PROOF: PERTURBATION-ROBUST NOISE FINETUNE VIA OPTIMAL TRANSPORT INFORMATION BOTTLENECK FOR HIGHLY-CORRELATED ASSET GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The diffusion model has provided a strong tool for implementing text-to-image (T2I) and image-to-image (I2I) generation. Recently, topology and texture control have been popular explorations. Explicit methods consider high-fidelity controllable editing based on external signals or diffusion feature manipulations. The implicit method naively conducts noise interpolation in manifold space. However, they suffer from low robustness of topology and texture under noise perturbations. In this paper, we first propose a plug-and-play **P**erturbation-**RO**bust n**O**ise **F**inetune (*PROOF*) module employed by Stable Diffusion to realize a trade-off between content preservation and controllable diversity for highly correlated asset generation. Information bottleneck (IB) and optimal transport (OT) are capable of producing high-fidelity image variations considering topology and texture alignments, respectively. We derive the closed-form solution of the optimal interpolation weight based on optimal-transported information bottleneck (OTIB), and design the corresponding architecture to fine-tune seed noise or inverse noise with around only 14K trainable parameters and 10 minutes of training. Comprehensive experiments and ablation studies demonstrate that PROOF provides a powerful unified latent manipulation module to efficiently fine-tune the 2D/3D assets with text or image guidance, based on multiple base model architectures.

1 INTRODUCTION

Controllable T2I and I2I are challenging and meaningful tasks for asset creation. Previous diffusion control models try to implement structure or appearance-aligned generation explicitly, mainly by feature-level modulation Lin et al. (2024); Mo et al. (2024); Epstein et al. (2023), adapter injection Mou et al. (2024); Zhao et al. (2023); Ye et al. (2023), and model fine-tuning based on external structure or appearance signals Zhang et al. (2023); Gal et al. (2023); Ruiz et al. (2023; 2024). Explicit methods are dependent on cumbersome user control guidance, which hinders topological diversity and appearance robustness as well. On the contrary, we pay attention to the implicit noise-level manipulation on the inherent latent space, where we conduct a trade-off of diversity, structure, and appearance simultaneously.

Recently, test-time noise searching Ma et al. (2025); Zhou et al. (2025) has proved that golden noise plays an important role in diffusion performance for semantic alignment. Other latent manipulation methods, e.g., UnCLIP Ramesh et al. (2022), Kwon et al. (2023), also focus on generating semantic-aligned variants. These works have a fundamental task distinction compared with PROOF. We assume the noise latent has been semantic-aligned, and conduct content-aligned variants with robust structures and textures preservation. We briefly introduce our motivation as follows.

Gaussian noise inherently encodes contextual information. It is supposed to adaptively inject diverse information into the source content while adversarially preserving the original content distribution. This fidelity-diversity trade-off needs to learn a pixel-wise minimal sufficient representation of the noise latent. Inspired by information bottleneck Tishby & Zaslavsky (2015); Schulz et al. (2020), we compress the content latent for topology alignment in an implicit view of the mutual information.

Furthermore, spatial attention is important to improve the contextual perception and appearance robustness. Noise features are distributed randomly without obviously recognizable patterns. There-

Figure 1: Content-diversity tradeoff: given a noise latent of a content, naive noise blending with interpolation weight λ generates uncontrollable topology and appearance. PROOF finetunes noise latent where adaptively injecting the perturbation based on the optimal transported information bottleneck. The structure and appearance statistics from the content are preserved well, with concurrently controllable diversity. **Res** means the optimized area of PROOF compared with naive blending.

fore, it is supposed to distribute attention in a coordinated manner to eliminate excessive local attention. However, traditional QKV attention uses Softmax, which lacks this global attention distribution ability. Inspired by Sinkhorn optimal transport Cuturi (2013); Kim et al. (2024), we apply the doubly stochastic activation constraint to better model the global feature relationships in noise space. This optimally transported attention exhibits significant appearance fidelity. More remarkably, we derive the closed-form solution of the Sinkhorn-regularized IB interpolation weight, which is the theoretical foundation of the PROOF architecture. More details are represented in Sec. 4.3.

As shown in Figure 1, the mainstream implicit approach, i.e., naive noise interpolation with a per-pixel constant weight λ for original noise and (1-λ) for another noise perturbation, fails to preserve the structure and appearance statistics of the original content. In our task definition, the assets for content and naive blending are not highly correlated due to substantial inconsistency of structure and appearance. In contrast, our PROOF adaptively blends pixel-wise perturbations via activation optimization in noise space, based on the proposed Optimal-Transported Information Bottleneck module, thereby facilitating precise asset variations. Our paper presents several significant contributions, mainly including three folds:

1. We first explore the structure and appearance-aligned 2D/3D asset generation by means of perturbation-robust noise representation learning rather than other explicit control manners, such as attention matrices, intermediate activations, or external control signals. Remarkably, test-time *PROOF* demands merely brief training while maintaining full disentanglement from the diffusion model's forward and denoising process.

2. We present an efficient and effective Optimal-Transported Information Bottleneck module that provides a trade-off between content preservation and mode variety. IB prevents the learning from mode collapse, and OT promotes higher faithfulness of textures. Moreover, we derive the closed-form solution of the Sinkhorn-regularized IB interpolation weight. This mathematical derivation is aligned with the information flow of OTIB, which provides a solid theoretical foundation for OTIB.

3. Our proposed PROOF is capable of being adaptive for multiple asset creation tasks, base architectures, and model checkpoints. Compared with state-of-the-art structure and appearance-aligned approaches, comprehensive experimental analyses demonstrate that PROOF is the first perturbation-robust plug-and-play implicit controller for pre-trained T2I models. Furthermore, PROOF is superior to other diversity-inducing methods, such as entropy regularization and contrastive objective.

