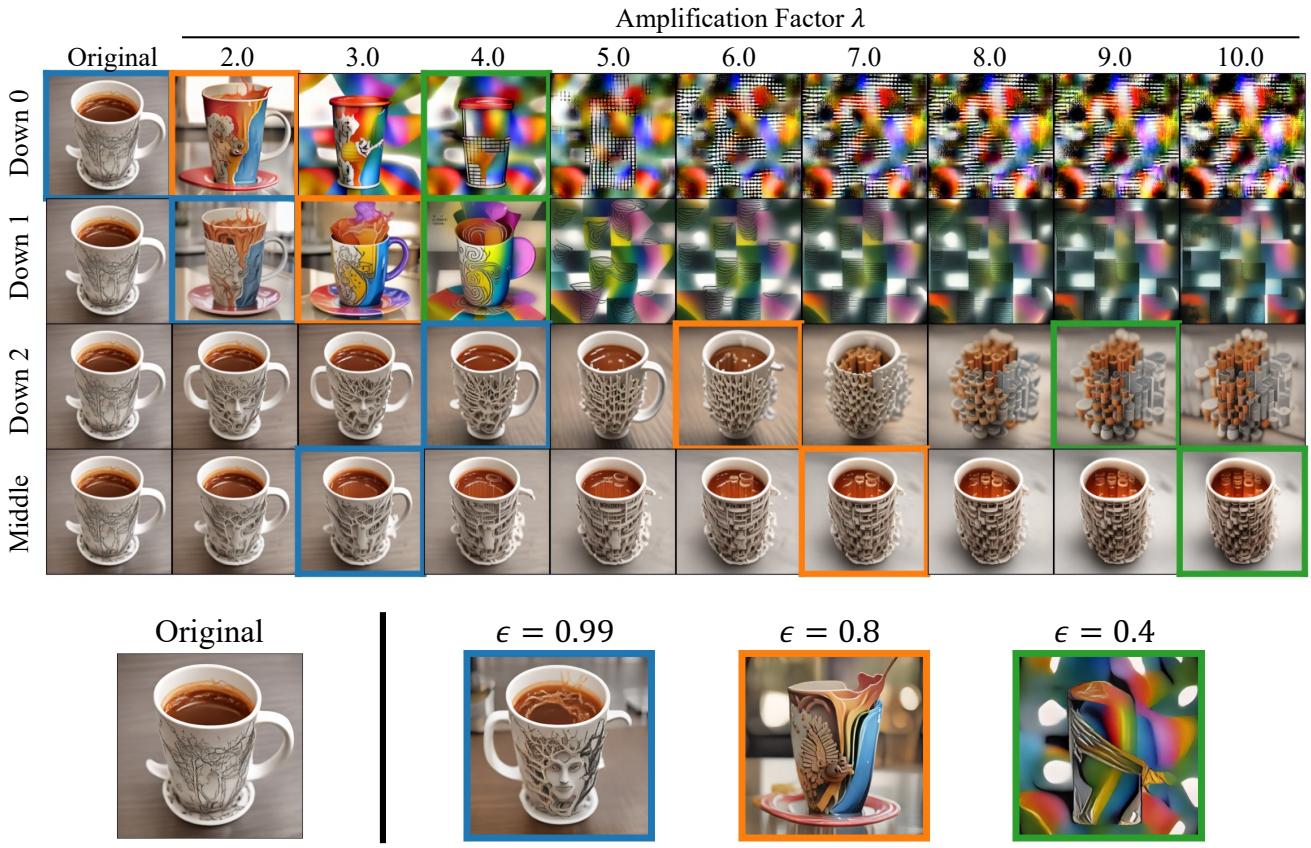


Figure L. The amplified results with the various usability buffer ϵ.

$$f^*(x_l) = s_l \cdot \lambda_l^* \cdot f_L(x_l) + f_H(x_l) \tag{8}$$

Empirically, we observe that bounding the sum of scaling factors across the blocks, denoted as $S = \sum_l s_l$, aids in preventing extensive parameter search for s_l. Within this sum constraint, the block-specific scaling factors can be adjusted in a user-controllable manner, allowing for flexible image generation. We observed that selecting an appropriate sum constraint prevents degradation in image quality, even as the scaling factors for individual blocks vary. (See Figure N)

In Figure M, we display the variations in images for each model and object as S changes. Results highlighted in red boxes represent the constraint value we used. (Specific block-wise scaling factor settings for the figures are summarized in Table C). For the SDXL-Lightning 1-step model, we observe that using a summation constraint of $S = 0.6$ resulted in the most creative images while maintaining the usability of the object in most cases. For the SDXL-Turbo model, a summation constraint of approximately $S = 1$ produces highly creative results while effectively preserving the structural integrity of objects like chairs. However, for more complex objects, such as buildings, the cumulative amplification tends to introduce additional noise, requiring a more conservative summation constraint to balance creativity and object clarity.

Furthermore, we quantitatively analyze the necessity of the scaling factor and its correlation with usability, which is measured using the BLIP score. As introduced earlier in Section 4.2, BLIP score represents the proportion of generated samples that receive a "yes" response from the BLIP VQA model when asked, "Is this image [object]?". For each sum of scaling factors, 100 images are generated with different scaling factors. These 100 cases were obtained by randomly sampling scaling factors for each block,s_l, within the given sum value.

The results reveal that as the sum of the scaling factors increases, the BLIP score decreases, indicating that larger scaling factors compromise usability. We set the scaling factor constraints to a value that ensures the model does not compromise its usability significantly, as highlighted with bold markers. Importantly, these findings are based on randomly selected scaling factors, demonstrating that the quality of the generated images remains robust within the specified sum constraint, regardless of how the scaling factors are distributed across blocks.

