SELF-CREATIVE TEXT-TO-OBJECT GENERATION USING SEMANTIC-AWARE SPATIAL WEIGHTING

Anonymous authors
Paper under double-blind review

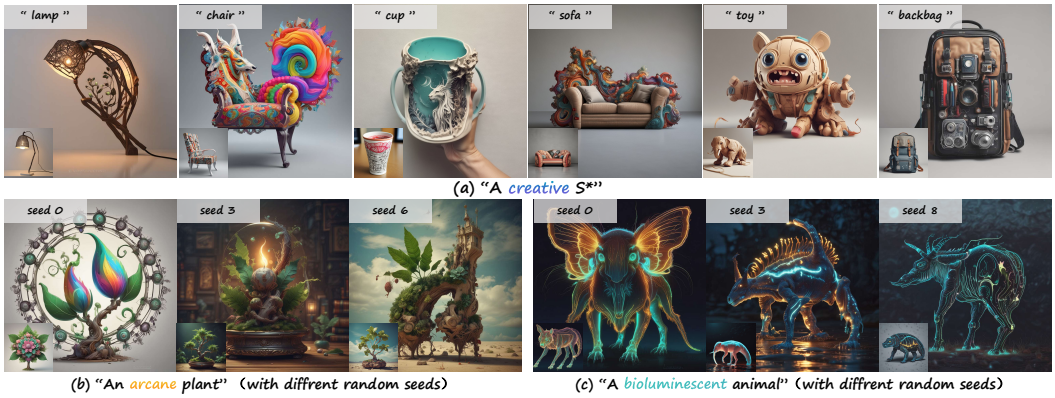

Figure 1: **Generating Creative Objects by our SCDiff**. (a) shows practical yet innovative designs using the prompt: *"A creative S*"*. (b) and (c) demonstrate its ability to generate distinct conceptual objects within a broad category by varying the random seed.

ABSTRACT

Instilling creativity in text-to-image (T2I) generation presents a significant challenge, as it requires synthesized images to exhibit not only visual novelty and surprise, but also artistic value. Current T2I models, however, are largely optimized for literal text-image alignment with their data distribution, and their noise prediction networks constrain the generation to high-probability regions, consequently generating outputs that lack authentic creativity. To address this, we propose a **S**elf-**C**reative **D**iffusion (**SCDiff**) model for meaningful T2I generations featuring two core modules: a learnable spatial weighting (LSW) module and a visual-semantic mixing loss (VSML). The LSW module designs a parametric Kaiser-Bessel window to reinforce central image features, fostering novel and surprising generation. The VSML module introduces a dual loss function: a similarity loss constrains that the new images align with its textual description, while a diversity loss maximizes its distinction from the original image, enhancing both semantic value and visual novelty. Extensive experiments demonstrate that our model substantially improves creativity, semantic alignment, and visual coherence, offering a simple yet powerful framework for generating creative objects.

1 INTRODUCTION

The rapid advancement of text-to-image (T2I) generation technologies Kingma & Welling (2014); Ramesh et al. (2021); Rombach et al. (2022) has revealed their vast potential across diverse fields including digital art, game design, film production and advertising Cao et al. (2023); Liao et al. (2025); Croitoru et al. (2023); Qin (2023). Consequently, the generation of creative content has gained significant attention Richardson et al. (2024); Li et al. (2024); Xiong et al. (2024); Elgammal et al. (2017); Han et al. (2025), as applications increasingly demand the production of new and meaningful visual objects. A central challenge in this pursuit is generating creative (*e.g., novel, surprising, and valuable* Boden (2004)) objects.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

While most T2I diffusion models Rombach et al. (2022); Lin et al. (2024); Sauer et al. (2024); Ho et al. (2020); Podell et al. (2024) can generate images that align with simple category texts, they often fail to produce creative objects Oppenlaender (2024); Malakouti & Kovashka (2025); Tan et al. (2024), even when prompts include words like "creative" or more complex descriptors. This stems from the standard denoising process, which converges toward high-density regions (*i.e.*, common samples) of the training data distribution, resulting in conventional and homogenized outputs.

One notable approach aimed at mitigating this issue is the Creative Concept Catalyst (C3) method Han et al. (2025). C3 operates by transforming intermediate U-Net features Ronneberger et al. (2015) into the Fourier domain using the Fast Fourier Transform (FFT) Bracewell (2000), selectively amplifying low-frequency components with a fixed window to foster creative elements. However, C3 suffers from an inherent mathematical limitation: according to the convolution theorem Gonzalez & Woods (2018), applying a hard cutoff in the frequency domain is equivalent to convolving the feature map with a Jinc function. As illustrated in Figure 2 with the prompt *"a colorful green lizard"*, this function possesses infinite spatial support characterized by oscillating side lobes. Consequently, the editing operation cannot be spatially localized, manifesting as severe global noise pollution in background regions.

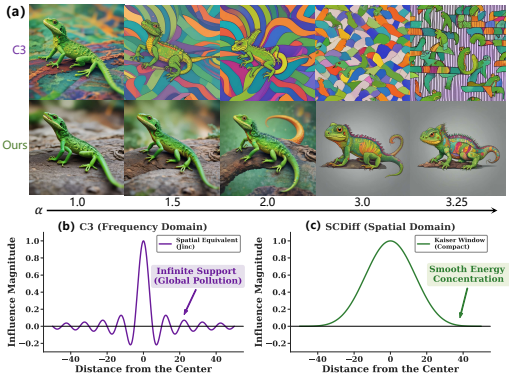

Figure 2: Comparison of **SCDiff** (Ours) against existing methods (e.g., **C3** Han et al. (2025)). Increasing the amplification factor in C3 introduces ghosting artifacts.

To overcome these issues, we propose a **Self-Creative Diffusion (SCDiff)** model to effectively produce creative objects. Guided by the three classical characteristics of creativity—novelty, surprise, and value Boden (2004), our model adaptively enhances spatial features to boost novelty and surprise, and leverages textual semantics to maintain meaningfulness. Specifically, SCDiff comprises two modules. The **Learnable Spatial Weighting (LSW)** module employs a learnable Kaiser-Bessel window Kaiser (1974) to adaptively reinforce central image features. While the radius R is user-defined to constrain the editing extent, the module dynamically adjusts the window profile parameter β and amplification factor α. This induces a distribution shift in the central region, effectively disrupting the model's tendency to converge towards the mean patterns of the training distribution. Constrained by the diversity and similarity losses, the model is guided to explore low-probability yet semantically plausible sub-manifolds, thereby generating outputs with enhanced novelty and surprise. A **Visual-Semantic Mixing Loss (VSML)** module introduces a dual-loss mechanism to balance the output. A similarity loss constrains the generated image to align with its textual description, ensuring semantic value and preventing meaningless outputs from the LSW. A diversity loss increases visual distinction from the original image, thereby explicitly promoting novelty. Our contributions are summarized as follows: **(1)** We propose a learnable spatial weighting method that performs controllable feature enhancement directly in the spatial domain. Its tunable parameters enable the model to accentuate target regions while maintaining global structural consistency. **(2)** We introduce a semantic constraint module that adjusts the amplification strength and window profile based on visual-semantic feedback, enabling a balance between visual novelty, surprise, and semantic meaningfulness. **(3)** Through extensive experiments on multiple T2I baselines, we demonstrate that our method improves semantic alignment, creativity, and visual coherence, and concurrently mitigates ringing artifacts, resulting in a powerful and robust solution for enhancing creativity.

2 Related Work

Recent diffusion models, such as Imagen Saharia et al. (2022), DALLE-2 Ramesh et al. (2021), and Stable Diffusion Sauer et al. (2024); Lin et al. (2024); Rombach et al. (2022), have achieved unprecedented levels of image synthesis quality, shifting the focus toward enhancing the diversity and creativity of generated images. Meanwhile, Large Language Models OpenAI (2024) enhance text-to-image (T2I) generation by leveraging its advanced language understanding, providing creative guidance to produce images that better align with the user's intent.

Recent studies focus on combinational visual creativity, generating novel objects by merging features from multiple concepts into a unified and coherent representation Kamb & Ganguli (2025); Vinker

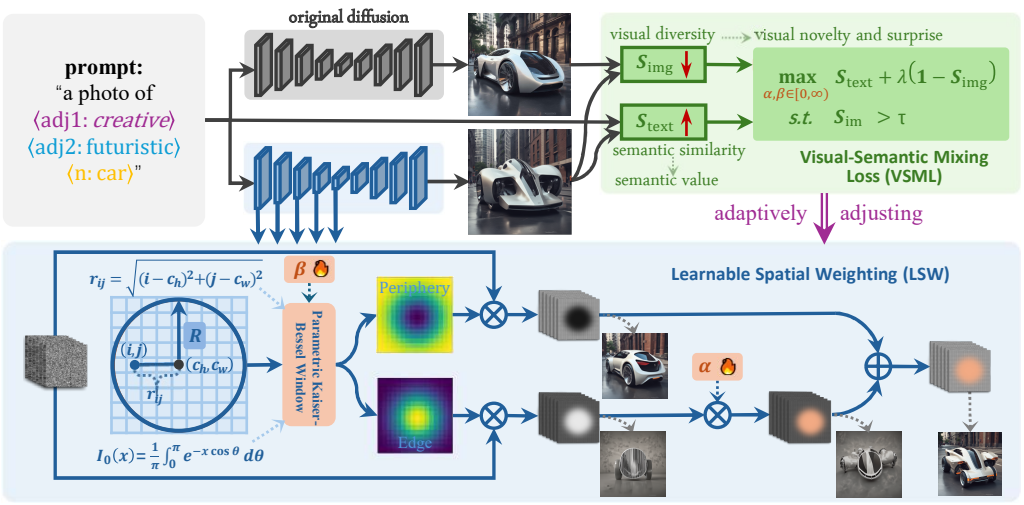

Figure 3: Overview of our SCDiff. It advances diffusion models through two novel modules: (1) a parametric Kaiser-Bessel window (LSW) modulating U-Net encoder features, and (2) a dual VSML loss that optimizes these parameters for self-creative generation.

et al. (2023); Chen et al. (2025b;a); Feng et al. (2025). PartCraft Ng et al. (2024) allows for the flexible composition of parts from diverse objects, producing novel yet holistically plausible results that maintain structural coherence. TP2O Li et al. (2024); Zhang et al. (2025) introduces a balanced swap-sampling strategy to creatively combine textual descriptions of two objects. ATIH Xiong et al. (2024) employs a balanced loss function along with a similarity score to fuse object texts with images, generating new distinctive objects.

Concurrently, another category of innovation methods are dedicated to enhancing creativity within the boundaries of given categories Richardson et al. (2024); Yu et al. (2025); Lu et al. (2024). ConceptLab Richardson et al. (2024) combines vision-language models with diffusion priors to drive the creation and hybrid exploration of novel concepts. C3 Han et al. (2025) enhances novelty and unconventional features by selectively amplifying intermediate activations during the denoising process, without requiring additional training. In contrast, we propose a feature enhancement paradigm based entirely on the spatial domain. We utilize a parameterized Kaiser-Bessel window to perform precise local modulation directly on intermediate features. This method effectively stimulates creative features while avoiding spectral aliasing and ringing artifacts caused by frequency truncation.

3 Self-Creative Diffusion Model

We introduce **Self-Creative Diffusion (SCDiff)** for generating creative objects (Fig. 3). Our model is built upon two key components: a *Learnable Spatial Weighting* (**LSW**) module that applies an adaptive Kaiser–Bessel window to U-Net features, and a *Visual-Semantic Mixing Loss* (**VSML**) module that combines a similarity loss enforcing text alignment with a diversity loss encouraging deviation from the original image, jointly enhancing semantic consistency and visual novelty.