2 Related Work

We briefly introduce diffusion control methods, diffusion seed manipulation, and information compression works in this section.

Diffusion control. On one hand, pre-trained T2I foundational models Rombach et al. (2022) are potentially able to generate diverse images taking advantage of the random noise initialization. On the other hand, uncertainty from the Gaussian noises makes it hard to synthesize credible images with a certain topology or texture. To address this matter, previous diffusion control methods com-

pose different adapters independently Mou et al. (2024); Zhao et al. (2023), or conduct adaptively feature modulations Zhang et al. (2023); Lin et al. (2024), and model finetune Gal et al. (2023); Ruiz et al. (2023) to facilitate alignment of internal diffusion knowledge and external control signals.

Topology alignment SD-based methods have demonstrated strong generalization capabilities and composability while maintaining high creation quality Li et al. (2023); Zhao et al. (2023); Yang et al. (2023); Avrahami et al. (2023b); Zheng et al. (2023); Wang et al. (2024); Zhou et al. (2024). External control signals include Canny edge, depth map, human pose, line drawing, HED edge drawing, normal map, segmentation mask (used in Zhang et al. (2023); Zhao et al. (2023)), as well as 3d mesh, point cloud, sketch (used in Lin et al. (2024)), etc. FreeControl Mo et al. (2024) manipulates the specific-class linear semantic subspace to employ structural guidance. Semantic signal usually possesses higher freedom than low-level vision signals. Note that our PROOF does not depend on any external structure control signal.

Texture alignment methods try to realize I2I by image prior embedding or few-shot weight adaptation. General I2I methods extract global semantic embedding from the referenced images Zhao et al. (2023); Ye et al. (2023); Mou et al. (2024). Personalized model concerning specific concept needs pretrained T2I diffusion finetuning based on a small set of image samples Ruiz et al. (2023); Gal et al. (2023); Avrahami et al. (2023a); Po et al. (2024); Ruiz et al. (2024). FreeControl Mo et al. (2024) uses intermediate activations as the appearance representation, similar to DSG Epstein et al. (2023). However, our PROOF achieves superior appearance alignment performance without personalized concept data or model fine-tuning.

Diffusion seed. Previous diffusion control methods only treat Gaussian noise as a flexible random generation seed Zhang et al. (2023); Zhao et al. (2023); Ye et al. (2023); Zheng et al. (2023); Wang et al. (2024); Zhou et al. (2024); Ruiz et al. (2023); Gal et al. (2023); Avrahami et al. (2023a); Po et al. (2024); Ruiz et al. (2024). They constrain the pre-trained diffusion model using external structure or textural data. Nevertheless, some diffusion inversion works Yang et al. (2025); Song et al. (2020); Mokady et al. (2023) show high-fidelity image reconstruction and editing. Seed searching Ma et al. (2025) is beyond the denoising steps for high-quality image generation. These methods establish the critical role of noise representation, which is demonstrated by Figure 1 as well. Therefore, we explore the implicit structure and appearance alignment based on noise in this paper.

Information bottleneck. Information bottleneck (IB) Tishby & Zaslavsky (2015) plays a representation trade-off between information compression and information preservation for neural learning tasks. Furthermore, VIB Alemi et al. (2017) leverages variational inference to facilitate the IB neural compression. IBA Schulz et al. (2020); Gao et al. (2021) polishes the attribution information based on KL divergence Csiszár (1975) to effectively disentangle relative and irrelative information concerning the classification task. We will introduce our information bottleneck in Section 3, 4.

3 PRELIMINARIES

3.1 PROBLEM SETTING

Given source noise N_{Orig} and injected noise N_{Div} are from a consistent distribution $\mathcal{N}(\mu_G, \sigma_G^2)$, where μ_G and σ_G represent the means and standard deviations. Then, the modulated manifold of 2D/3D asset can be formulated as follows Schulz et al. (2020):

$$N_{Out} = \lambda N_{Orig} + (1 - \lambda) N_{Div}, \tag{1}$$

where λ is the blending weight as the hyperparameter or learned prior, N_{Div} is the noise perturbation. Given N_{Out} as z_t, the latent diffusion model Rombach et al. (2022) conducts a denoising process on the compressed latent from the Gaussian noise distribution. The denoised manifold of the pre-trained diffusion model is calculated as follows:

$$\tilde{z}_0 = \frac{z_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(z_t, c, t)}{\sqrt{\bar{\alpha}_t}}. \tag{2}$$

where ϵ_θ is the denoising propagation parameter, t is the diffusion timestep, c means prompt, α_t means the noise scheduling parameter at timestep t, while $\bar{\alpha}_t$ indicates the cumulative product of α from step 1 to t. Given $\tilde{z}_0$, we obtain highly correlated assets via the Decoder of VAE.

Naive noise interpolation based on a constant λ and other diversity-inducing methods (e.g., entropy regularization, contrastive objective) are not robust to perturbation from N_{Div}. Our PROOF learns