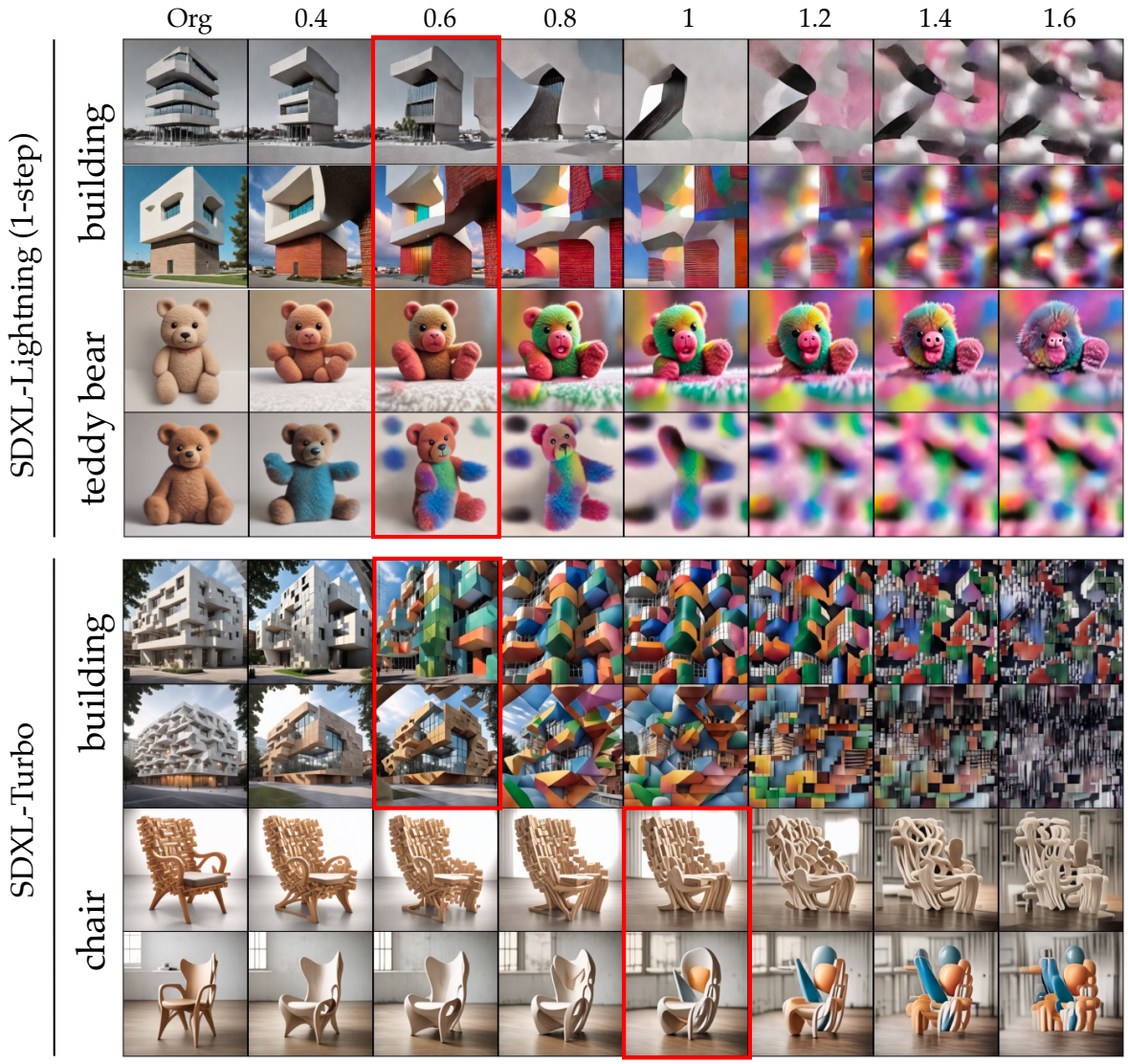


Figure M. Amplification results for various scaling factors with the sum constraint. The red box presents the sum value we used for each object.

B.4. Step-wise Analysis

Diffusion models operate through a multi-step denoising process. In this section, we examine the effects of applying C3 at various stages within this denoising process. We observe the changes in generated images by applying our method at six distinct points across a total of 50 steps. To quantify the degree of image change, we use LPIPS, a perceptual similarity metric, to compare the results at each stage with those generated by the proposed method. The LPIPS scores, averaged across 100 different images and random seeds, are shown in Figure O-(top). Both the LPIPS scores and exemplar images show that when C3 is applied after the fifth step (Figure O-(c)), the resulting images increasingly resemble the original, diverging from those generated with C3 applied continuously at each step. This analysis demonstrates that the impact of our method is most pronounced when applied in the early stages of the diffusion process, aligning with prior analyses on diffusion models that suggest structural content is primarily established in the earlier timesteps [30].

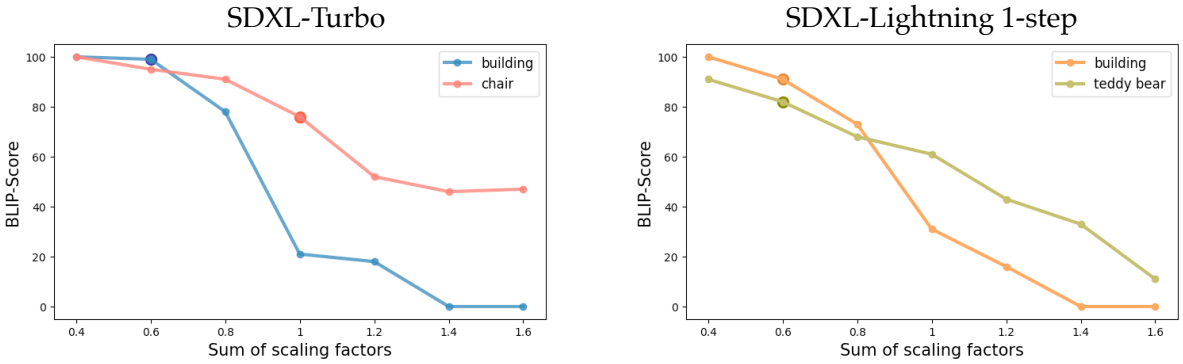


Figure N. BLIP scores of various scaling factors.