3.1 Learnable Spatial Weighting

Creativity research formalizes generation as search within a constrained conceptual space Boden (2004). In the Creative Systems Framework (CSF) Wiggins (2006), the conceptual space $\mathcal{C}$ is governed by a rule set $\mathcal{R}$ and defined as the subset of a space $\mathcal{U}$ satisfying an acceptability condition:

$$\mathcal{C} = \{c \mid c \in \mathcal{U} \wedge \mathcal{A}(\mathcal{R}, c) \text{ is true}\}, \tag{1}$$

where $\mathcal{A}$ denotes the acceptability operator. The transition from an existing concept c_{in} to a novel concept c_{out} is realized by a Traversal operator $\mathcal{T}$ and an Evaluation operator $\mathcal{E}$:

$$c_{\text{out}} = \mathcal{T}(\mathcal{R}, \mathcal{E})(c_{\text{in}}). \tag{2}$$

We instantiate the CSF framework within the latent manifold of a diffusion model. Let the intermediate feature maps $\mathbf{x}$ of the denoising U-Net (or DiT blocks) constitute the universal space $\mathcal{U}$. To induce structural novelty, we propose an LSW module that modulates the encoder feature $\mathbf{x}$ spatially to

operationalize the rule set $\mathcal{R}$, while α acts as the traversal operator $\mathcal{T}$ to drive generation from the original concept c_{in} (I_{org}) to novel structural variants c_{out}.

Given the structured prompt template T_p, *"a photo of* $< adj_1 :$ *creative or* $adj_2 >$ *<object name>."*, the base diffusion model generates an original image I_{org}. We extract feature maps $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$ from one encoder block selected from $\{Down_0, Down_1, Down_2, Mid\}$, excluding upsampling layers to avoid high-frequency noise (Appendix B). The modulated features pass through skip connections to the decoder to generate diverse objects.

To realize the traversal operator $\mathcal{T}$ within the CSF framework, we introduce spatial feature modulation; however, traditional hard-threshold windows with fixed radius Han et al. (2025) often introduce abrupt truncation, causing boundary discontinuities and spectral ringing, as illustrated in Fig. 2. To mitigate this, our LSW module adopts a parametric smooth Kaiser–Bessel window, defined as:

$$\mathbf{x}^{\text{KB}}(\alpha, \beta) = \mathbf{x} \odot (1 - \mathbf{w}(\beta)) + \alpha \cdot \mathbf{x} \odot \mathbf{w}(\beta), \tag{3}$$

where $\mathbf{w}(\beta) = \textsc{Repeat}(w(\beta), BC)$ replicates the window $w(\beta)$ across BC feature maps, α is a learnable amplification factor, and β controls the window shape. The window $w(\beta) \in \mathbb{R}^{H\times W}$ is:

$$w_{ij}(\beta) = \begin{cases} \dfrac{I_0\left(\beta\sqrt{1-(r_{ij}/R)^2}\right)}{I_0(\beta)}, & r_{ij} \le R, \\ 0, & r_{ij} > R, \end{cases} \tag{4}$$

where $r_{ij} = \sqrt{(i - H/2)^2 + (j - W/2)^2}$ denotes the radial distance of pixel (i, j) from the image center, R is a cutoff radius hyper-parameter, and $I_0(\cdot)$ is the zeroth-order modified Bessel function of the first kind, $I_0(x) = \frac{1}{\pi}\int_0^{\pi} e^{-x\cos(\theta)}\, d\theta$ Watson (1995). The smooth tapering of w_β provides a differentiable spatial weighting for feature maps, allowing adjustment of large-scale structures.

The rationale for this spatial-domain design can be justified by the theoretical limitations of frequency-domain structure isolation. *Frequency-domain editing methods* (e.g., C3 Han et al. (2025)) typically isolate the structure by applying a circular low-pass filter (mask M) in the frequency domain. However, by the convolution theorem Gonzalez & Woods (2018), an ideal circular cutoff in continuous frequency in the spatial domain corresponds to convolution with a *Jinc function*:

$$\mathcal{F}^{-1}M(u, v) \propto J_1(2\pi f_c r)/r, \tag{5}$$

where J_1 is the first-order Bessel function Watson (1995), f_c denotes the spectral cutoff frequency, and r represents the spatial radial distance. This Jinc function has **infinite spatial support** and exhibits oscillating side-lobes ("ringing") that decay asymptotically but never vanish completely (Fig. 2(b)). Crucially, the spatial support of this kernel is infinite:

$$\text{Supp}(\mathcal{F}^{-1}M) = \{r \mid r \in [0, \infty)\}, \tag{6}$$

As a result, such hard-cutoff filtering can affect the entire spatial domain, causing semantic drift in background regions (C3 in Fig. 2(a)). In contrast, *Spatial-Domain methods* like our Kaiser-Bessel window $w(\beta)$ are **compactly supported**, satisfying the following boundary condition: $w_{ij}(\beta) \equiv 0, \quad \forall r_{ij} > R$. This mathematical locality guarantees that $\mathbf{x}^{\text{KB}}(\alpha, \beta) \equiv \mathbf{x}$ outside the local region of interest (Fig. 2(c)). This strict locality helps limit background perturbations, thereby mitigating the noise and "ghosting" artifacts observed with frequency-based modulation (ours in Fig. 2(a)).

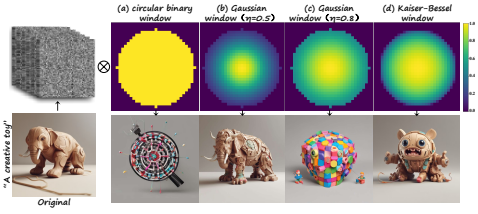

Figure 4: Results using different spatial window profiles: (a) circular binary, (b) Gaussian (η=0.5), (c) Gaussian (η=0.8), and (d) our Kaiser-Bessel window.

While compact support suppresses global artifacts, the profile within $r_{ij} \le R$ determines whether the decoder interprets the modulation as geometric transformation or mere texture variation. We therefore examine three candidate windows: (1) A naive *circular binary window* creates a sharp discontinuity which the diffusion model often misinterprets as a semantic boundary, causing unnatural radial discontinuities that prevent the object from blending seamlessly with the background (Fig. 4(a)). (2) *Gaussian windows* provide a smooth baseline but concentrates energy less effectively under the same spatial support: to ensure a gradual decay to zero at the boundary R, a small width η is required, which concentrates amplification in a narrow central zone(Fig. 4(b)).

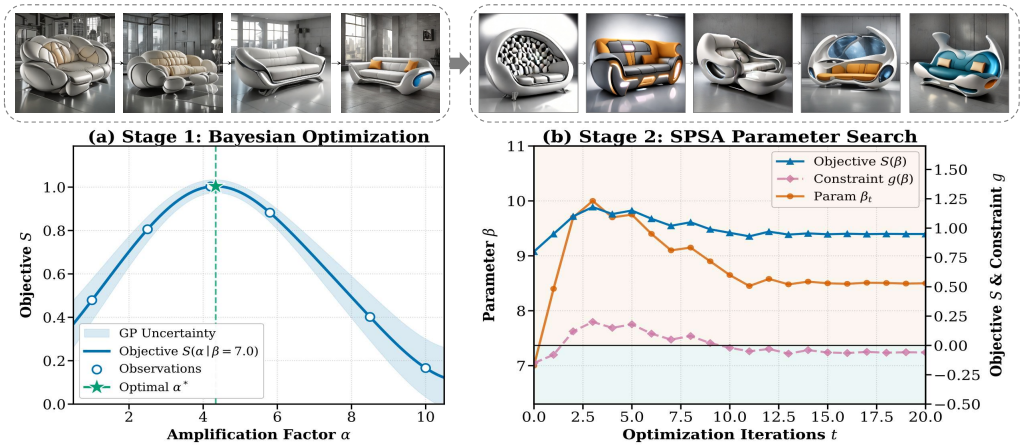

Figure 5: Visualizing the updated process of our VSML using the prompt *'a futuristic sofa'*.

The decoder typically interprets this localized energy peak as texture variation rather than geometric transformation. Conversely, increasing η to flatten the profile reintroduces boundary discontinuities and thus boundary artifacts (Fig. 4(c)). (3) The *Kaiser-Bessel window*, inspired by the classical Kaiser approximation to Discrete Prolate Spheroidal Sequences (DPSS), in contrast, provides a tunable profile for spatial feature modulation Kaiser (1974); Harris (1978). By adjusting the shape parameter β, the window controls the radial taper within the prescribed spatial support. This shape promotes global structural reorganization beyond local texture changes (Fig. 4(d)).

3.2 Visual-Semantic Mixing Loss

The LSW module employs a tunable Kaiser–Bessel window (Eq. 3) for spatial modulation, parameterized by the amplification factor α, the shape parameter β, and the user-defined hyperparameter R. Its centroid (x_c, y_c) defaults to the image center but is adjustable for localized off-center enhancement. This module generates diverse object images, denoted as $I(\alpha, \beta)$, through a complete denoising trajectory. To dynamically optimize α and β—thereby allowing the model to automatically balance energy distribution and spectral smoothness—we introduce a Visual-Semantic Mixing Loss (VSML) in this subsection, which serves as the evaluation operator $\mathcal{E}$ in the CSF framework. The objective of VSML is twofold: first, to enhance the visual novelty and surprise by increasing the diversity between $I(\alpha, \beta)$ and the original image I_{org}; and second, to preserve semantic value by constraining $I(\alpha, \beta)$ using its similarity to the original text prompt T_p. Formally, VSML is defined as follows:

$$\max_{\alpha,\beta} \mathcal{S}(\alpha, \beta), \quad \text{s.t. } g(\alpha, \beta) \leq 0, \tag{7}$$

where $\mathcal{S}(\alpha, \beta) = S_{\text{text}}(T_p, I(\alpha, \beta)) + \lambda \cdot (1 - S_{\text{img}}(I_{\text{org}}, I(\alpha, \beta)))$, $g(\alpha, \beta) = \tau - S_{\text{img}}(I_{\text{org}}, I(\alpha, \beta))$, $S(\cdot, \cdot)$ represents the CLIP Radford et al. (2021) cosine similarity, where lower similarity to the original generated image indicates higher novelty following Elgammal and Saleh Elgammal & Saleh (2015), τ acts as a safety margin and is set to $\tau = 0.7$. This constraint ensures the model does not mistakenly reward severely distorted, low-similarity images with high creative scores. Empirically finding parameters is challenging when simultaneously adjusting α and β as α varies widely. Here, we introduce a hierarchical adjustment strategy to learn the parameters α, β. The procedure proceeds as follows. First, with β fixed at its empirically determined value ($\beta = 7.0$), we consider the VSML objective $\mathcal{S}(\alpha, \beta)$ without imposing the constraint $g(\alpha, \beta) \leq 0$. Given the unimodal yet expensive-to-evaluate response of α observed in Fig. 5, we employ Bayesian optimization Snoek et al. (2012) to efficiently determine the optimal amplification factor α^*:

$$\alpha^* = \arg \max_{\alpha \in [1.5, 10.0]} \mathcal{S}(\alpha \mid \beta = 7.0). \tag{8}$$

Since each query of α requires a complete diffusion forward pass, we employ a Gaussian Process surrogate model with the Expected Improvement (EI) acquisition function.