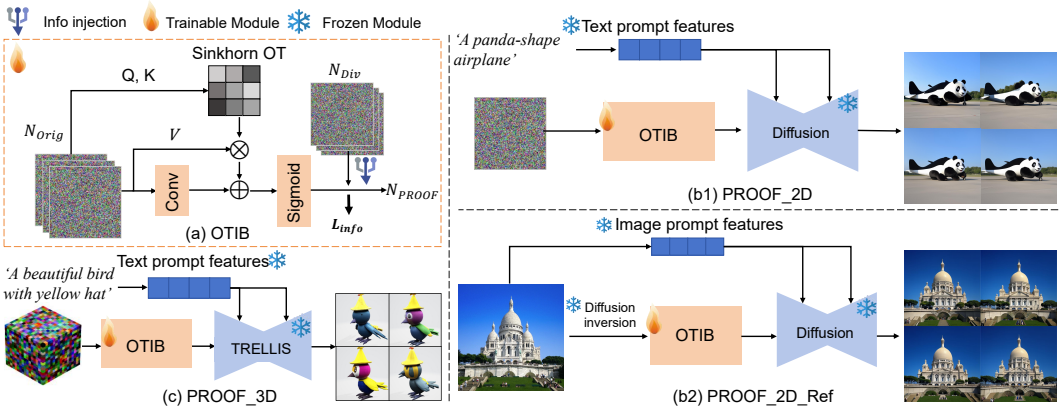

Figure 2: Method overview: as a plug-and-play content controller, PROOF can be employed for 2D/3D generation tasks, different architectures and model checkpoints. OTIB consists of a Sinkhorn Attention module and an information bottleneck module. We obtain N_{PROOF} by information compression of N_{Orig} and information modulation of N_{Div}. More details are introduced in Section 4.

the adaptive interpolation weight based on the closed-form solution of OTIB. We define our noise finetuning as:

$$\theta^* = argmin_\theta \mathbb{E}_{N_{Orig},N_{Div}}[\mathcal{L}_{noise}(PROOF_\theta(N_{Oirg}, N_{Div}), N_{Orig}) + \mathcal{L}_{info}(PROOF_\theta(N_{Orig}), \lambda)], \tag{3}$$

where $PROOF_\theta$ is the generator of PROOF, $\mathcal{L}_{noise}$ aims to provide pixel-level regularization for structure and appearance alignment with N_{Orig}, and $\mathcal{L}_{info}$ explores controlling appropriate neural feature leakage with consideration of contextual preservation, which learns the minimal sufficient representation to avoid the diversity collapse.

3.2 Information Bottleneck Revisiting

Let's denote the original input data, the corresponding label, and compressed information by X, Y, and Z. The information compression principle Tishby & Zaslavsky (2015) is a trade-off between task-related information preservation and the minimal sufficient information compression, by maximizing the sharable information of Z and Y while minimizing that of Z and X:

$$\max_Z \mathbb{I}(Y;Z) - \beta\mathbb{I}(X;Z), \tag{4}$$

where $\mathbb{I}$ means the mutual information and β is a trade-off weight. Let R denote the feature representations of X, and the information loss is formulated as:

$$\mathbb{I}(X;Z) \triangleq \mathbb{I}(R;Z) \triangleq \mathcal{D}_{KL}[p(Z|R)\|q(Z)], \tag{5}$$

where $q(Z)$ with Gaussian distribution is a variational approximation of $p(Z)$ Schulz et al. (2020). $\mathcal{D}_{KL}$ is the KL divergence Csiszár (1975) used to represent the distance between two distributions.

In our problem setting, R is the noise latent N_{Orig} and Z is the compressed latent N_{Out}.

3.3 Optimal Transport Revisiting

We revisit the Optimal Transport that provides a mathematical framework for transporting probability distributions from the source to the target. Given discrete distributions as:

$$\mu = \sum_{i=1}^{M} \mu_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{N} \nu_j \delta_{y_j} \tag{6}$$

where μ, ν are discrete probability measures, $\mu_i \geq 0$, $\nu_j \geq 0$ are probability masses ($\sum_i \mu_i = \sum_j \nu_j = 1$), δ_x denotes the Dirac delta function centered at point x, M and N are the number of support points. The original OT problem finds a transport plan $\mathbf{T}^*$ that minimizes the total

transportation cost, which is computationally intensive. The Sinkhorn algorithm Cuturi (2013); Kim et al. (2024) equips OT with an entropy regularization term:

$$\mathbf{T}^* = \arg\min_{\mathbf{T}\in\Pi(\mu,\nu)} \langle \mathbf{T}, \mathbf{C}\rangle_F - \epsilon H(\mathbf{T}), \tag{7}$$

where $\mathbf{T} \in \mathbb{R}^{M\times N}$ is the transport matrix with $\mathbf{T}_{ij}$ specifying how much mass moves from x_i to y_j, $\mathbf{C} \in \mathbb{R}^{M\times N}$ is the cost matrix where $\mathbf{C}_{ij} = d(x_i, y_j)$, $\Pi(\mu,\nu) = \{\mathbf{T} \geq 0 \mid \mathbf{T1^N} = \mu, \mathbf{T}^\top \mathbf{1^M} = \nu\}$ defines the set of admissible transport plans, $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product. Moreover, $\epsilon > 0$ is the regularization strength, $H(\mathbf{T}) = -\sum_{ij} \mathbf{T}_{ij} \log \mathbf{T}_{ij}$ is the entropy of the transport plan.

4 Approach

In this section, we provide a detailed introduction to our proposed PROOF, including the overall pipeline in Section 4.1, OTIB module architecture in Section 4.2, the closed-form theoretical solution in Section 4.3, along with the training loss in Section 4.4.

4.1 Overall pipeline

As shown in Fig. 2, PROOF can manipulate random noise with text or image conditions in 2D Rombach et al. (2022); Esser et al. (2024) or 3D data Xiang et al. (2025) distribution.

4.1.1 PROOF_2D

As for none-referenced PROOF_2D, given a text prompt denoted by 'S', diverse images can be synthesized based on:

$$I_{PROOF} = G_\phi^{2D*}(PROOF_\theta^{2D}(N_{Oirg}, N_{Div}), \text{'S'}), \tag{8}$$

where G_ϕ^{2D*} is the frozen generator of diffusion model Rombach et al. (2022).