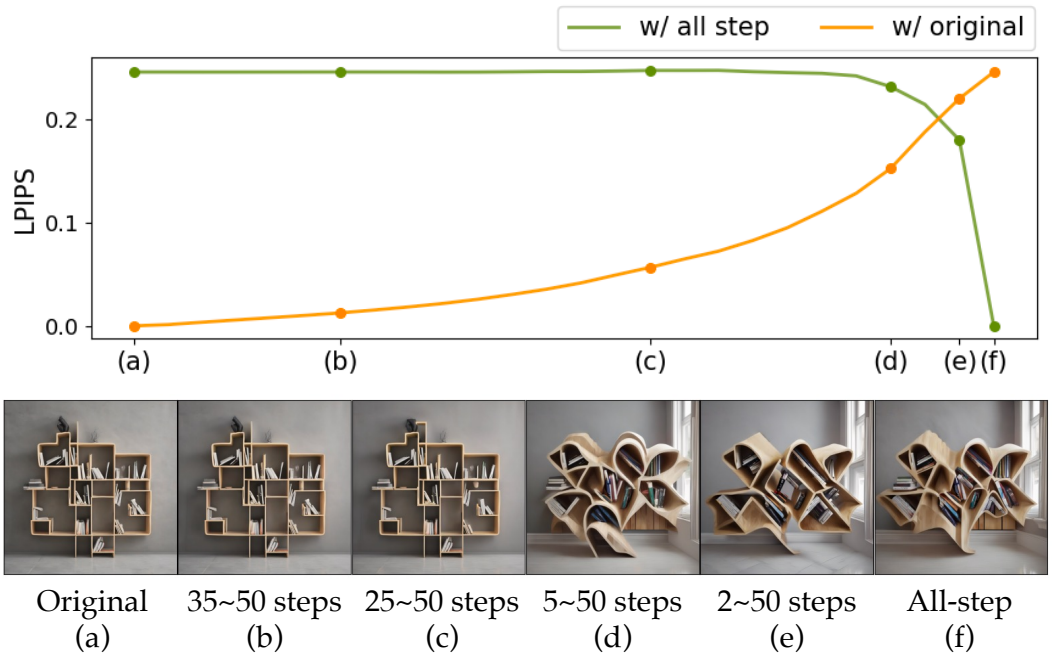


Figure O. Step-wise amplification results on SDXL (step=50). Significant changes are observed in the earlier steps.

C. Detailed Experimental Results

C.1. Qualitative Results

C.1.1. Uncurated Samples for SDXL-Lightning (1-step)

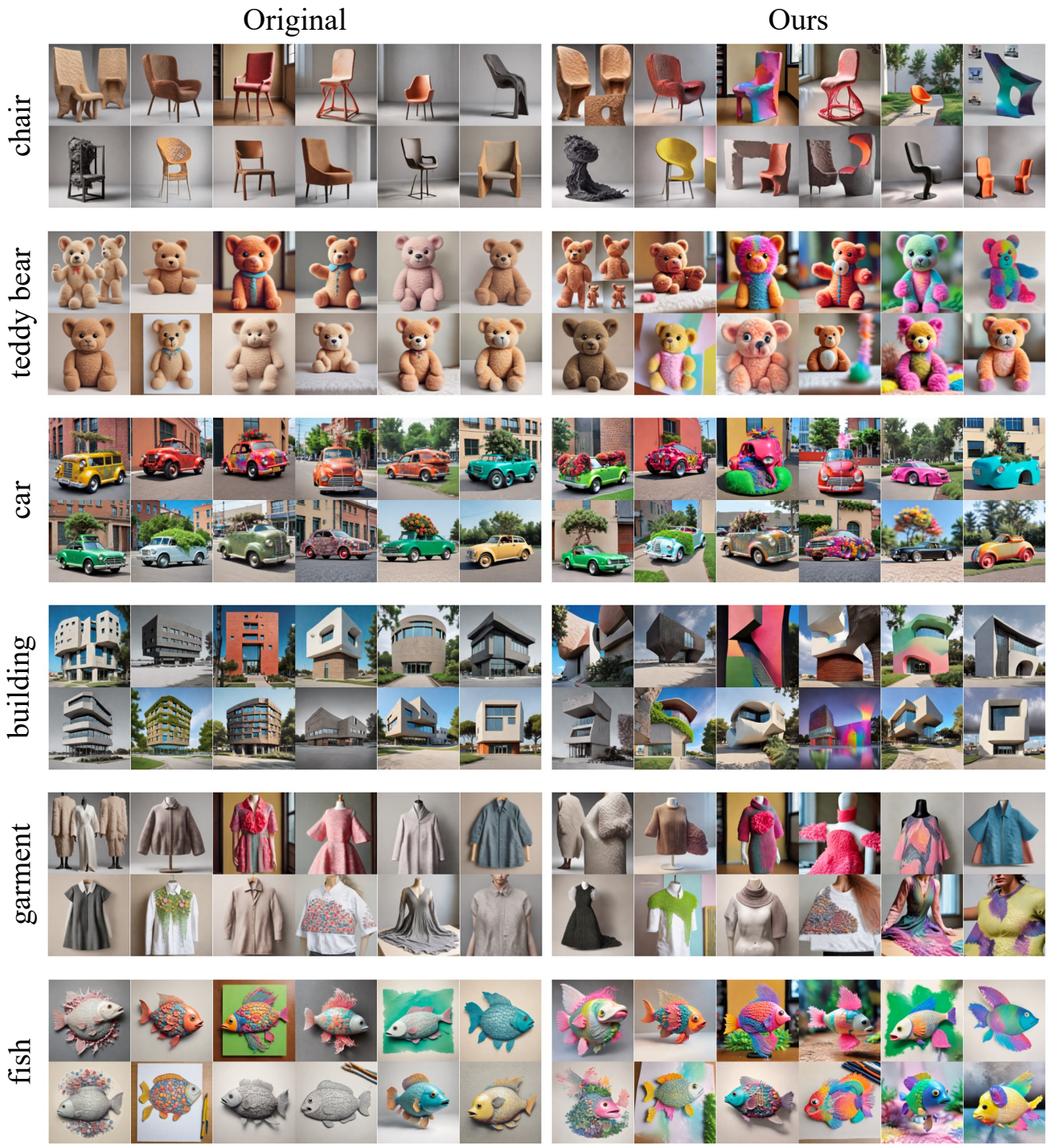


Figure P. Uncurated samples generated from SDXL-Lightning (1-step). The samples are generated by manually setting the random seed to values ranging from 0 to 11.