Next, using the resulting α^* together with the user-specified R, we refine the window profile parameter β by solving the constrained optimization problem:

$$\max_{\beta} \mathcal{S}(\beta \mid \alpha^*), \ \ s.t. \ g(\beta \mid \alpha^*) \leq 0. \tag{9}$$

This problem is solved using the Simultaneous Perturbation Stochastic Approximation (SPSA) Spall (1998), a zeroth-order method that estimates pseudo-gradients via simultaneous random perturbations. Further details for Bayesian optimization and SPSA are consolidated in Appendix C. Fig. 5 illustrates this hierarchical optimization process. By strictly adhering to the user-defined spatial constraint ($R = 15$), this adaptive mechanism automatically selects the amplification factor and window profile, ensuring a balance between text alignment and visual deviation from the original generation.

3.3 Optional Parameter Prediction.

The two-stage optimization requires repeated image generation to evaluate candidate parameters. As a lower-cost alternative, we train a lightweight MLP supervised by parameters from Bayesian optimization and SPSA to directly predict the LSW amplification and window-shape parameters from a prompt's text embedding. Details and comparison with search-based selection are in Appendix D.

4 Experiments

4.1 Experiment Settings

Datasets. We extend the COF dataset Zhang et al. (2025) in two ways. First, we build *CreaCOF* by prepending "creative" to each of the 950 original class names. Second, we construct *AdjCOF* by assigning 10 tailored adjectives to each of the 95 superclasses and pairing them with the 10 class names within that superclass, yielding 9,500 prompts with more complex syntax and higher generative difficulty. Details are in Appendix F.

Implementation Details. We evaluated SCDiff using the one-step version of SDXL-Lightning Lin et al. (2024) and SDXL-Turbo Sauer et al. (2024). All experiments were conducted on a single NVIDIA RTX 3090 GPU. The two-stage parameter search required approximately 38 seconds per prompt, after which image generation took 1–2 seconds.

Evaluation metrics. We evaluate SCDiff in terms of ***Semantic Alignment (SA)***, ***diversity***, and ***creativity***. For semantic alignment, we use CLIP similarity Radford et al. (2021) to measure text–image alignment and LLaVA-VQA Liu et al. (2023) to measure object recognition. We assess within-prompt diversity using two higher-is-better metrics: mean LPIPS distance in the AlexNet feature space Zhang et al. (2018) and Vendi Score Askari Hemmat et al. (2024), with Vendi computed over 20 images per prompt; both metrics are computed separately for each prompt. We primarily assess creativity using CREward Han et al. (2026), which reports geometry, material, texture, and overall scores, while GPT-4o ratings and human preferences provide complementary evidence.

Table 1: Quantitative comparison on CreaCOF and AdjCOF.

Model	Method	Semantic Alignment		Within-prompt Diversity		CREward Han et al. (2026)				GPT-4o
		CLIP↑	LLaVA↑	LPIPS↑	Vendi↑	Geometry↑	Material↑	Texture↑	Overall↑	Overall↑
Turbo (CreaCOF)	Orig. Sauer et al. (2024)	0.290	0.742	0.520	1.752	-4.601	-0.301	-1.686	-3.674	6.0
	C3 Han et al. (2025)	0.295	0.784	**0.577**	2.818	-3.472	0.853	-1.209	-2.090	6.4
	Our SCDiff	**0.308**	**0.911**	0.556	**3.411**	**-2.713**	**0.984**	**-0.759**	**-1.499**	**7.8**
Lightning (CreaCOF)	Orig. Lin et al. (2024)	0.282	0.503	0.525	2.097	-2.661	1.247	-1.304	-1.544	8.0
	C3 Han et al. (2025)	0.287	0.629	**0.611**	4.149	-2.441	1.328	-1.120	-1.423	7.8
	Our SCDiff	**0.290**	**0.743**	0.596	**4.232**	**-2.171**	**1.413**	**-0.780**	**-0.799**	**8.8**
Reference (CreaCOF)	SD3.5 Large Esser et al. (2024)	0.305	0.747	0.563	3.088	-4.599	-0.167	-1.471	-3.804	8.0
	FLUX.1 Dev Labs (2024)	0.303	0.820	0.541	3.178	-4.211	-0.002	-2.009	-3.582	7.7
	Qwen Image 2.1 Qwen Team (2026)	0.276	0.713	0.540	2.314	-5.411	0.401	-2.466	-4.513	7.6
	ConceptLab Richardson et al. (2024)	0.220	0.291	0.679	2.562	-3.513	0.684	-1.359	-2.600	8.1
Turbo (AdjCOF)	Orig. Sauer et al. (2024)	0.292	0.515	0.547	2.600	-5.447	-0.818	-2.454	-4.645	7.7
	C3 Han et al. (2025)	0.297	0.612	0.561	3.146	-3.850	0.379	**-1.528**	-2.384	7.1
	Our SCDiff	**0.305**	**0.658**	**0.579**	**3.275**	**-3.813**	**0.459**	-1.605	**-2.132**	**8.7**

4.2 Main Results

To ensure a comprehensive evaluation of our SCDiff, we compare it against C3 Han et al. (2025), ConceptLab Richardson et al. (2024), along with relevant baseline models Sauer et al. (2024); Lin et al. (2024). For a fair comparison, we also adaptively tuned the parameter α of C3.

Quantitative Results. Table 1 reports results on CreaCOF and AdjCOF. For semantic alignment, SCDiff consistently achieves higher CLIP and LLaVA-VQA scores than Orig. and C3 across all three backbone–dataset pairings, supporting better text–image alignment and target-object recognizability. On CreaCOF, SCDiff improves within-prompt diversity over Orig. under both Turbo and Lightning. Compared with C3, Turbo-SCDiff achieves a higher Vendi Score (3.411 vs. 2.818), while its LPIPS is slightly lower (0.556 vs. 0.577) but remains above Orig. (0.520). Lightning-SCDiff shows the same pattern, with a higher Vendi Score (4.232 vs. 4.149) but a slightly lower LPIPS (0.596 vs.



Figure 6: Qualitative Comparisons on *CreaCOF* demonstrate that our method delivers creative and meaningful outputs.

0.611). For creativity, SCDiff leads all four CREward metrics within each backbone group and also exceeds all reference methods on these metrics. On AdjCOF, adjective-modified object prompts steer generation toward specific attributes or styles. While improving text–image alignment, SCDiff achieves the highest overall CREward score (-2.132), LPIPS (0.579), and Vendi Score (3.275) . On the texture subscore, however, C3 scores higher (-1.528 vs. -1.605), though both improve over the Orig. baseline (-2.454). GPT-4o also assigns SCDiff higher overall ratings than Orig. and C3 across all three settings (7.8, 8.8, and 8.7).

To complement automated metrics, we conducted a human preference evaluation on CreaCOF and AdjCOF. SCDiff received the largest share of selections in both settings, 79.1% and 81.4%, respectively, as shown in Table 2. Further details are provided in Appendix G.

Table 2: Human preference results.

Method	CreaCOF	AdjCOF
Orig.	3.6%	1.9%
C3	5.3%	10.4%
GPT-4o	9.5%	3.9%
ConceptLab	2.5%	2.4%
SCDiff	**79.1%**	**81.4%**

Qualitative Results. Figs. 6 and 7 present visual comparisons on both the CreaCOF and AdjCOF benchmarks, respectively. The results demonstrate that our method, SCDiff, generates creative, meaningful, and visually impressive outputs, surpassing related state-of-the-art approaches. Specifically, C3 exhibits noticeable visual artifacts under high amplification factors, substantially compromising both fidelity and creative expression. ConceptLab relies on sub-category extraction from broad parent classes, which often leads to inaccurate semantic alignment with target descriptors and a significant deviation from the intended creative concepts. Meanwhile, although GPT-4o, SD 3.5 Large, and Flux.1 Dev maintain high content coherence, their outputs lack the necessary distributional shift required for meaningful creative generation, thereby limiting their imaginative potential.

In addition, we explored complex textual prompts with contextual expressions. As shown in Fig. 8, these prompts improve image detail and semantic consistency without sacrificing realism, showing

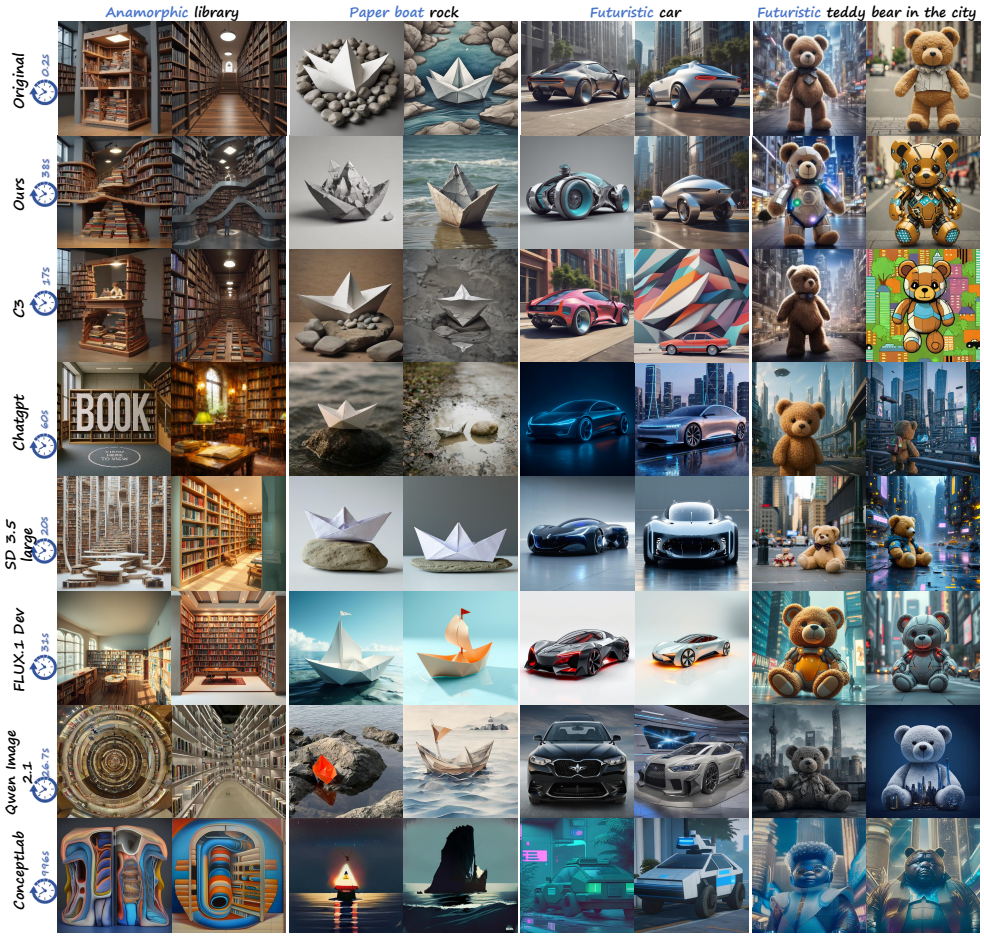

Figure 7: Qualitative Comparisons on *AdjCOF* demonstrate that our method consistently produces visually impressive results.

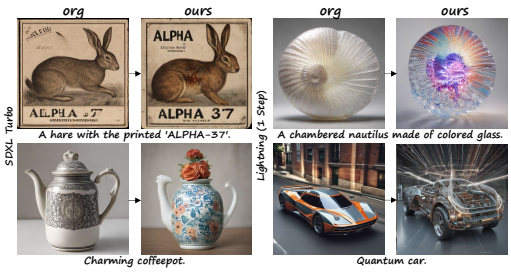

Figure 8: Qualitative results of complex textual inputs using different baseline models.