As for referenced PROOF_2D, given a reference image I_{Ref}, we extract the image prompt using IP-Adapter Ye et al. (2023) for consistent appearance transfer. Furthermore, we utilize the diffusion inversion method Mokady et al. (2023) to recover the corresponding contextual latent of I_{Ref}. $PROOF_\theta^{2D}$ perturbs the inversed noise to generate diverse images:

$$I_{PROOF} = G_\phi^{2D*}(PROOF_\theta^{2D}(Inv(I_{Ref}), N_{Div}), I_{Ref}) \tag{9}$$

4.1.2 PROOF_3D

TRELLIS Xiang et al. (2025) compresses the 3D asset representation into a structured 3D latent similar to Latent Diffusion Rombach et al. (2022). It's possible for $PROOF_\theta^{3D}$ to implement the 3D tradeoff considering structural and textural preservation, along with the distribution diversity of 3D models and neural rendering Mildenhall et al. (2021); Kerbl et al. (2023); Lu et al. (2024):

$$M_{PROOF} = G_\phi^{3D*}(PROOF_\theta^{3D}(N_{Oirg}, N_{Div}), \text{'S'}), \tag{10}$$

where G_ϕ^{3D*} is the frozen generator of TRELLIS Xiang et al. (2025).

4.2 OTIB architecture

As mentioned in Section 3, implicit neural compression of information can be formulated as follows:

$$\min_Z \beta \mathbb{I}(N_{Orig}; Z), \tag{11}$$

where $\mathbb{I}$ denotes the mutual information function, Z is the manipulated latent derived from N_{Orig} via Equ. 1. To realize high-fidelity content preservation and generation diversity, we adaptively learn a neural information filter λ of OTIB.

$$\lambda = Sigmoid(Conv(N_{Orig}) + \mathcal{F}_{SA}(N_{Orig})), \tag{12}$$

where $\mathcal{F}_{SA}$ is a Sinkhorn Attention module, as shown in Figure 2. The intent of PROOF is to improve representation diversity while implicitly adhering to the global content attributes of a certain scenario. If λ is 0, the whole manifold will be replaced by N_{Div}, which results in entire structure and appearance leakages. If λ is 1, Z excludes any form of diversity-inducing perturbations, which results in mode collapse. Qualitative analyses are illustrated in Sec. 5.

4.3 CLOSED-FORM SINKHORN-IB SOLUTION

We impose a Sinkhorn Attention module $\mathcal{F}_{SA}$ in a spatial-OT view to improve contextual preservation of PROOF. The Sinkhorn Attention algorithm is as follows:

Algorithm 1 Sinkhorn-Attention Forward Pass

Input: Feature map $X \in \mathbb{R}^{B\times C\times H\times W}$
$Q = \text{Conv\_Nd}(X), K = \text{Conv\_Nd}(X), V = \text{Conv\_Nd}(X)$ ▷ Learnable projections
$A = QK^\top/\sqrt{C}$ ▷ Attention logits
for $k = 1$ to n_{iters} **do**
  $A = A - \text{LogSumExp}(A, \text{dim} = 2)$ ▷ Row normalization
  $A = A - \text{LogSumExp}(A, \text{dim} = 1)$ ▷ Column normalization
end for
$\mathbf{T} = \exp(A)$ ▷ Optimal attention weights
return $\mathbf{T}V$ ▷ Transport applied to values

where $Q, K, V \in \mathbb{R}^{B\times(HW)\times C}$ are Query, Key, Value tensors, respectively. $A \in \mathbb{R}^{B\times(HW)\times(HW)}$ is Attention logits matrix, $\text{LogSumExp}(A)_i = \log\sum_j \exp(A_{ij})$, and $\mathbf{T}$ is Doubly-stochastic attention matrix. Our transport solution is established through:

$$\mathbf{T}_{ij} = \exp(\underbrace{\frac{q_i^\top k_j}{\sqrt{C}}}_{\text{Transport cost}} - \underbrace{\alpha_{OT}^i - \beta_{OT}^j}_{\text{Sinkhorn scalars}}) \tag{13}$$

where α_{OT} and β_{OT} are row and column normalization factors, respectively. The division by $\sqrt{C}$ stabilizes gradient flow. We consider the joint optimization objective of OTIB:

$$\min_\lambda \underbrace{I(R;Z)}_{\text{IB term}} + \gamma \underbrace{< A^*, \mathbf{C} >}_{\text{Sinkhorn OT term}} + \epsilon H(A^*), \tag{14}$$

where $Z = \lambda \odot N_{Orig} + (1-\lambda)\odot N_{Div}$, $A^* = \text{Sinkhorn}(\mathbf{C})$, where $\mathbf{C}_{ij} = \frac{\langle q_i, k_j\rangle}{\sqrt{d}}$, $d = C$.

We assume that: $N_{Orig} \sim \mathcal{N}(0, \sigma_R^2 I)$, $N_{Div} \sim \mathcal{N}(0, \sigma_{N_{Div}}^2 I)$. N_{Orig} and N_{Div} are independent.

Step 1: Information Bottleneck Term Simplification. Under Gaussian assumptions, the mutual information and the gradient calculation are formulated as:

$$I(R;Z) = \frac{1}{2}\log\left(1 + \frac{\lambda^2\sigma_R^2}{(1-\lambda)^2\sigma_{N_{Div}}^2}\right), \frac{\partial I}{\partial\lambda} = \frac{\lambda\sigma_R^2 - (1-\lambda)\sigma_{N_{Div}}^2}{\lambda^2\sigma_R^2 + (1-\lambda)^2\sigma_{N_{Div}}^2} \tag{15}$$