C.1.2. Uncurated Samples for SDXL-Turbo

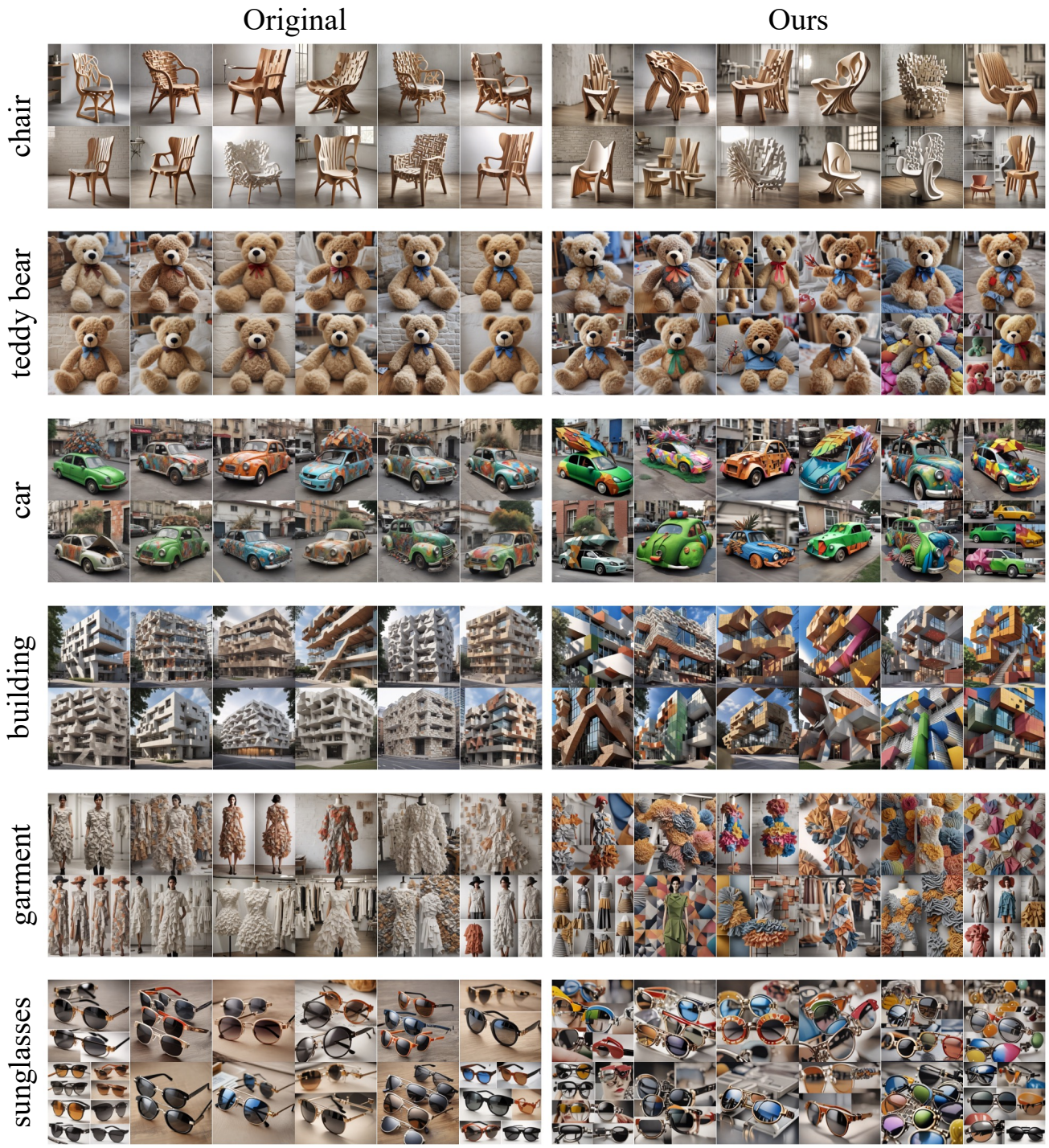


Figure Q. Uncurated samples generated from SDXL-Turbo. The samples are generated by manually setting the random seed to values ranging from 0 to 11.

C.1.3. Uncurated Samples for SDXL-Lightning (4-step)

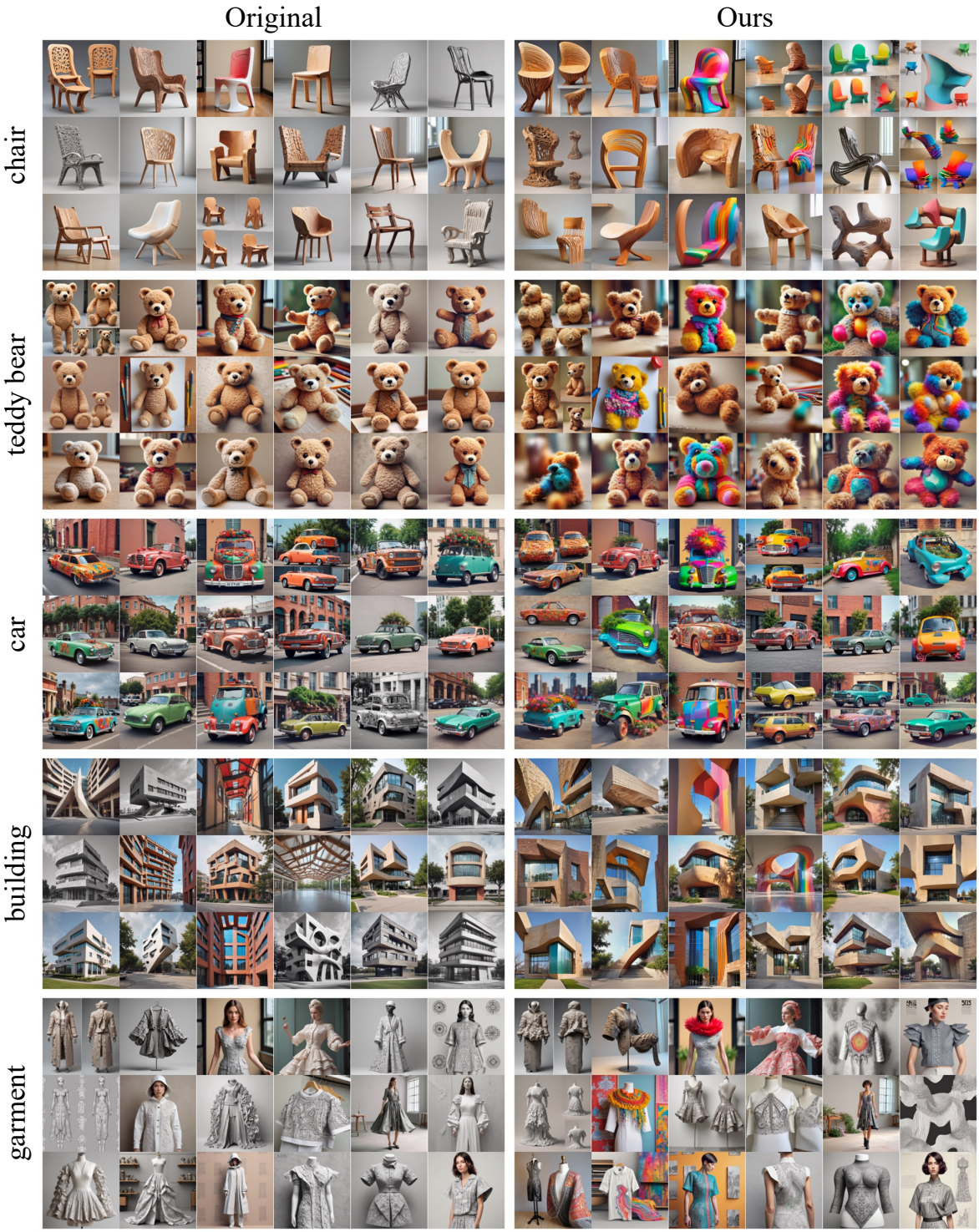


Figure R. Uncurated samples generated from SDXL-Lightning (4-step). The samples are generated by manually setting the random seed to values ranging from 0 to 17.