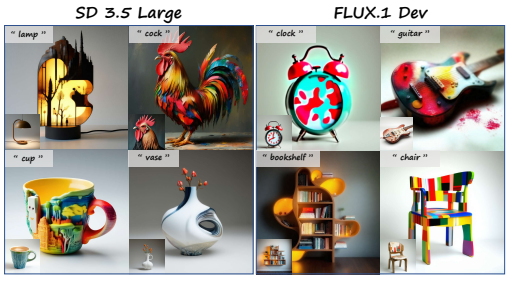

Figure 9: Our method on DiT-based models: (a) SD 3.5, (b) FLUX.

potential for semantically grounded generation. Appendix H provides more results. Specifically, we provide a visualization of the effects of feature amplification across different network layers in Figs. 22-26, and demonstrate the generative diversity of fixed prompts under random seeds in Figs. 18-19.

4.3 Extension to Diffusion Transformers

We evaluate our method on Transformer-based diffusion backbones, with qualitative results shown in Fig. 9. Specifically, for SD 3.5 Esser et al. (2024) and FLUX Labs (2024), latent tokens are first reshaped from sequence format to 2D spatial feature maps for masking, then restored to sequence format for subsequent transformer layers. Despite the absence of skip-connections and hierarchical features in DiT blocks, our method preserves semantic consistency while enhancing diversity, verifying that the core modulation principle is architecture-agnostic.

4.4 Ablation Study

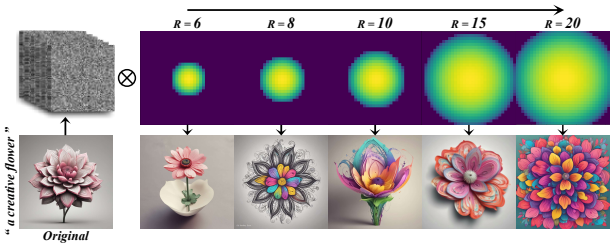

Figure 10: Visual comparisons of different sampling radii (R) under the optimal magnification factor $\alpha = 3.1$ and shape parameter $\beta = 6.7$.



Figure 11: Visual comparisons of the global amplification factors α across different fixed spatial parameters.

Models	Semantic Alignment		Within-prompt Diversity		CREward				GPT-4o
	CLIP↑	LLaVA↑	LPIPS↑	Vendi↑	Geometry↑	Material↑	Texture↑	Overall↑	Overall↑
Baseline w/o LSW and VSML	0.290	0.742	0.520	1.752	-4.601	-0.301	-1.686	-3.674	6.0
w/ LSW w/ Fixed α	0.291	0.815	0.548	1.982	-3.214	-0.350	-0.925	-2.356	6.4
w/ LSW w/ Fixed β	0.301	0.834	0.539	1.645	-2.870	0.724	-1.067	-2.181	6.9
w/ LSW and VSML (**our Full Model**)	**0.308**	**0.911**	**0.556**	**3.411**	**-2.713**	**0.984**	**-0.759**	**-1.499**	**7.8**

Table 3: Quantitative ablation study of our proposed SCDiff.



Figure 12: Ablation analysis.

We conducted an ablation study to evaluate the contributions of each key component in SCDiff, as shown in Fig. 12 and Table 3. For the configuration with $R = 15$, the experiments proceeded through the following progressive modifications, shown from left to right: (i) baseline model without the LSW and VSML modules; (ii) With the LSW module (non adaptive amplification factor α); (iii) With the LSW module (non adaptive Kaiser–Bessel window (β)); (iv) Full model with both LSW and VSML modules. Experimental results confirm that the adaptive design of each component meaningfully improves both image quality and creative effectiveness.

4.5 Parameter Analysis

Sampling radius R. As illustrated in Fig. 10, we visualize the impact of varying the sampling radius R on the *"flower"* concept, while fixing the global amplification factor $\alpha = 3.1$ and the shape parameter $\beta = 6.7$ (derived from our optimization process). Specifically, a large radius causes the flower to occupy the entire canvas, whereas a small radius restricts the deformation to only the central circular region. This observation confirms that R effectively serves as a flexible control knob, allowing users to customize the spatial extent of creative editing according to their design intent, thereby producing diverse visual outcomes.

Amplification factor α. Can optimizing the global amplification factor α compensate for a fixed but suboptimal spatial window? As illustrated in Fig. 11, fixing the spatial parameters (R, β) and increasing α alone reveals limitations. If the fixed window is too small for the concept, edits remain strictly localized even with large α, preventing the global structural reconstruction. Conversely, an oversized window that intrudes into the background leads to structural degradation as α increases, driving identity similarity $\mathcal{S}_{\text{img}}$ below the safety threshold τ. Thus, our two-stage strategy enables multi-scale feature manipulation, enhancing visual novelty while maintaining structural integrity.

5 Conclusion

In this paper, we proposed SCDiff, an adaptive and semantic-aware framework for creative object generation in text-to-image synthesis. We incorporate the U-Net architecture by inserting a Learnable Spatial Weighting (LSW) module into the shallow encoder layers, where a parametric Kaiser–Bessel window selectively amplifies core structural features and fosters the emergence of novel visual patterns. In parallel, we introduce the Visual-Semantic Mixing Loss (VSML) module, which employs a dual-objective loss to jointly encourage semantic alignment and controlled visual divergence, enabling more expressive and creative generations. Extensive experiments across 950 object categories validate that SCDiff significantly outperforms existing methods in terms of visual creativity, semantic faithfulness, and structural coherence. This establishes a new direction for post-alignment creative generation and delivers an efficient tool for digital art and game design.

AI USE STATEMENT

In this work, we used GPT-4o as an evaluator to assess images generated by text-to-image models in terms of novelty, surprise, value, and overall creativity. The resulting scores were reported as quantitative evaluation metrics in our experiments. We reviewed the prompts and scoring criteria, checked the scoring outputs and aggregation procedures, and verified that the reported results matched the original evaluation records. Additionally, we used generative AI tools to assist with manuscript editing and refinement of the presentation of our arguments. Except for the image generation models studied and compared in this work and the automated evaluation described above, we did not use generative AI tools to generate training data, develop or implement the proposed method, or produce other experimental results. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Human evaluation. Our human evaluation involved voluntary participants who evaluated the generated images. Participants were recruited voluntarily through online invitations and received a modest monetary compensation for their participation; this compensation was not contingent on the content of their responses. Before participation, all participants were informed of the study purpose and procedure and provided consent to take part. Participation was voluntary, and participants could stop at any time without penalty. We collected only the evaluation responses necessary for our analysis and did not collect or retain personally identifiable or sensitive information. The study involved no deception, or foreseeable risks beyond those associated with ordinary image-evaluation tasks.

Potential misuse. SCDiff may support creative exploration in art and design. However, like other image generation methods, it could also be misused to produce deceptive visual content or images that misrepresent the appearance or feasibility of products. For deployment, we recommend clearly disclosing AI-generated content and using appropriate provenance and labeling mechanisms to support transparency and traceability.

REFERENCES

Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *ECCV*, pp. 213–229, 2024.

Margaret A. Boden. *The creative mind-Myths and mechanisms*. 2004.

Ronald N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill, third edition, 2000.

Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, and Kangxue Yin. Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4146–4158, 2023.

Haibo Chen, Zhoujie Wang, Lei Zhao, Jun Li, and Jian Yang. Trtst: Arbitrary high-quality text-guided style transfer with transformers. *IEEE Transactions on Image Processing*, 34:759–771, 2025a.

Haibo Chen, Zhiwen Zuo, Lei Zhao, Jun Li, and Jian Yang. Conceptcraft: One-shot personalized text-to-image generation via object-background disentanglement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025b.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869, 2023.

A. Elgammal and Babak Saleh. Quantifying creativity in art networks. In *ICCC*, 2015.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633, 2024.

Fu Feng, Yucheng Xie, Xu Yang, Jing Wang, and Xin Geng. Redefining in dictionary: Towards an enhanced semantic understanding of creative generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18444–18454, 2025.

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 4th edition, 2018.

Jiyeon Han, Dahee Kwon, Gayoung Lee, Junho Kim, and Jaesik Choi. Enhancing creative generation on stable diffusion-based models. In *CVPR*, pp. 28609–28618, 2025.

Jiyeon Han, Ali Mahdavi-Amiri, Hao Zhang, and Haedong Jeong. Creward: A type-specific creativity reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21932–21941, June 2026.

Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

J. F. Kaiser. Nonrecursive digital filter design using the I_0-sinh window function. In *Proceedings of the 1974 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1974.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Black Forest Labs. Flux.1-dev. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024.

Jun Li, Zedong Zhang, and Jian Yang. Tp2o: Creative text pair-to-object generation using balance swap-sampling. In *ECCV*, pp. 92–111, 2024.

Chin-Wen Liao, Hsiang-Wei Chen, Bo-Siang Chen, I-Chi Wang, Wei-Sho Ho, and Wei-Lun Huang. Exploring the application of text-to-image generation technology in art education at vocational senior high schools in taiwan. *Information*, 16(5), 2025.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pp. 34892–34916, 2023.

Jack Lu, Ryan Teehan, and Mengye Ren. Procreate, don't reproduce! propulsive energy diffusion for creative generation. In *ECCV*, 2024.

Sina Malakouti and Adriana Kovashka. Role bias in diffusion models: Diagnosing and mitigating through intermediate decomposition. In *Advances in Neural Information Processing Systems*, volume 38, pp. 84668–84699, 2025.

Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Partcraft: Crafting creative objects by parts. In *ECCV*, pp. 420–437, 2024.

OpenAI. Hello gpt-4o, 2024. URL `https://openai.com/index/hello-gpt-4o/`. Accessed: 2026-01-19.

Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, pp. 3763–3776, 2024.

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 *International Conference on Learning Representations*, 2024.

Jiayang Qin. How does text-to-image ai affect indie game designers and artists? *Journal of Innovation and Development*, pp. 107–111, 2023.

Qwen Team. Qwen-Image-2.1. `https://huggingface.co/Qwen/Qwen-Image-2.1`, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8821–8831, 2021.

Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics(TOG)*, 43 (3), 2024. ISSN 0730-0301.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention(MICCAI)*, pp. 234–241, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, pp. 87–103, 2024.

Jasper Snoek, H. Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Neural Information Processing Systems*, 2012.

J.C. Spall. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems*, 34(3):817–823, 1998.

Zhiyu Tan, Mengping Yang, Luozheng Qin, Hao Yang, Ye Qian, Qiang Zhou, Cheng Zhang, and Hao Li. An empirical study andanalysis of text-to-image generation using large language model-powered textual representation. In *ECCV*, pp. 472–489, 2024.

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics*, 42(6), 2023. ISSN 0730-0301.

George Neville Watson. *A treatise on the theory of Bessel functions*. Cambridge university press, 1995.

Geraint A. Wiggins. Searching for computational creativity. *New Generation Computing*, 24(3): 209–222, 2006.

Zeren Xiong, Zedong Zhang, Zikun Chen, Shuo Chen, Xiang Li, Gan Sun, Jian Yang, and Jun Li. Novel object synthesis via adaptive text-image harmony. In *Proceedings of the 38th Conference on Neural Information Processing Systems*, volume 37, pp. 139085–139113, 2024.