Step 2: Sinkhorn Term Gradient. Using the Envelope Theorem and chain rule:

$$\frac{\partial\mathcal{L}_{OT}}{\partial\lambda} = \left\langle\frac{\partial A^*}{\partial\lambda}, \mathbf{C}\right\rangle + \left\langle A^*, \frac{\partial\mathbf{C}}{\partial\lambda}\right\rangle \approx \left\langle A^*, \frac{\partial\mathbf{C}}{\partial\lambda}\right\rangle, \tag{16}$$

where $A^* = \text{diag}(u)K\text{diag}(v)$ with $K = e^{-\mathbf{C}/\epsilon}$. $\frac{\partial\mathbf{C}_{ij}}{\partial\lambda} = \frac{\partial}{\partial\lambda}(\frac{<q_i,k_j>}{\sqrt{d}}) = \frac{1}{\sqrt{d}}\langle q_i, \frac{\partial k_j}{\partial Z_j}\cdot\frac{\partial Z_j}{\partial\lambda}\rangle \approx \frac{1}{\sqrt{d}}\langle q_i, \frac{\partial k_j}{\partial N_{Orig}^j}\cdot\frac{\partial Z_j}{\partial\lambda}\rangle = \frac{1}{\sqrt{d}}\langle q_i, W_K(N_{Orig}^j - N_{Div}^j)\rangle$.

Step 3: First-Order Optimality Condition Setting. The total gradient to zero:

$$\frac{\lambda\sigma_R^2 - (1-\lambda)\sigma_{N_{Div}}^2}{\lambda^2\sigma_R^2 + (1-\lambda)^2\sigma_{N_{Div}}^2} + \gamma\mathbb{E}_{A^*}\left[\frac{\partial\mathbf{C}_{ij}}{\partial\lambda}\right] = 0 \tag{17}$$

Step 4: Closed-Form OTIB Solution. The optimal weights take the form (More details are in Appendix A):

$$\lambda^* = \sigma\left(\frac{1}{\eta}\left(\gamma\cdot\text{Align}(N_{Orig}, N_{Div}) - \frac{\sigma_{N_{Div}}^2}{\sigma_R^2}\right)\right), \tag{18}$$

where $\text{Align}(\cdot) = \mathbb{E}_{A^*}\left[\frac{\partial\mathbf{C}_{ij}}{\partial\lambda}\right]$, $\sigma(\cdot)$ is the sigmoid function, and η is a hyperparameter. The closed-form solution is aligned with Equ. 12, where $Conv$ approximates σ^2 ratio, and $\mathcal{F}_{SA}$ approximates Align term.

Figure 3: Qualitative results of PROOF_2D, ControlNet + IP Adapter Zhang et al. (2023); Ye et al. (2023), FreeControl Mo et al. (2024), Ctrl-X Lin et al. (2024), Uni-ControlNet Zhao et al. (2023), T2I-Adapter + IP Adapter Mou et al. (2024); Ye et al. (2023), and Reimagine AI (2023). Zoom in for better observation. PROOF realizes more controllable image variations with high-fidelity content.

4.4 Training Loss

Training losses contain pixel-level reconstruction loss and manifold-level information compression loss. As for noise consistency loss, the pixel-level supervision for N_{PROOF} is MSE loss that demonstrates a powerful content preservation function Rombach et al. (2022); Ruiz et al. (2023):

$$\mathcal{L}_{noise} = ||N_{PROOF} - N_{Orig}||_2^2. \tag{19}$$

For Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(0, 1)$, KL divergence is formulated as:

$$\mathcal{D}_{KL}[N(\mu, \sigma^2)\|N(0, 1)] = -\frac{1}{2}[log(\sigma)^2 - (\sigma)^2 - (\mu)^2 + 1]. \tag{20}$$

Our framework eliminates the need for feature mean/variance pre-calculation by leveraging the predefined properties of Gaussian noise (μ_G=0, σ_G=1). As for our case mentioned in Equ. 5, the distribution of $p(Z|R)$ is accessed as $\mathcal{N}[\lambda R, (1-\lambda)^2]$ according to Equ. 1. We normalize $p(Z|R)$ along with $q(Z)$ using μ_G and σ_G, then the information compression metric of PROOF is:

$$\mathcal{L}_{info} = \mathbb{I}(Z; R) = KL[p(Z|R)\|q(Z)] = -\frac{1}{2}[log(1-\lambda)^2 - (1-\lambda)^2 - (\lambda R)^2 + 1], \tag{21}$$

Finally, the total loss of PROOF is formulated as:

$$\mathcal{L}_{PROOF} = \beta\mathcal{L}_{info} + \mathcal{L}_{noise}, \tag{22}$$

where β is the content-diversity tradeoff weight (Fig. 8a). Higher β usually intentionally relaxes contextual constraints but boosts the diversity (Fig. 10, Fig. 12).

Table 1: PROOF outperforms other SOTA methods in structure and appearance alignments and robustness, measured by DINO ViT self-similarity and DINO-I. We report the inference time of PROOF_2D and PROOF_2D_Ref, where diffusion inversion Mokady et al. (2023) is time-consuming. We assess image quality (PickScore, HPSv2, AES) and diversity (LPIPS, L1).

Methods	Training	Inference time (s)	self-sim ↓	DINO-I ↑	PickScore↑	HPSv2↑	AES↑	L1	LPIPS
Uni-ControlNet Zhao et al. (2023)	✓	10.6	0.045	0.555	6.49	25.33	6.26	56.41	0.5500
ControlNet + IP Adapter Zhang et al. (2023)	✓	8.1	0.068	0.656	15.08	25.02	6.29	46.06	0.4334
T2I-Adapter + IP Adapter Mou et al. (2024)	✓	4.2	0.055	0.603	12.39	25.45	6.28	50.45	0.4436
Ctrl-X Lin et al. (2024)	✗	14.9	0.057	0.686	11.65	24.63	6.27	37.07	0.4812
FreeControl Mo et al. (2024)	✗	21.5	0.058	0.572	18.13	26.13	6.19	**85.45**	**0.636**
Reimagine AI (2023)	✓	10.1	0.073	0.753	15.14	25.27	6.34	64.12	0.6192
PROOF (ours)	✓	7.3 / 27.2	**0.038**	**0.841**	16.61	25.67	6.29	41.58	0.4342

5 EXPERIMENTS

Comprehensive qualitative and quantitative evaluations validate PROOF's dual capability in maintaining content fidelity while enhancing generation diversity for digital asset creation. Additional results, e.g., golden noise Zhou et al. (2025) finetune (Fig. 11), are shown in Appendix D.