C.1.4. Uncurated Samples for SDXL

Figure S. Uncurated samples generated from SDXL. The samples are generated by manually setting the random seed to values ranging from 0 to 17.

C.1.5. Uncurated Samples for ConceptLab

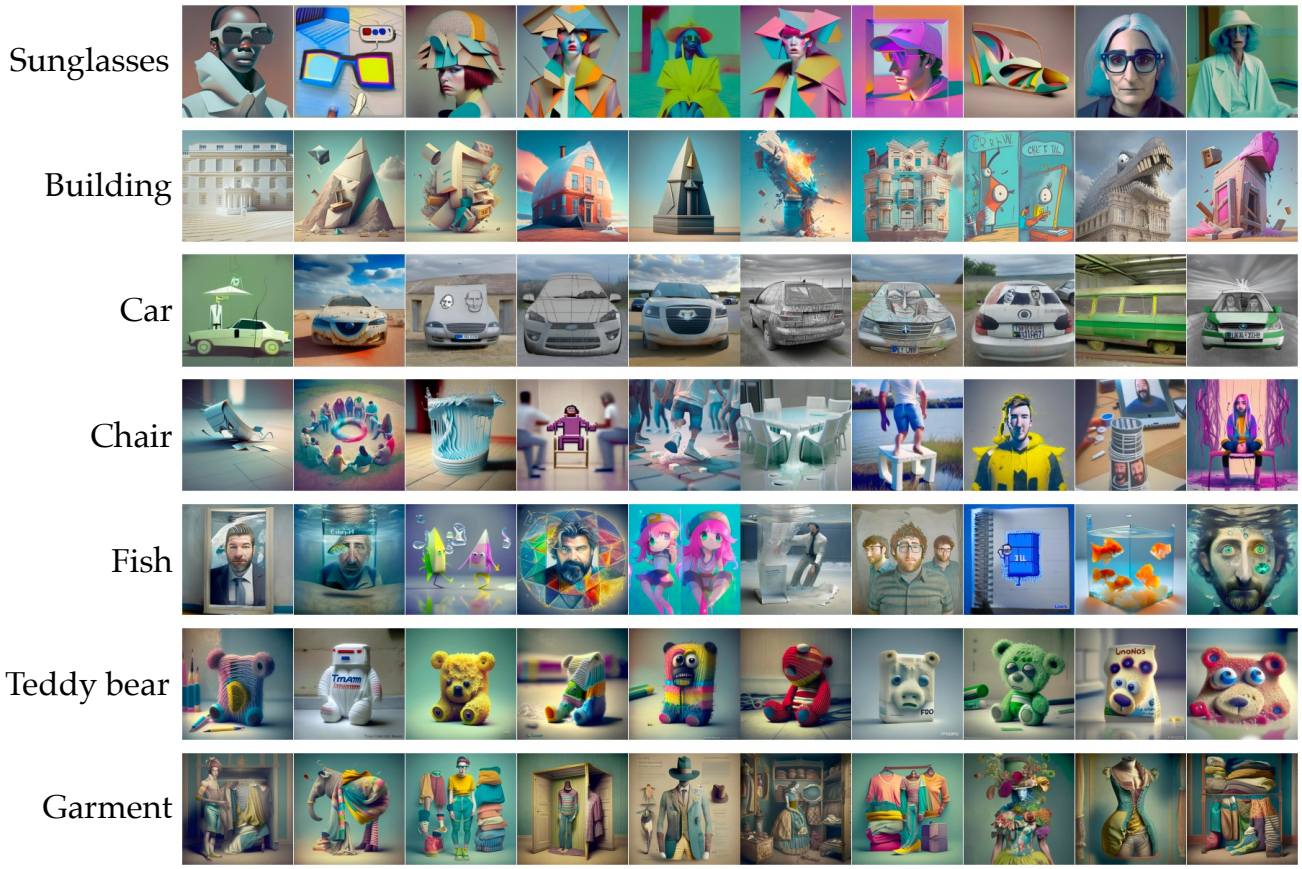


Figure T. Uncurated samples generated from ConceptLab. Each column of samples is trained with a different initial seed.

C.2. Quantitative Results

We list the detailed quantitative results in Table D, corresponding to each object we use in Section 4.2. FID, precision and recall are computed based on two sets of images, conventionally say, real dataset and fake dataset. For the real dataset, we use 100 images generated from SDXL with the prompt of "a [object]". For the fake dataset, we use 100 images generated from each model and method. Here, we underscore again that FID and precision are interpreted as the opposite of the conventional way as our aim is to produce the object images that are distinct from the typical target object. While the image generated with our method are to be distinctive, it should also be recognized as the target object. Thus, we provide the reference scores which are computed using the fake dataset, generated from SDXL with the prompt of "a [reference-object]" for each object.

The overall trends are shared across the objects. For novelty metrics, our method outperforms the original generation in all objects. Especially for FID, our method does not exceed the reference score, indicating that the generated images are novel yet perceived as different objects. For 'chair', our method shows low precision scores compared to the reference score, as the outlook of chairs are significantly different from the ordinary chair images. However, the high BLIP score as a usability metric defenses that the generated images with our method still look as chairs. Conversely, ConceptLab, a baseline method for comparison, presents significantly low BLIP scores for some objects as illustrated in Figure 6. This limitation of ConceptLab arises from the increased difficulty in defining sub-categories within a specific category.

Notably, our method also increases the diversity within the generated creative samples. Both LPIPS scores, which compute the distances between the generated samples in the feature space, and Vendi scores, which represent the effective number of modes among the samples, show notable improvement over the original generation across the objects. Recall scores, which are considered a measure of mode coverage within the real dataset, are comparable in most cases and show significant improvement for the Turbo model, which suffers from the mode collapse issue.