C. Yu, C. Han, and C. Zhang. Dmfft: Improving the generation quality of diffusion models using fast fourier transform. *Scientific Reports*, 15:10200, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zedong Zhang, Ying ji Tai, Jianjun Qian, Jian Yang, and Jun Li. Agswap: Overcoming category boundaries in object fusion via adaptive group swapping. *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025.

A Appendix

This supplementary material provides additional technical details and extended results to support the main paper. First, in Section B, we analyze the impact of feature enhancement across different U-Net blocks. Section C details the hierarchical Bayesian-SPSA optimization procedure to solve the constrained optimization problem. Section D describes the optional parameter predictor, comparing predicted and search-based parameters under identical generation settings. In Section E, we discuss the limitations of our method. Section F describes the COF taxonomy and the strategy for assigning tailored adjectival prompts for complex generation tasks. Section G provides specific details of the human preference study and a detailed breakdown of the final vote counts for each question. Finally, in Section H, we showcase additional qualitative results.



Figure 13: The images generated using the adaptively optimized parameters for each block.

B Discussion on Feature Enhancement: Shallow vs. Deep.

As shown in Figure 13, experiments demonstrate that enhancing downsampling and middle blocks yields the best gains in structural creativity. Applying the module in upsampling layers adds noise because these layers focus on texture refinement rather than altering spatial structure. Therefore, we apply feature enhancement only to downsampling and middle blocks.

C Hierarchical Optimization

C.1 Bayesian Optimization for Amplification Factor α

Surrogate model. We employ a Gaussian Process (GP) as the surrogate for the implicit objective $\tilde{f}(\alpha_k)$, using the Matérn-5/2 kernel:

$$k(\alpha, \alpha') = \sigma^2 \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right), \tag{10}$$

where $r = |\alpha - \alpha'|$. We assume a zero-mean GP prior and independent Gaussian observation noise $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. The length scale l, signal variance σ^2, and noise variance σ_n^2 are jointly inferred by maximizing the marginal likelihood.

Acquisition function. We adopt Expected Improvement (EI), which under the GP posterior $\tilde{f}(\alpha) \mid \mathcal{D}_n \sim \mathcal{N}(\mu_n(\alpha), \sigma_n^2(\alpha))$ admits the closed-form expression:

$$\mathrm{EI}(\alpha) = (\mu_n(\alpha) - \tilde{f}^+)\Phi(Z) + \sigma_n(\alpha)\phi(Z), \tag{11}$$

where $\tilde{f}^+ = \max_{1 \le i \le n} \tilde{f}(\alpha_i)$ is the current best observation, $Z = (\mu_n(\alpha) - \tilde{f}^+)/\sigma_n(\alpha)$, and $\Phi(\cdot), \phi(\cdot)$ are the standard normal CDF and PDF, respectively. EI automatically balances exploitation of promising regions and exploration of high-uncertainty areas.

Optimization procedure. The search space is $[1.5, 10.0]$. We initialize with $n_{\text{init}} = 5$ random samples via Latin hypercube sampling to ensure space coverage, followed by $n_{\text{iter}} = 10$ sequential acquisition iterations. At each iteration, we maximize EI via L-BFGS-B with 10 random restarts to avoid local optima, query the point with maximal EI, observe $\tilde{f}(\alpha_{n+1})$, and update the GP posterior. The optimal amplification factor is selected as the point with maximum posterior mean: $\alpha_k^* = \arg\max_\alpha \mu_n(\alpha)$. With only $n = 15$ total observations, exact GP inference via Cholesky decomposition ($\mathcal{O}(n^3)$) is computationally negligible.

C.2 SPSA for Constrained Window Parameter β

Algorithm. Simultaneous Perturbation Stochastic Approximation (SPSA) Spall (1998) is a zeroth-order optimization method that estimates gradients via simultaneous random perturbations, requiring only two function evaluations per iteration regardless of parameter dimensionality.

At iteration t, we generate a random perturbation vector Δ_t with entries drawn independently from a symmetric Bernoulli distribution (± 1 with equal probability). The pseudo-gradient is estimated as:

$$\hat{\nabla} f(\beta_t) = \frac{f(\beta_t + c_t\Delta_t) - f(\beta_t - c_t\Delta_t)}{2c_t} \odot \Delta_t, \tag{12}$$

where $c_t = c/t^\gamma$ is the decaying perturbation size with $c = 0.1$ and $\gamma = 0.101$, and $\odot$ denotes element-wise product (since $1/\Delta_t^{(i)} = \Delta_t^{(i)}$ for $\Delta_t^{(i)} \in \{-1, +1\}$). The parameter is updated via projected gradient ascent:

$$\beta_{t+1} = \Pi_{[\beta_{\min}, \beta_{\max}]} \left(\beta_t + a_t \hat{\nabla} f(\beta_t) \right), \tag{13}$$

where $a_t = a/t^\alpha$ is the decaying learning rate with $a = 0.5$, and $\Pi(\cdot)$ denotes projection onto the feasible box.

Constraint handling. If the updated β_{t+1} violates $g(\beta_{t+1}) > 0$ (i.e., image similarity below the threshold), we halve the step size $a_t \leftarrow a_t/2$ and resample the perturbation direction Δ_t until a feasible point is found. This ensures that the optimization trajectory remains within the feasible region while maximizing text alignment.

Implementation details. We set the initial value $\beta_0 = 8.0$, with search interval $[\beta_{\min}, \beta_{\max}] = [6.0, 12.0]$ and run for $T = 20$ iterations. During the optimization, β is iteratively updated according to the gradient direction estimated by SPSA, and the final optimized parameter is denoted by β^*.

D Optional Parameter Prediction

Architecture and Training. We use an MLP with two hidden layers of 512 and 256 units to predict LSW parameters from text embeddings. Scaled sigmoid activations constrain each output to the corresponding parameter range used by the search procedure. The MLP is supervised by parameters obtained through the Bayesian optimization and SPSA procedure described in the main text, using a Smooth L1 loss on normalized targets. We train with AdamW at a learning rate of 10^{-3} and a batch size of 32, selecting the checkpoint based on validation loss.

Dataset and Evaluation. We split the 950 development prompts into 808 training prompts and 142 validation prompts. The target parameters used to supervise the parameter predictor are obtained through the aforementioned Bayesian optimization and SPSA procedure. The training set spans

Table 4: Comparison of search-based and MLP-predicted modulation parameters on an independent test set covering 100 unseen object categories. Text-sim and Img-sim denote CLIP similarities to the input prompt and the original generation, respectively. CREward evaluates geometry, material, texture, and overall creativity.

	CLIP similarity		CREward ↑			
Parameter selection	Text-sim	Img-sim	Geometry	Material	Texture	Overall
Search (BO + SPSA)	0.293	0.725	-2.727	1.527	0.165	-0.341
MLP prediction	0.287	0.721	-2.751	1.501	-0.076	-0.381

animals (mammals, birds, fish, reptiles, amphibians, and insects), vehicles (boats, buses, trucks, airplanes, and bicycles), objects and appliances (furniture, containers, musical instruments, and tools), and abstract textures (gradients, symmetry, and geometric patterns). The validation set is used for model checkpoint selection. Final evaluation is conducted on a separate test set covering 100 object categories that are absent from both the training and validation sets. Under identical generation settings, we compare generations obtained using predicted parameters and those obtained using search-based parameters.

Results. As shown in Table 4, the parameters predicted by the MLP achieve text–image and image–image similarity scores of 0.287 and 0.721, respectively, closely matching the corresponding scores of 0.293 and 0.725 obtained using search-based parameters. These results indicate comparable prompt alignment and image similarity between the two parameter-selection approaches. We further evaluate the effect of replacing search with prediction on creative generation using CREward, which assesses geometry, material, texture, and overall creativity. The predicted and search-based parameters achieve overall CREward scores of -0.381 and -0.341, respectively. Their geometry scores (-2.751 vs. -2.727) and material scores (1.501 vs. 1.527) are closely matched, while prediction yields a noticeably lower texture score (-0.076 vs. 0.165). In terms of efficiency, search-based parameter selection takes 38 s on average, whereas MLP prediction requires only 0.02 s. Together, these results characterize the trade-off between computational cost and generation performance when using the optional predictor.



Figure 14: Qualitative results for intra-class diversity under the prompt "a creative animal".

E Limitation

E.1 Structural Constraints in Fine-Grained Generation

Our method operates under the constraint of semantic fidelity, exhibiting limitations when processing fine-grained categories with strong structural priors. As shown in Fig. 14, under broad concepts (e.g., "a creative animal"), the model generates pronounced diversity in both structure and appearance; however, when the prompt is constrained to a specific species, the model attempts structural variations yet these tend to converge to similar silhouette patterns, rendering structural differences across samples barely perceptible to human observers, with innovations primarily reflected in texture pattern variations. To alleviate this tension, future work could introduce explicit cross-category blending mechanisms, enabling more directed creative exploration while maintaining semantic consistency.

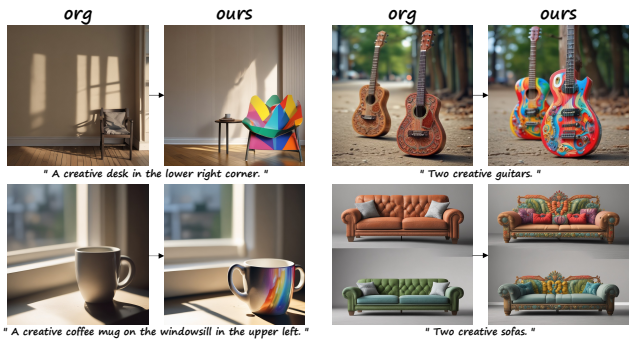

Figure 15: Qualitative examples of SCDiff for off-center objects and multi-object scenes. Left: enhancement of a corner-positioned object using an adjusted spatial anchor. Right: generation in a multi-object scene with moderate amplification.

E.2 Spatial Prior Assumptions

To accommodate the inherent central-cropping bias of dominant T2I backbones, **SCDiff** defaults to a center-aligned window, with quantitative evaluation focused on single-object creativity. The spatial anchor (x_c, y_c) is user-adjustable and supports off-center object enhancement. Fig. 15 (left) shows precise enhancement for corner-placed objects; (right) shows coherence in multi-object scenes under moderate amplification factors. Generalizing this localized control to complex multi-subject environments remains future work, which could be facilitated by integrating autonomous object-localization priors to drive the LSW module.

F Datasets

We adopt the COF dataset introduced by AGSwap Zhang et al. (2025) as our evaluation benchmark. This dataset effectively reorganizes ImageNet-1k categories based on the WordNet hierarchy and manual curation to ensure visual similarity. It comprises 95 visually consistent superclasses (e.g., retaining 'dog' while excluding 'wild dog') with 10 balanced subclasses each. To rigorously assess the model's generative novelty across varying difficulty levels, we derive two distinct benchmarks from this structure: a Simple Dataset, constructed by appending the prefix "creative" to each of the 950 subclasses to evaluate basic diversity; and a Complex Dataset, which assigns 10 tailored adjectives to each superclass, yielding 9,500 prompt variants to challenge the model with higher structural complexity.