Training Protocol. We train our PROOF on Gaussian noise tensors with corresponding dimension shape of different architectures, e.g., $4 * 64 * 64$ Rombach et al. (2022), $16 * 128 * 128$ Esser et al. (2024), $8 * 16 * 16 * 16$ Xiang et al. (2025). N_{Orig} and N_{Div} are random noises in each training step. As for PROOF_3D, we utilize 3D convolutions for SA and IB modules. We train PROOF for 20k iterations with one NVIDIA RTX 4090 GPU. The training batch size is set to 1. During training, we employ Adam Kingma & Ba (2014) with $2 * 10^{-3}$ learning rate. We set $\beta = 0.01$ for mild diversity (Figure 3a), $\beta = 0.1$ for substantial diversity (Fig. 3b, Fig. 10), and $\beta = 1$ for diversity with reference constraints (Fig. 3c).

Baselines. There are several state-of-the-art controllable synthesis methods based on diffusion models. ControlNet Zhang et al. (2023) and T2I-Adapter Mou et al. (2024) align diffusion priors to the external control structures. We further apply IP-Adapter Ye et al. (2023) to them for better textural transfer. These methods present low topological flexibility with restriction by the explicit structure alignment, and limited textural fidelity with global appearance control. FreeControl Mo et al. (2024) has large-scale content variance due to imprecise structure and appearance representations (col 4 in Fig. 3). Ctrl-X Lin et al. (2024) provides too-strict structure and appearance alignments, and there are texture distortions. Uni-ControlNet Zhao et al. (2023) also suffers from the global appearance representation (col 6 in Fig. 3). Reimagine AI (2023) produces uncontrollable content layout, despite high image quality and diversity (col 8 in Fig. 3). We evaluate all methods on SDXL v1.0 Podell et al. (2024) when workable and on their pre-configured base models otherwise.

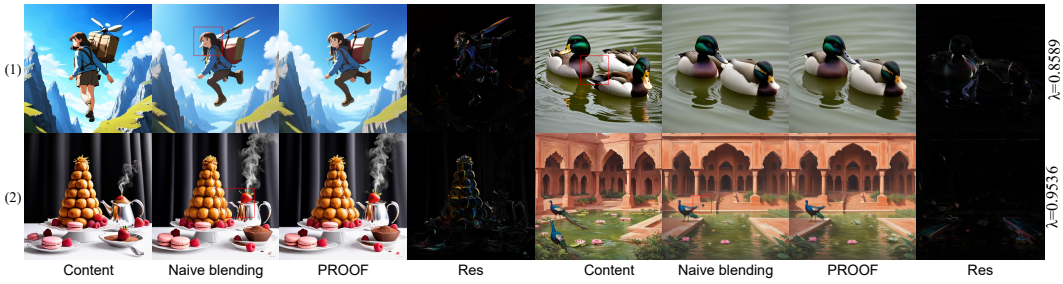

Figure 4: PROOF w/ $\beta = 0.1$ (Row 1) and $\beta = 0.05$ (Row 2) are corresponding with naive blending w/ $\lambda = 0.8589$ and $\lambda = 0.9536$, based on the mean value across the channel and spatial dimensions of PROOF's neural λ. PROOF preserves fine-grained structure and appearance features.

Evaluation metrics. Tab. 1 shows a quantitative comparison of natural images of datasets Lin et al. (2024). The content alignment metrics include DINO ViT self-similarity Tumanyan et al. (2022), DINO-I Ruiz et al. (2023) (details are explained in Appendix B). Note that PROOF shows consistent superiority on self-sim and DINO-I scores. As for image quality, we utilize PickScore Kirstain et al. (2023), HPSv2 Wu et al. (2023), and Aesthetic Score (AES) Schuhmann (2023). We assess the diversity via LPIPS Zhang et al. (2018) between the source image and the generated image. Meanwhile, the subjective metrics consist of quality, fidelity, and diversity without compromising fidelity. PROOF achieves comparable user preference (Tab. 3).

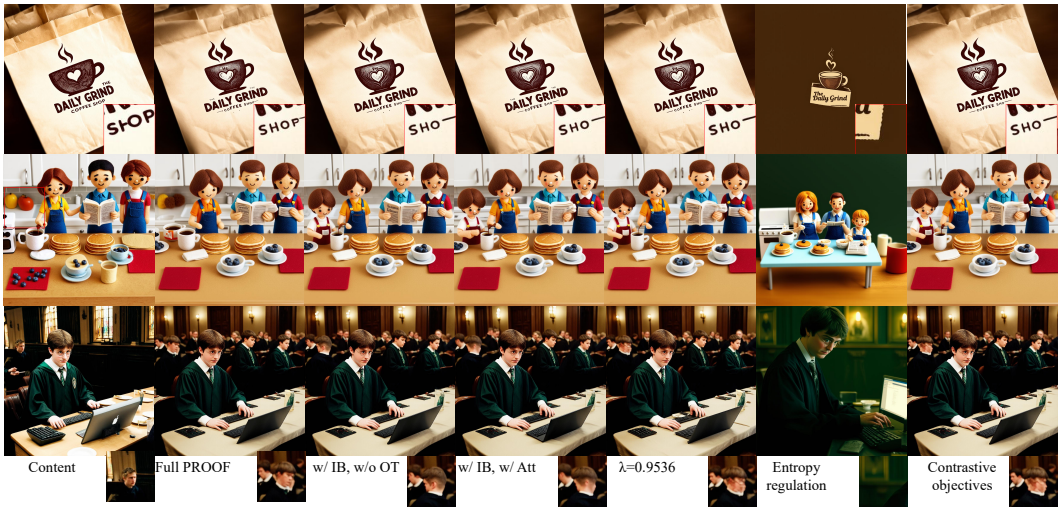

Figure 5: PROOF sufficiently preserves the global structure and appearance based on OTIB, e.g., the word 'SHOP', no-man's land on the left of Row 2, and the far-distance face of Row 3, while other variants show lower content fidelity. More results are illustrated in Fig. 6.