Object	Model	Method	Novelty		Diversity			Usability	
			FID* (↑)	Prcs* (↓)	Rcl (↑)	LPIPS (↑)	Vendi (↑)	CLIP (↑)	BLIP (↑)
Chair	Lightning (1-step)	Orig	108.91	0.81	**0.93**	0.17	6.58	**0.29**	**0.97**
		Ours	**185.18**	**0.35**	0.84	**0.27**	**8.54**	0.27	0.89
	Turbo	Orig	94.95	0.96	0.51	0.20	3.76	**0.29**	**1.00**
		Ours	**132.72**	**0.56**	**0.62**	**0.24**	**5.92**	0.29	0.99
	Lightning (4-step)	Orig	91.55	0.79	**1.00**	0.20	5.72	**0.29**	**0.99**
		Ours	**178.05**	**0.34**	0.76	**0.30**	**8.45**	0.28	0.82
	SDXL	Orig	104.94	0.84	**0.97**	0.18	7.77	**0.29**	**0.96**
		Ours	**158.88**	**0.60**	0.91	**0.25**	**8.57**	0.28	0.87
	Real-to-Ref	-	207.47	0.87	0.52	-	-	-	-
	ConceptLab	-	266.58	0.59	0.66	0.37	10.46	0.23	0.01
Teddy Bear	Lightning (1-step)	Orig	65.55	0.99	0.28	0.10	1.98	**0.29**	**1.00**
		Ours	**82.58**	**0.80**	**0.69**	**0.23**	**2.82**	0.28	0.79
	Turbo	Orig	84.89	0.89	0.08	0.14	1.33	**0.30**	**1.00**
		Ours	**85.11**	**0.79**	**0.53**	**0.28**	**1.71**	0.29	**1.00**
	Lightning (4-step)	Orig	78.07	0.91	**0.87**	0.20	1.76	0.29	**1.00**
		Ours	**86.96**	**0.50**	0.78	**0.30**	**3.08**	**0.29**	0.95
	SDXL	Orig	67.26	0.91	**0.98**	0.19	3.03	**0.29**	**1.00**
		Ours	**97.21**	**0.84**	0.96	**0.31**	**4.07**	0.28	0.89
	Real-to-Ref	-	297.82	0.71	0.31	-	-	-	-
	ConceptLab	-	337.86	0.87	0.29	0.33	8.12	0.26	0.07
Garment	Lightning (1-step)	Orig	172.54	1.00	0.76	0.31	7.07	**0.27**	**1.00**
		Ours	**193.78**	**0.93**	**0.87**	**0.39**	**8.52**	**0.27**	0.93
	Turbo	Orig	212.69	0.87	0.15	0.20	5.21	**0.26**	**1.00**
		Ours	**214.95**	**0.68**	**0.47**	**0.36**	**6.95**	0.26	0.93
	Lightning (4-step)	Orig	165.04	0.91	**0.93**	0.29	7.58	**0.26**	**0.98**
		Ours	**176.02**	**0.72**	0.92	**0.37**	**8.78**	0.26	0.89
	SDXL	Orig	167.05	0.89	**0.95**	0.22	8.31	**0.27**	**0.94**
		Ours	**196.13**	**0.74**	0.94	**0.38**	**9.10**	0.26	0.81
	Real-to-Ref	-	232.91	0.83	0.80	-	-	-	-
	ConceptLab	-	225.85	0.89	0.71	0.37	7.79	0.26	0.66
Car	Lightning (1-step)	Orig	92.83	0.84	0.90	0.36	4.78	0.25	**0.89**
		Ours	**111.43**	**0.61**	**0.93**	**0.42**	**5.25**	**0.26**	0.88
	Turbo	Orig	131.62	0.77	0.45	0.30	3.08	0.26	**1.00**
		Ours	**150.14**	**0.22**	**0.96**	**0.46**	**4.89**	**0.26**	0.84
	Lightning (4-step)	Orig	86.82	0.85	0.89	0.39	3.90	0.25	**1.00**
		Ours	**110.63**	**0.55**	**0.95**	**0.43**	**4.70**	**0.26**	0.95
	SDXL	Orig	119.38	0.54	0.96	0.30	5.47	0.26	**0.88**
		Ours	**137.24**	**0.39**	**0.98**	**0.34**	**6.24**	**0.26**	0.77
	Real-to-Ref	-	220.79	0.31	0.38	-	-	-	-
	ConceptLab	-	151.87	0.47	0.79	0.40	6.70	0.27	0.71
Building	Lightning (1-step)	Orig	179.90	0.83	0.56	0.34	**6.12**	**0.26**	**1.00**
		Ours	**242.10**	**0.57**	**0.64**	**0.37**	6.07	0.24	0.97
	Turbo	Orig	208.01	0.86	0.17	0.28	4.31	**0.24**	**1.00**
		Ours	**237.44**	**0.32**	**0.82**	**0.48**	**5.19**	**0.24**	**1.00**
	Lightning (4-step)	Orig	165.12	0.84	**0.89**	0.35	6.17	**0.25**	0.98
		Ours	**222.89**	**0.65**	0.75	**0.37**	**6.37**	0.24	**0.99**
	SDXL	Orig	182.38	0.76	**0.96**	0.30	8.00	**0.24**	**0.97**
		Ours	**200.18**	**0.74**	0.87	**0.32**	**8.64**	0.24	**0.97**
	Real-to-Ref	-	281.64	0.09	0.80	-	-	-	-
	ConceptLab	-	274.19	0.43	0.77	0.37	8.98	0.23	0.13

Table D. Object-wise quantitative results. 'Prcs' and 'Rcl' refer to precision and recall, respectively.