G Human Preference Study

To evaluate the effectiveness of our method and users' perceptions of creativity, we conducted a comprehensive human preference study comparing our method against state-of-the-art baselines: C3, Org, GPT-4o, and ConceptLab. The evaluation focused on two distinct categories: OOD Adjectives, which involve complex attribute bindings and material conflicts (e.g., "Frozen flame vase", "Rock paper boat"), and Creative Generation, which involves open-ended prompts (e.g., "Creative bed"). The detailed vote distributions for these categories are presented in Table 6 and Table 5, respectively. Figure 17 further illustrates the percentage distribution of votes across all compared methods. Figure 16 presents examples of the human evaluation questionnaires. A total of 193 participants completed the survey, each evaluating four CreaCOF and six AdjCOF prompt sets, yielding 1,930 selections. Participants were asked to select the image that best combined creativity with alignment to the text prompt. Overall, SCDiff received 943 of 1,158 selections (81.43%) on AdjCOF and 611 of 772 selections (79.15%) on CreaCOF. In the OOD Adjectives category (Table 6), our method demonstrated exceptional capability in resolving conflicting semantic attributes. For instance, in the "Futuristic car" case, our method secured a dominant **85.49%** of the votes, significantly outperforming the second-best method, C3 (8.29%). Similarly, for abstract material combinations such as "Frozen flame vase" and "Rock paper boat," our method garnered **80.31%** and **82.38%** of the preferences, respectively, attesting to its robustness in synthesizing coherent structures

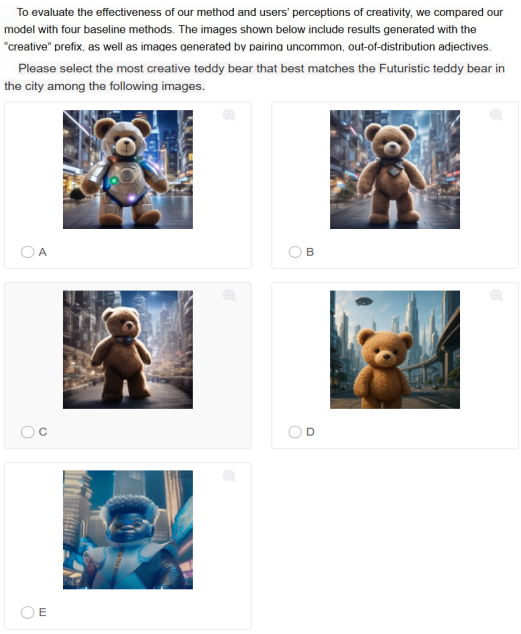

Figure 16: The user interface used for the human evaluation task. Participants selected the image that best combined creativity with alignment to the text prompt.

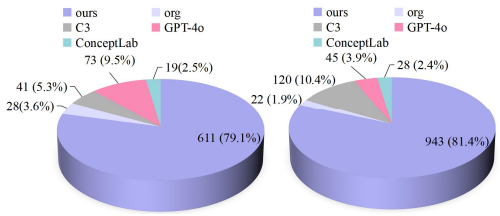

Figure 17: Human preference comparison of different methods on creative prompts (left, CreaCOF) and adjective prompts (right, AdjCOF).

Table 5: Human preference results on **Creative Generation**. This category assesses the visual novelty and imagination of objects generated from open-ended "Creative [class]" prompts.

Method / prompt	A(Ours)	B(C3)	C(Org)	D(GPT-4o)	E(ConceptLab)
Creative table	159(82.38%)	9(4.66%)	5(2.59%)	19(9.84%)	1(0.52%)
Creative cactus	156(80.83%)	11(5.7%)	9(4.66%)	11(5.7%)	6(3.11%)
Creative bed	131(67.88%)	9(4.66%)	12(6.22%)	33(17.1%)	8(4.15%)
Creative mirror	165(85.49%)	12(6.22%)	2(1.04%)	10(5.18%)	4(2.07%)

from contradictory textual descriptions. Although C3 occasionally ranked second (e.g., 17.62% for "Mountain architecture"), it remained substantially lower than our approach (73.58%).

In the Creative Generation category (Table 5), our method maintained a strong lead. Specifically, for prompts requiring significant structural deviation such as "Creative mirror" and "Creative table," our approach garnered an overwhelming 85.49% and 82.38% of user votes, respectively. Even in categories with broad design spaces like "Creative bed," where GPT-4o showed relatively stronger performance (17.10%), our method maintained a dominant lead with 67.88%. Results show our framework aligns better with human aesthetic preferences and creative expectations than existing state-of-the-art methods.

Table 6: Human preference results on **OOD Adjectives**. This category evaluates the model's capability in handling complex attribute bindings and material conflicts.

Method / prompt	A(Ours)	B(C3)	C(Org)	D(GPT-4o)	E(ConceptLab)
Frozen flame vase	155(80.31%)	29(15.03%)	5(2.59%)	4(2.07%)	0(0%)
Futuristic teddy bear in the city	163(84.46%)	12(6.22%)	1(0.52%)	10(5.18%)	7(3.63%)
Iridescent storm clouds	159(82.38%)	19(9.84%)	3(1.55%)	4(2.07%)	8(4.15%)
Mountain architecture	142(73.58%)	34(17.62%)	2(1.04%)	12(6.22%)	3(1.55%)
Rock paper boat	159(82.38%)	10(5.18%)	8(4.15%)	7(3.63%)	9(4.66%)
Futuristic car	165(85.49%)	16(8.29%)	3(1.55%)	8(4.15%)	1(0.52%)



Figure 18: More visual results. Additional samples generated using the prompt "a creative toy" are presented.

H More Results

We present an extensive gallery of generated samples to further demonstrate the capabilities and interpretability of our framework. First, in Fig. 18–21, we showcase the generative diversity of SCDiff. By fixing the textual prompt while varying the random seed, these results illustrate the model's ability to produce a wide range of distinct, creative interpretations for a single concept. Second, in Figures 22–26, we provide a layer-wise visualization of the feature amplification process. By applying the LSW module to different encoder layers (e.g., Down and Mid blocks) under a fixed seed and prompt, we demonstrate how specific modulations influence the output—thereby validating the effectiveness of our hierarchical spatial control.

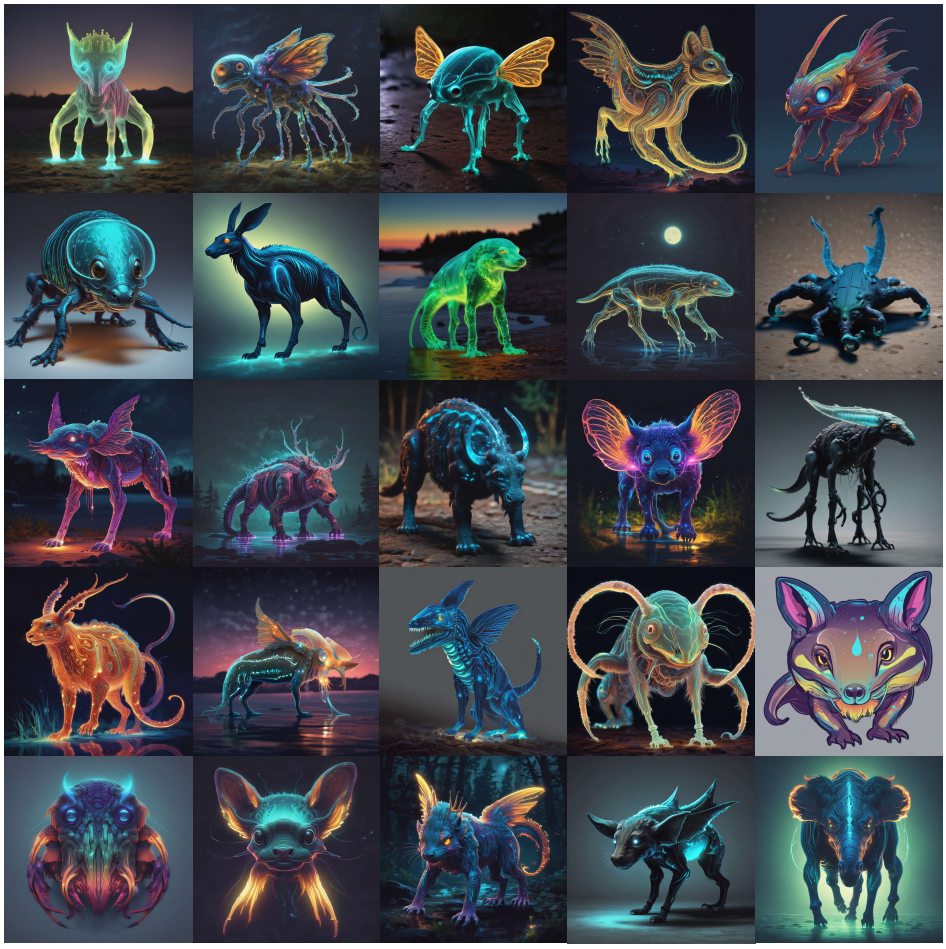

Figure 19: More visual results. Additional samples generated using the prompt "a bioluminescent animal" are presented.

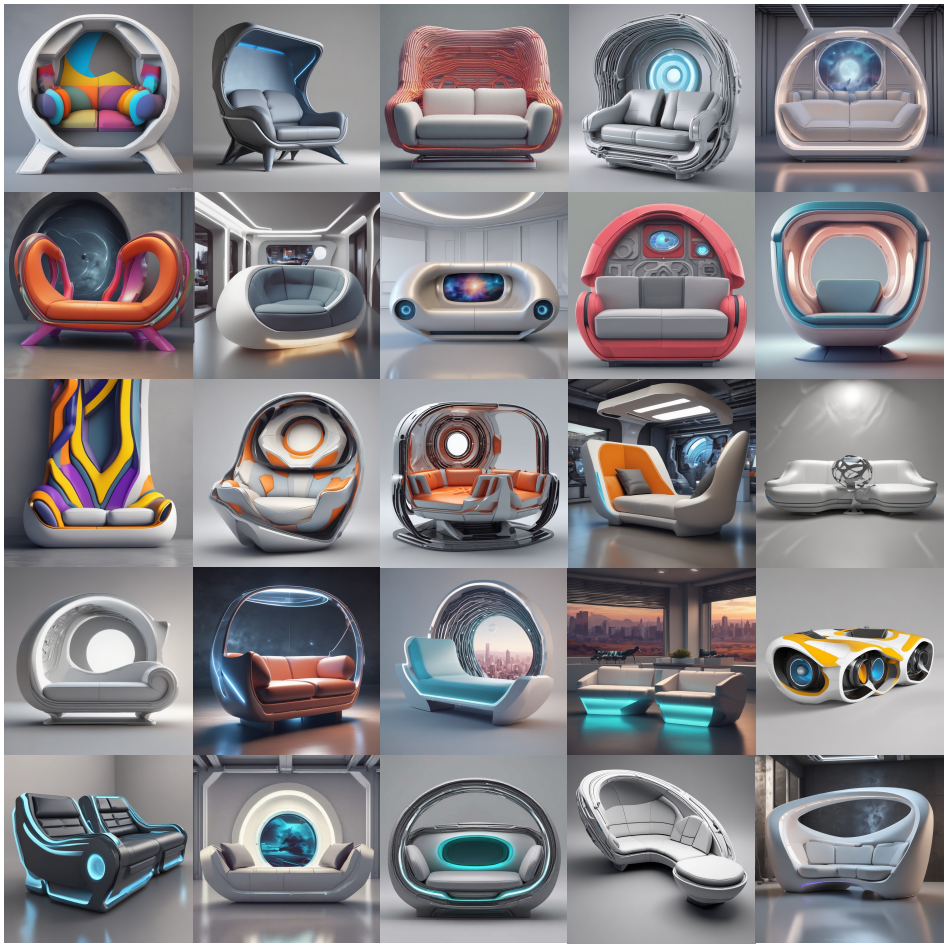

Figure 20: More visual results. Additional samples generated using the prompt "a futuristic sofa" are presented.