Table 2: Quantitative validation for PROOF_2D generation with random noise initialization on the dataset Lin et al. (2024). PROOF outperforms other ablation configurations and diversity-inducing methods in structure and appearance alignments. w is the loss weight aligned with λ.

Configuration	self-sim ↓	DINO-I ↑	PickScore↑	HPSv2↑	AES↑	LPIPS
w/o IB ≜ Content	0	0.9999	20.95	34.64	5.50	0
Full PROOF β=0.05	**0.0314**	**0.9026**	18.43	33.38	5.43	0.4551
w/ IB, w/o OT	0.0333	0.8974	18.20	33.35	5.36	0.4562
w/ IB, w/ AttentionBlock	0.0331	0.8968	**18.81**	**33.80**	5.37	0.4590
Naive interpolation λ=0.9536	0.0423	0.8650	14.85	33.15	5.39	0.4549
Entropy regularization w=0.45	0.0947	0.6299	12.30	31.20	**5.76**	**0.6790**
Contrast objective w=0.085	0.0320	0.9012	17.38	33.45	5.41	0.4565

Qualitative results. PROOF only learn noise representation supervised by itself based on OTIB. Visually comparable results demonstrate that our implicit PROOF is a better workbench for highly correlated asset editing (Fig. 3, Fig. 11, more examples in Appendix D).

Ablation Study Fig. 4 and Fig. 5 demonstrate substantial benefits of PROOF over other alternatives, e.g., naive interpolation, entropy regularization, contrastive objective (loss details in Appendix C). PROOF exhibits best structure and appearance fidelity, and comparative perceptual quality in Tab. 2. Without the Information Bottleneck, the model will suffer from mode collapse due to a lack of mode diversity. Moreover, as shown in Fig. 8 (a), the PROOF variants without Sinkhorn Attention fail to capture local structure and appearance patterns (red boxes in col 3&4). The context-diversity tradeoff weight β controls the structure and appearance leakage in an adaptive way (Fig. 12).

Limitations Large-scale compression with small weight λ may result in background leakage to a certain extent, as shown in Figure 8 (b). Nevertheless, the pose and identity of the original content are preserved well.

6 Conclusion

Our proposed PROOF conducts perturbation-robust asset creation with a trade-off of fidelity and diversity. We derive the closed-form solution of the optimal transported information bottleneck and design an efficient and effective OTIB module. Compared with explicit content alignment methods, along with other diversity-inducing alternatives, PROOF preserves topology and texture better. Comprehensive experimental analyses demonstrate that PROOF is promising to be the first plug-and-play implicit controller for pre-trained conditional 2D/3D generation models with remarkable context consistency and controllable diversity.

Broader impacts. Our method provides a robust editor for both images and 3D models. While its primary advantage lies in assisting designers, animators, and 3D modelers in asset creation, the potential for malicious manipulation of visual assets necessitates mandatory watermarking in practical applications.

REFERENCES

Stability AI. Clipdrop reimagine. Web Service, 2023. URL https://clipdrop.co/reimagine. AI-powered image regeneration tool.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *ICLR*, 2017.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–12, 2023a.

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18370–18380, 2023b.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3404–3413, 2021.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation. In *European Conference on Computer Vision*, pp. 200–217. Springer, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22511–22521, 2023.

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. In *Advances in Neural Information Processing Systems*, 2024.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, and et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *CVPR*, 2024.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7964–7973, 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.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6527–6536, 2024.

Christoph Schuhmann. Improved aesthetic predictor. 2023.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *ICLR*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10748–10757, 2022.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6232–6242, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

Xiaofeng Yang, Chen Cheng, Xulei Yang, Fayao Liu, and Guosheng Lin. Text-to-image rectified flow as plug-and-play priors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SzPZK856iI.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–14255, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 2023.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22490–22499, 2023.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6818–6828, 2024.

Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *International Conference on Computer Vision*, 2025.

A Appendix A: Detailed Derivation of Closed-form Solution

1. **Initial optimality condition:** Based on step 3 of Section 4.3, the optimization problem of OTIB gives us:

$$\frac{\lambda\sigma_R^2 - (1-\lambda)\sigma_{N_{Div}}^2}{\lambda^2\sigma_R^2 + (1-\lambda)^2\sigma_{N_{Div}}^2} + \gamma\text{Align} = 0 \tag{23}$$

This equation balances the information bottleneck term with the optimal transport term.

2. **Rearrange optimality condition:** We multiply both sides by the denominator to eliminate the fraction:

$$\lambda\sigma_R^2 - (1-\lambda)\sigma_{N_{Div}}^2 = -\gamma\text{Align}(\lambda^2\sigma_R^2 + (1-\lambda)^2\sigma_{N_{Div}}^2) \tag{24}$$

This form removes the denominator but introduces quadratic terms in λ.

3. **Auxiliary function definition:** To analyze this equation, we define:

$$f(\lambda) = \lambda\sigma_R^2 - (1-\lambda)\sigma_{N_{Div}}^2 + \gamma\text{Align}\left[\lambda^2\sigma_R^2 + (1-\lambda)^2\sigma_{N_{Div}}^2\right] \tag{25}$$

The optimal solution occurs when $f(\lambda) = 0$.