C.3. User Study

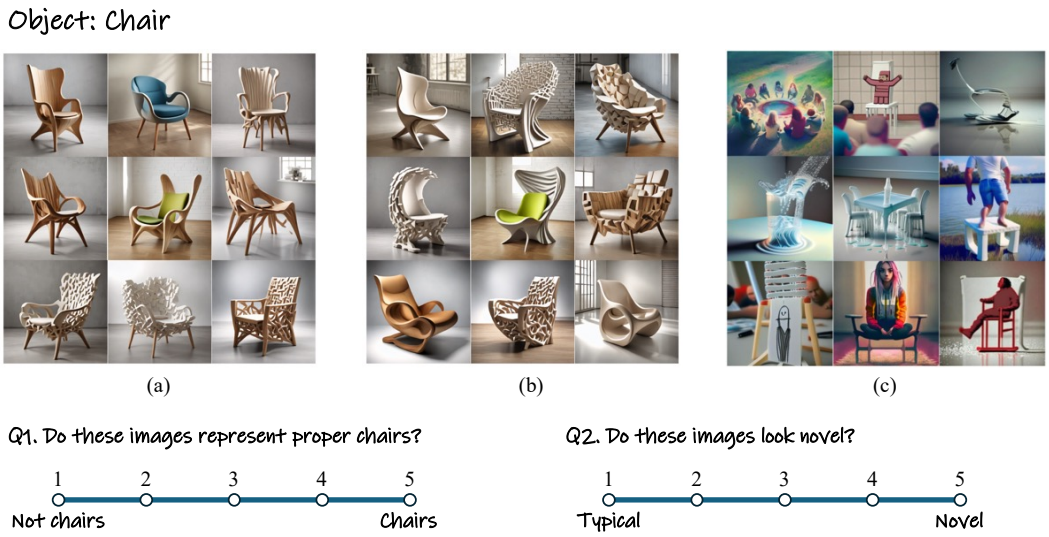


Figure U. Example of user study questionnaires for the object 'Chair'. The questions are posed repeatedly for each set of images. (a) SDXL-Turbo (Original). (b) SDXL-Turbo (Ours). (c) ConceptLab. Images are carefully curated as the best to represent each method.

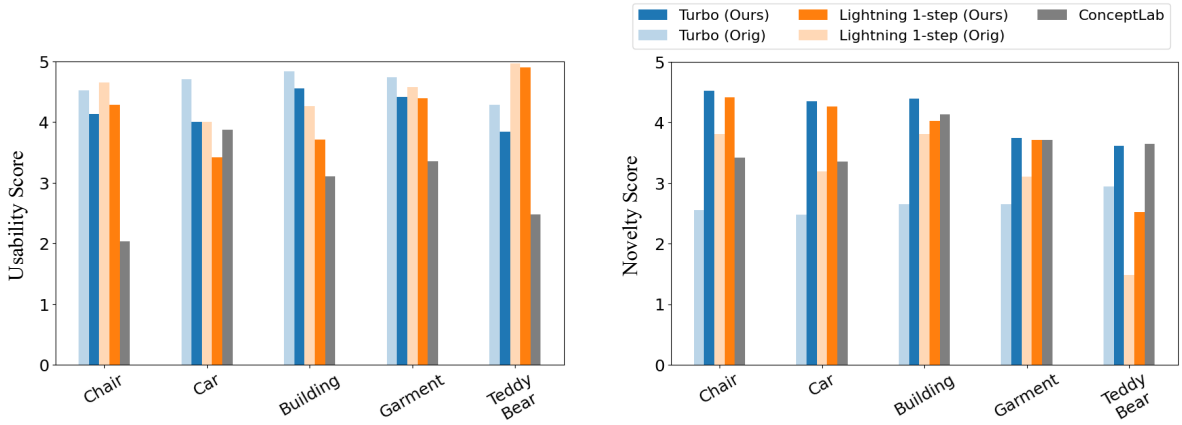


Figure V. Object-wise results of user study. Our method improves the original results in terms of both usability and novelty for all objects.

We conduct a user study to evaluate the creativity of generated samples for each method with human perception. Given a set of images, we ask participants questions regarding two main aspects of creativity: usability and novelty. As an example illustrated in Figure U, with a target object 'chair', we ask (1) whether these images represent proper chairs and (2) whether these images look novel. The responses are collected using a 5-level Likert scale. For a target object, images are generated from the model's default setting, marked as 'Orig', and from **C3**, marked as 'Ours'. Additionally, images are generated from ConceptLab for baseline comparison. For each image set, we carefully select 9 images that appear to be the most creative within the generated images for each method. While each method is anonymized in the questions, we denote each method in the result as 'Turbo (Orig/Ours)', 'Lightning (1-step) (Orig/Ours)', and 'ConceptLab', respectively.

In total, 31 participants have responded. We summarize the responses for each object in Figure V. The blue bar plots represent results corresponding to Turbo, while the yellow bar plots represent results corresponding to Lightning (1-step). The darker color represents **C3** marked as 'Ours', while the lighter color represents the default generation marked as 'Orig'. The results corresponding to ConceptLab are presented with the gray bar plots. Our method significantly enhances novelty in all cases, with only a relatively small reduction in usability scores. Notably, our method outperforms ConceptLab in usability scores and achieves higher or comparable novelty scores.

D. Types of Creativity

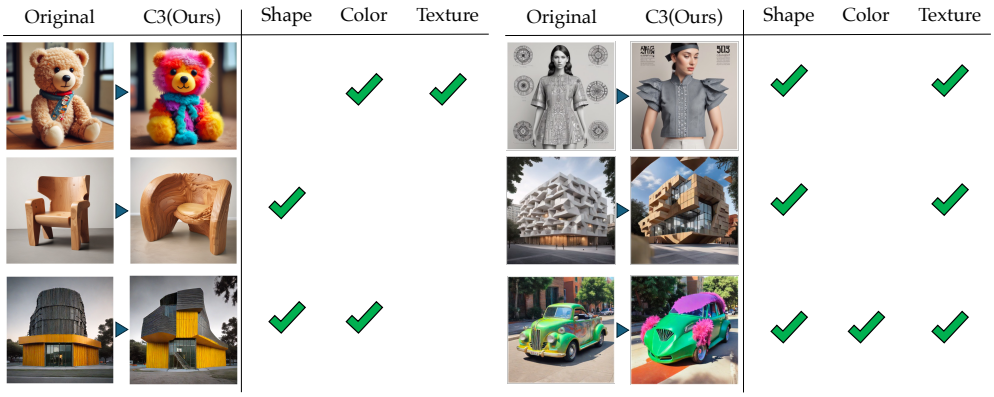


Figure W. Types of creativity classified by GPT 4o of samples generated by the proposed method compared to samples generated by the original models. Responses are multiple-choice, among 'Shape', 'Texture', and 'Color'.

Here, we present the categories of creativity amplified in samples generated using the proposed method. All images are generated using the prompt "a creative [object]." The teddy bear, garment, and chair images are generated based on the backbone model Lightning (4-step), while a building image (left) and the car images are generated using the Lightning (1-step) model. Another building image (right) is generated with the Turbo.