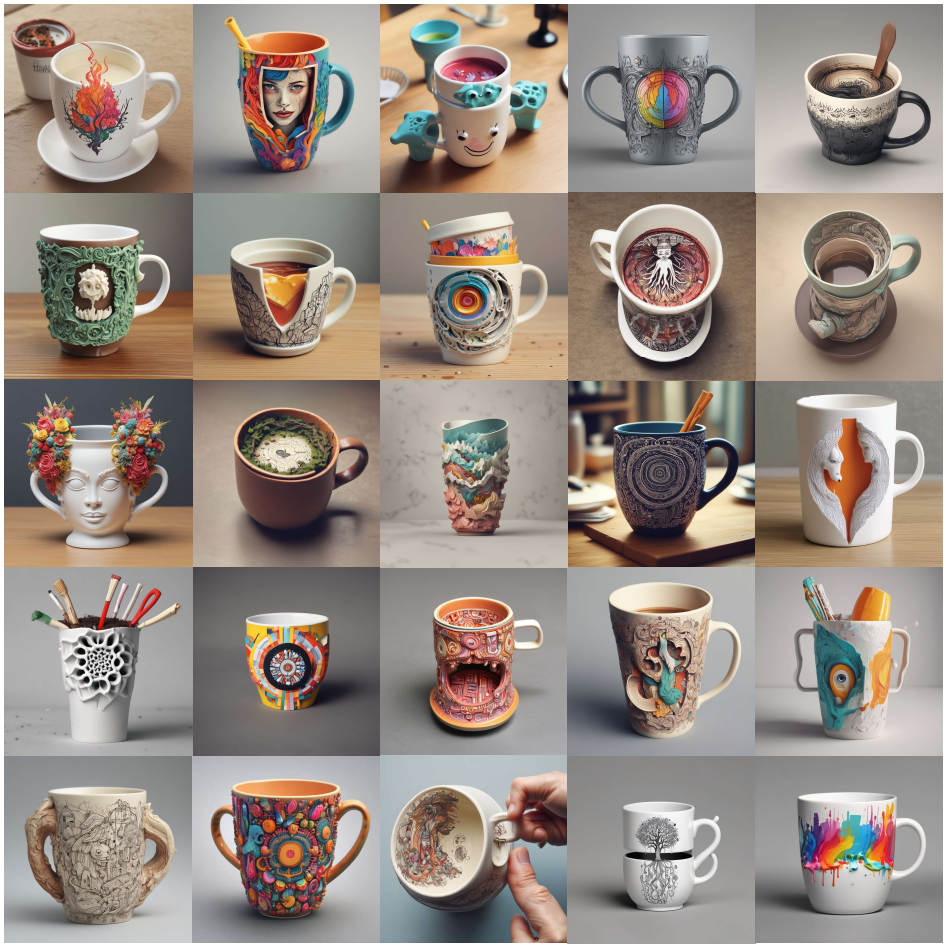

Figure 21: More visual results. Additional samples generated using the prompt "a creative cup" are presented.

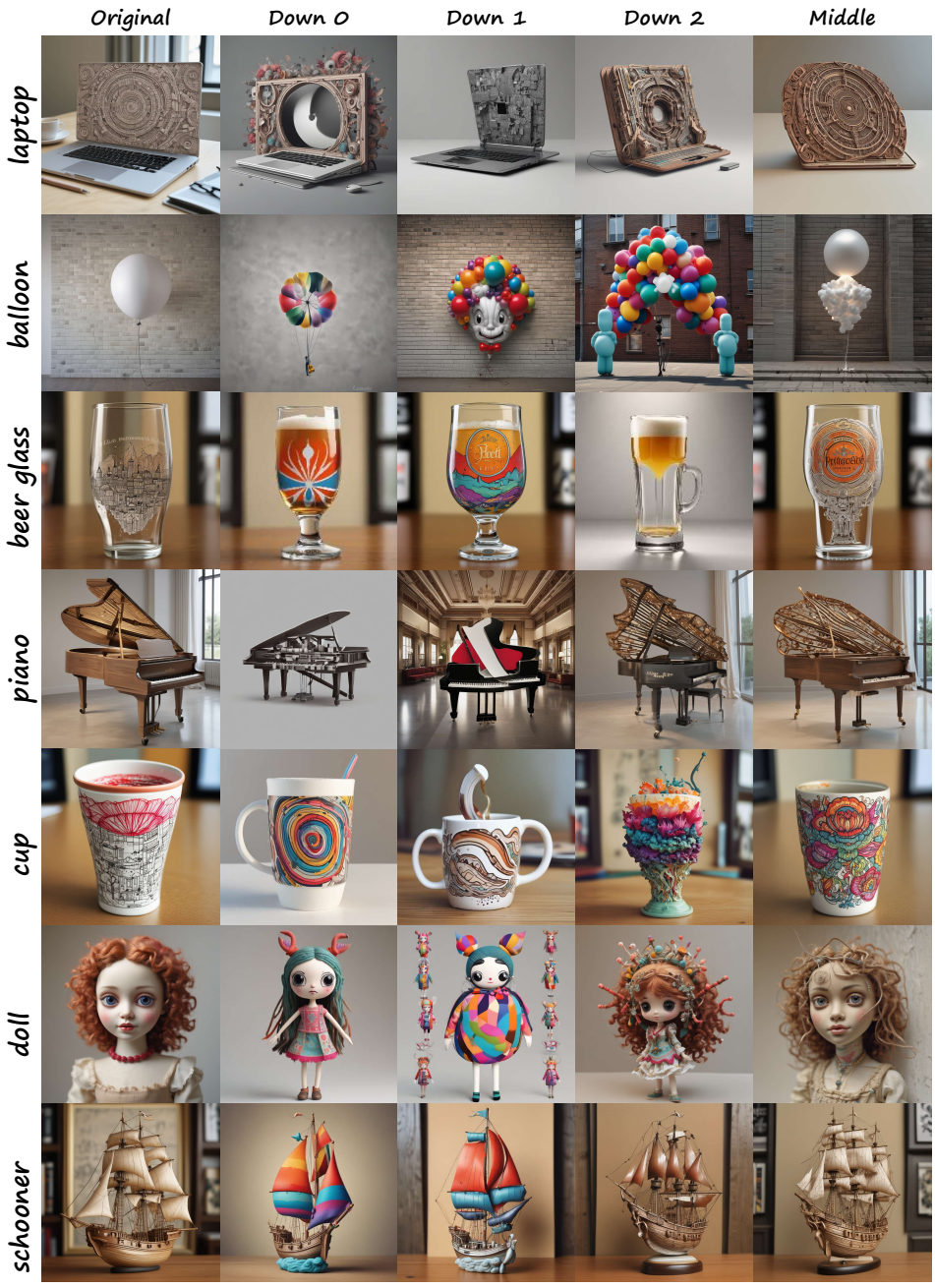

Figure 22: More visual results. Visualization of feature amplification across different U-Net layers, confirming the model's versatility in generating novel content at every structural level.

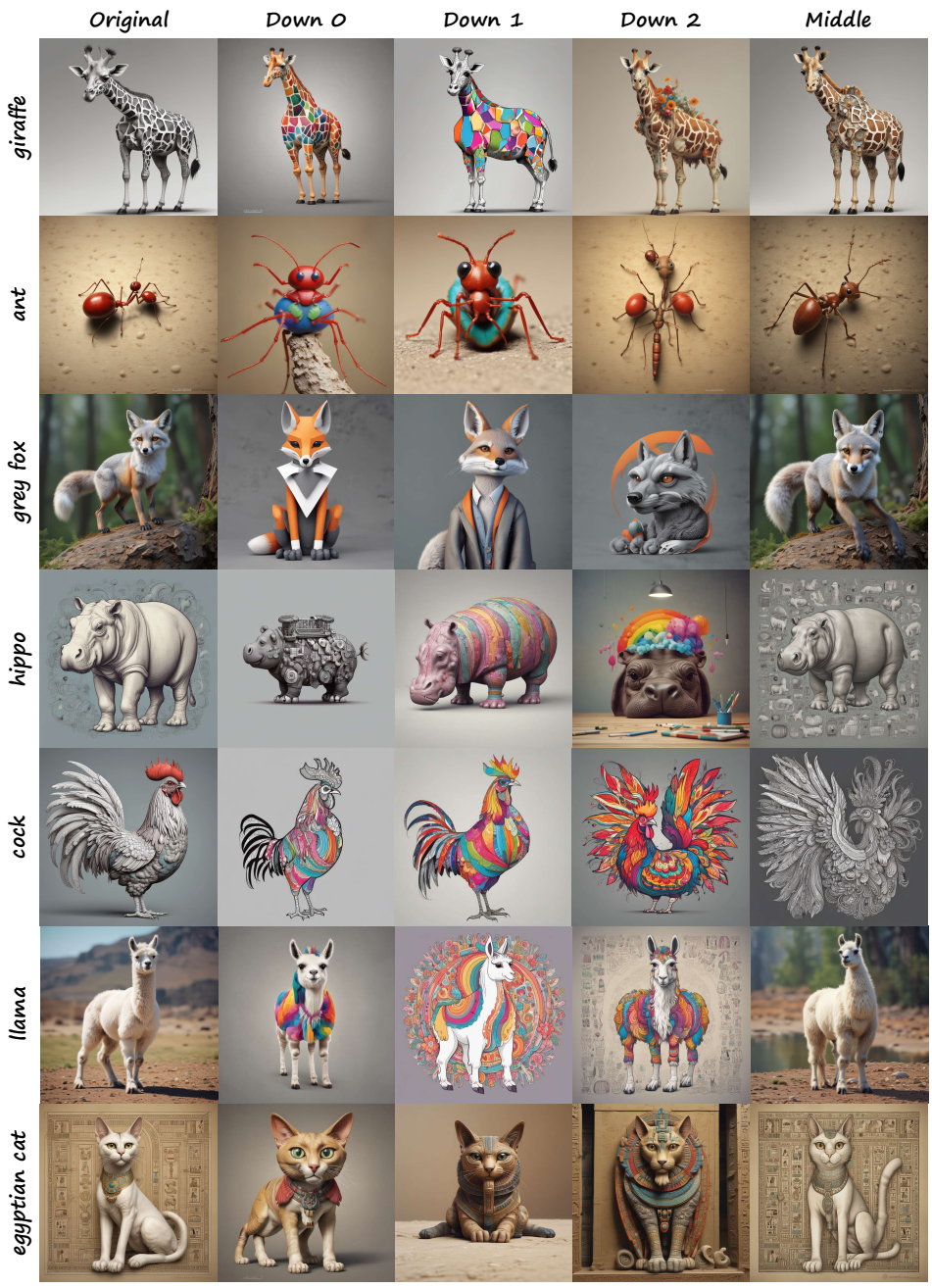

Figure 23: More visual results. Visualization of feature amplification across different U-Net layers, confirming the model's versatility in generating novel content at every structural level.