Same as the settings of Section 5.2, we utilize GPT4o [1] to obtain responses. The exact question posed is:

"Please identify the components that contribute to the creativity of the second image (ours) compared to the first image (original). The components can be selected from shapes, colors, and textures. If none apply, state no."

As illustrated in Figure W, images generated with **C3** demonstrate enhancements in various aspects of creativity. For instance, in the case of "a creative teddy bear," the generated image becomes more creative through enriched color (a more vibrant, colorful body and scarf) and texture (a fluffy appearance).

E. Failure Cases

(a) Case 1: Degrade in Functionality

(b) Case 2: Limited Creative Features

Figure X. Failure cases of **C3**. Lightning (1-step) is used for 'bicycle', 'victorian painting', and 'bracelet' and Turbo is used for 'teapot', 'donut' and 'jacket'.

While we show **C3** successfully enhances the creative generations of pretrained Stable Diffsuion-based models, there exist failure cases. We present two main failure cases in Figure X. In Case 1, **C3** produces genuinely novel images, but this comes at the cost of reduced functionality. This occurs primarily due to the limitations of the current usability score, which is inadequate for assessing detailed functionality. In Case 2, the generated images exhibit enhanced creativity (e.g., modern fashion in Victorian-style painting), but they do not significantly deviate from common patterns: a woman in a painting, a round-shape donut, a colorful bracelet, and a color-patched vinyl jacket. We presume that these patterns originate from biases present in the pre-trained models. We leave the generation of creative outputs that mitigate model bias for future work.

F. Extension on Non-SD Models

Unet-based Model. In Figure Y, we applied C3 to Kandinsky 3.0. Similarly to Stable Diffusion XL (SDXL), structural/color variation occurs more on down blocks, while up blocks increase the filter effect, such as contrast.

Figure Y. (Left) Block-wise amplified images on Kandinsky 3.0. (Right) Results of C3 applied on Down blocks of Kandinsky 3.0.

Transformer-based Model. In Figure Z, we applied C3 on Hunyuan-DiT. Amplifying blocks 5–20 yields the most variation, while blocks beyond 20 show minimal change even at maximum amplification. Nevertheless, the impact of C3 on transformer-based models requires careful study, which we leave for future work.

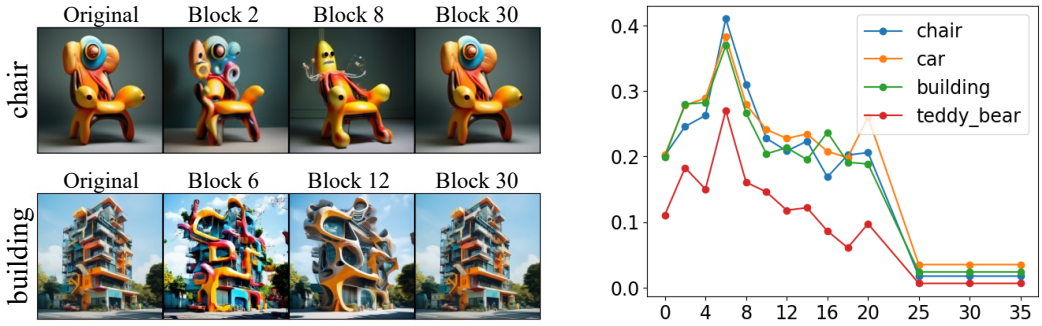


Figure Z. (Left) Images generated with amplified features for the ith transformer block of Hunyuan-DiT. (Right) Block-wise LPIPS score between the original image and the amplified image.

G. Related Work

G.1. Stable Diffusion Models

Diffusion models [9, 24] learn to generate images from random noise through a denoising process, demonstrating stable training and remarkable performance in image and video generation compared to GANs. Latent Diffusion model [20], instead of directly processing images during the denoising process, learns the encoded latent vectors of the images, successfully reducing the size of the model. Stable Diffusion XL (SDXL) [17], a widely used Latent Diffusion model, is publicly available with accessible source code and trained models. Diffusion models undergo denoising processes in T steps in order to generate a sample. To accelerate high-quality sample generation, distilled models have been developed. SDXL-Turbo (Turbo) [21] employs Adversarial Diffusion Distillation (ADD) to condense the multi-step denoising process of a large pre-trained teacher model into 1-4 steps while maintaining high quality. However, due to the limitations of adversarial training, Turbo cannot prevent mode collapse. SDXL-Lightning (Lightning) [12] combines ADD with progressive distillation to address mode collapse while quickly generating high-quality samples in one or a few steps.

SDXL and its distilled variants share a U-net structure in the backbone to generate a latent noise. The U-net structure is composed of three down blocks, decreasing the resolution of the internal feature maps while increasing the number of

channels, a middle block, and three up blocks, increasing the resolution of the feature maps again.

Recent research [10] indicates that up blocks in Stable Diffusion models are primarily associated with style while the structure is preserved. This aligns with our observation that up blocks minimally alter the creative style when amplified, although some filter effects are introduced. Other models may exhibit different block characteristics; for instance, Disco-diff [28], influenced by StyleGAN, employs discrete latents for each block and trains end-to-end, potentially redefining block roles.

G.2. Creative Generation

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 been shown to enhance 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 significant advancements in generating creatively inspired outputs, they necessitate burdensome additional training or optimization. To the best of our knowledge, there is no training-free approach for generating creative samples.

G.3. Feature Map Manipulation

GAN Dissection [3] pioneered techniques to visualize and control the inner workings of GANs by identifying "interpretable units" that correspond to distinct objects within generated images. This approach enables precise feature control, allowing specific objects to be added or removed within scenes, making it effective for targeted scene composition. Expanding on internal feature manipulation to the text-to-image diffusion models, research into internal features of diffusion models is advancing rapidly [4, 6, 11, 26]. Especially, P+ [26] takes a step further by introducing multi-layered conditioning, enabling flexible visual manipulation and enhanced image customization through layer-specific control. Our work builds on these foundations by exploring feature manipulation in the Fourier domain for even greater creativity control. FreeU [23] also operates within the Fourier domain, leveraging Fourier transforms on skip connections to reduce low-frequency information, ultimately improving image fidelity in diffusion models. In contrast, our approach amplifies creativity by applying Fourier-based manipulation directly to the backbone features in the specific blocks rather than skip connections. This distinction allows our method to focus on enhancing creative aspects, making it suited for generating novel and expressive images.