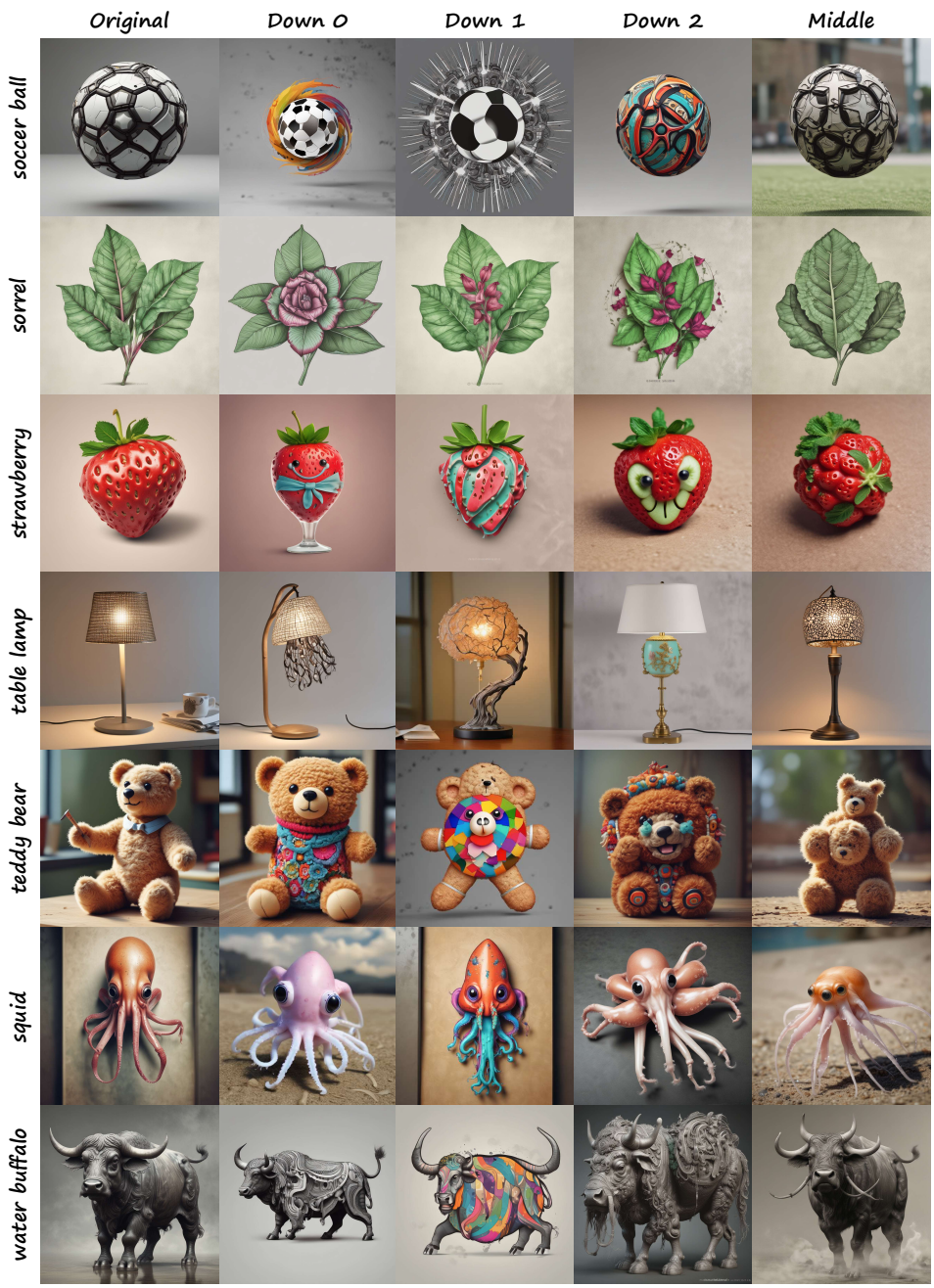

Figure 24: More visual results. Visualization of feature amplification across different U-Net layers, confirming the model's versatility in generating novel content at every structural level.

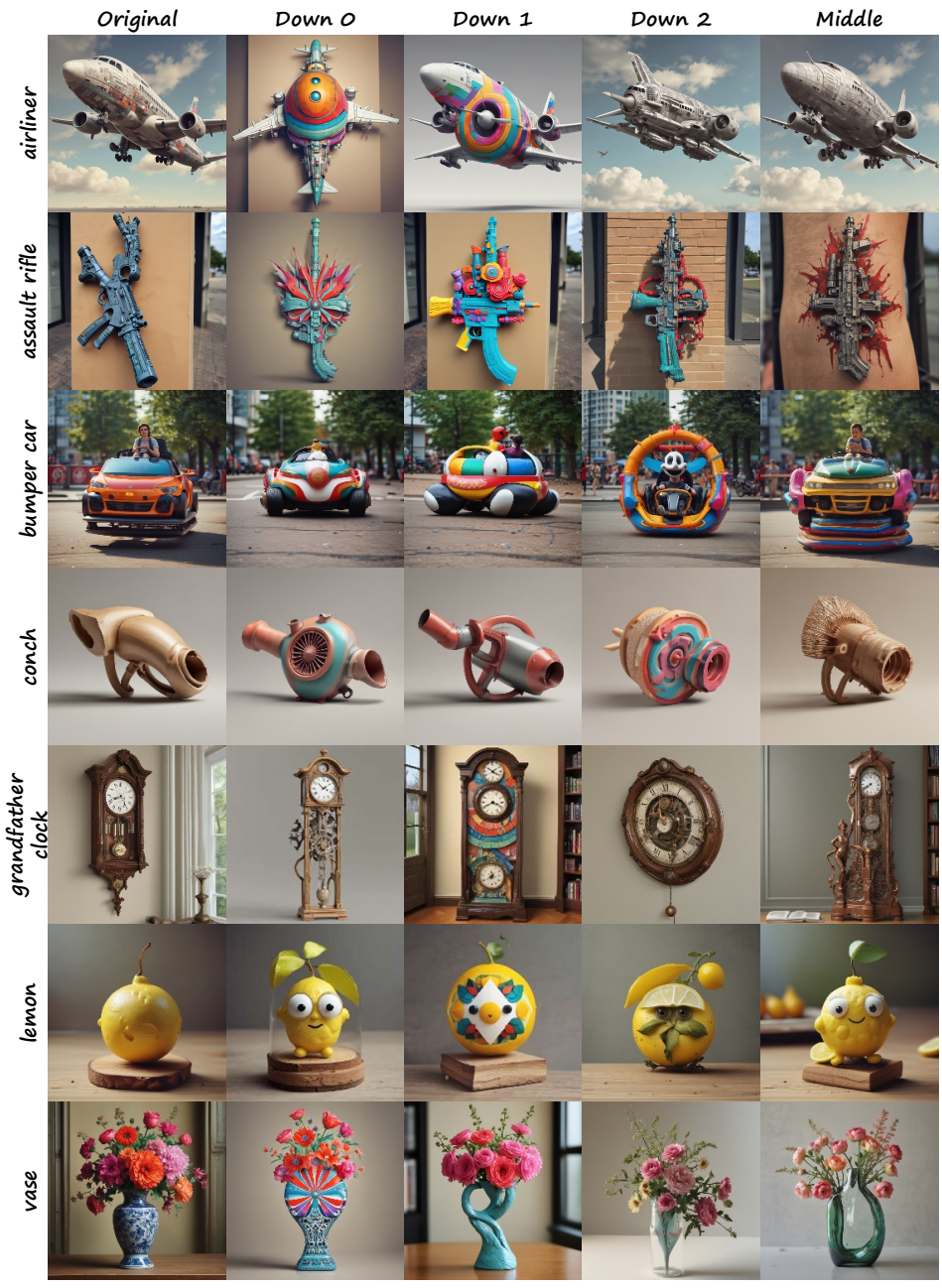

Figure 25: More visual results. Visualization of feature amplification across different U-Net layers, confirming the model's versatility in generating novel content at every structural level.

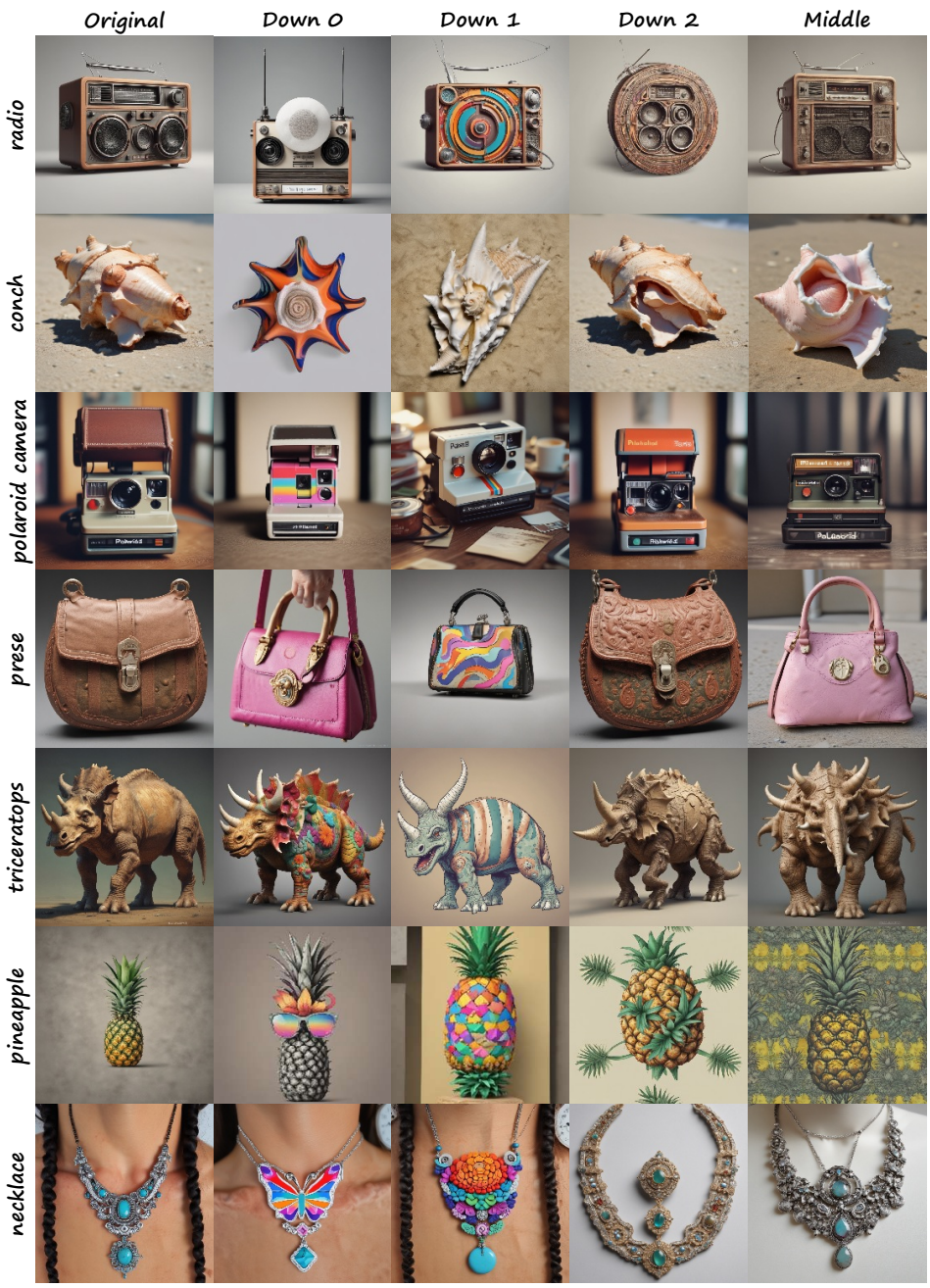

Figure 26: More visual results. Visualization of feature amplification across different U-Net layers, confirming the model's versatility in generating novel content at every structural level.