4. **Taylor expansion at $\lambda = 0.5$:** We linearize around $\lambda = 0.5$ because:

- It's the midpoint of possible λ values
- The function is most linear in this region
- Higher-order terms are minimized here

4.1. Function value at $\lambda = 0.5$:

$$f(0.5) = 0.5(\sigma_R^2 - \sigma_{N_{Div}}^2) + 0.25\gamma\text{Align}(\sigma_R^2 + \sigma_{N_{Div}}^2) \tag{26}$$

This combines the linear difference and quadratic alignment terms.

4.2. First derivative:

$$f'(\lambda) = \sigma_R^2 + \sigma_{N_{Div}}^2 + \gamma\text{Align}\left[2\lambda\sigma_R^2 - 2(1-\lambda)\sigma_{N_{Div}}^2\right] \tag{27}$$

$$f'(0.5) = \sigma_R^2 + \sigma_{N_{Div}}^2 + \gamma\text{Align}(\sigma_R^2 - \sigma_{N_{Div}}^2) \tag{28}$$

The derivative shows how sensitive the function is to λ changes.

4.3. Linear approximation solution: Using Taylor expansion:

$$\lambda \approx 0.5 - \frac{f(0.5)}{f'(0.5)} = 0.5 - \frac{0.5(\sigma_R^2 - \sigma_{N_{Div}}^2) + 0.25\gamma\text{Align}(\sigma_R^2 + \sigma_{N_{Div}}^2)}{\sigma_R^2 + \sigma_{N_{Div}}^2 + \gamma\text{Align}(\sigma_R^2 - \sigma_{N_{Div}}^2)} \tag{29}$$

This gives us a first-order approximation of the optimal λ.

4.4. Simplified linear expression: When γAlign is relatively small compared to the variance terms:

$$\lambda \approx \underbrace{\frac{\sigma_{N_{Div}}^2}{\sigma_R^2 + \sigma_{N_{Div}}^2}}_{C} + \underbrace{0.25\gamma}_{K}\cdot\text{Align}, \tag{30}$$

where C represents the baseline compression ratio, and K determines how strongly alignment affects the result.

5. **Identify limitations of the linear form:** The linear expression has two critical flaws:

- When Align is too large, λ may exceed 1
- When Align is too small, λ may be less than 0

However, λ must be a weight coefficient strictly between 0 and 1. Therefore, we need a function that constrains the output to (0,1) while preserving the positive correlation between λ and Align.

6. **Choose sigmoid function for constraint:** The sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ is ideal because:

- Its output is strictly bounded between (0,1)
- It's monotonically increasing, preserving the positive correlation
- It provides smooth, differentiable transitions

7. **Match the baseline value at Align $= 0$:** When there's no alignment (Align $= 0$), the linear expression gives $\lambda \approx C$. To maintain consistency:

$$\sigma(x_0) = C \quad \text{where } x_0 = \sigma^{-1}(C) \tag{31}$$

Using the inverse of the sigmoid function (logit function) $\sigma^{-1}(y) = \ln\left(\frac{y}{1-y}\right)$, we get:

$$\sigma^{-1}(C) = \ln\left(\frac{\sigma^2_{N_{Div}}}{\sigma^2_R}\right) \tag{32}$$

This ensures the sigmoid preserves the baseline behavior when Align $= 0$.

8. **Final sigmoid parameterization:** To maintain the positive correlation while adding flexibility, we introduce:

$$x = \frac{1}{\eta}\left(\gamma \cdot \text{Align} - \frac{\sigma^2_{N_{Div}}}{\sigma^2_R}\right), \tag{33}$$

where $\eta > 0$ controls the steepness of the transition. The final solution becomes:

$$\lambda^* = \sigma\left(\frac{1}{\eta}\left(\gamma \cdot \text{Align} - \frac{\sigma^2_{N_{Div}}}{\sigma^2_R}\right)\right) \tag{34}$$

This closed-form solution is presented in step 4 of Section 4.3, and satisfies all our requirements:

- Strictly bounded output (0,1)
- Preserves positive correlation
- Matches baseline when Align $= 0$
- Allows tuning via η and γ

B EVALUATION METRIC

Below is the explicit explanation of how DINO ViT self-similarity and DINO-I are calculated:

1. The structural consistency is quantified as:

$$\text{Self-sim} = \frac{1}{N}\sum_{i=1}^{N} \|\phi_{\text{DINO}}(I_{\text{Ref}})_i - \phi_{\text{DINO}}(I_{\text{Out}})_i\|_2^2, \tag{35}$$

where ϕ_{DINO}: DINO-ViT base model (patch size=8) feature extractor, I_{Ref}: Reference input image, I_{Out}: Generated output image, N: Number of feature vectors (layer_num=11).

2. The appearance similarity is computed as:

$$\text{DINO-I} = \frac{\mathbf{v}_{\text{ref}} \cdot \mathbf{v}_{\text{out}}}{\|\mathbf{v}_{\text{ref}}\|_2 \|\mathbf{v}_{\text{out}}\|_2}, \tag{36}$$

where $\mathbf{v}_{\text{ref}} = \phi^{[\text{CLS}]}_{\text{DINO}}(I_{\text{ref}})$: DINO-ViT [CLS] token embedding of reference image, $\mathbf{v}_{\text{out}} = \phi^{[\text{CLS}]}_{\text{DINO}}(I_{\text{out}})$: DINO-ViT [CLS] token embedding of output image, ϕ_{DINO}: DINO-ViT small model (patch size=16) feature extractor, '$\cdot$' denotes dot product.

C DIVERSITY-BOOSTING METHODS

In these diversity-inducing settings, we maintain the $\mathcal{L}_{\text{noise}}$ of Equ. 22 to conduct global content preservation.

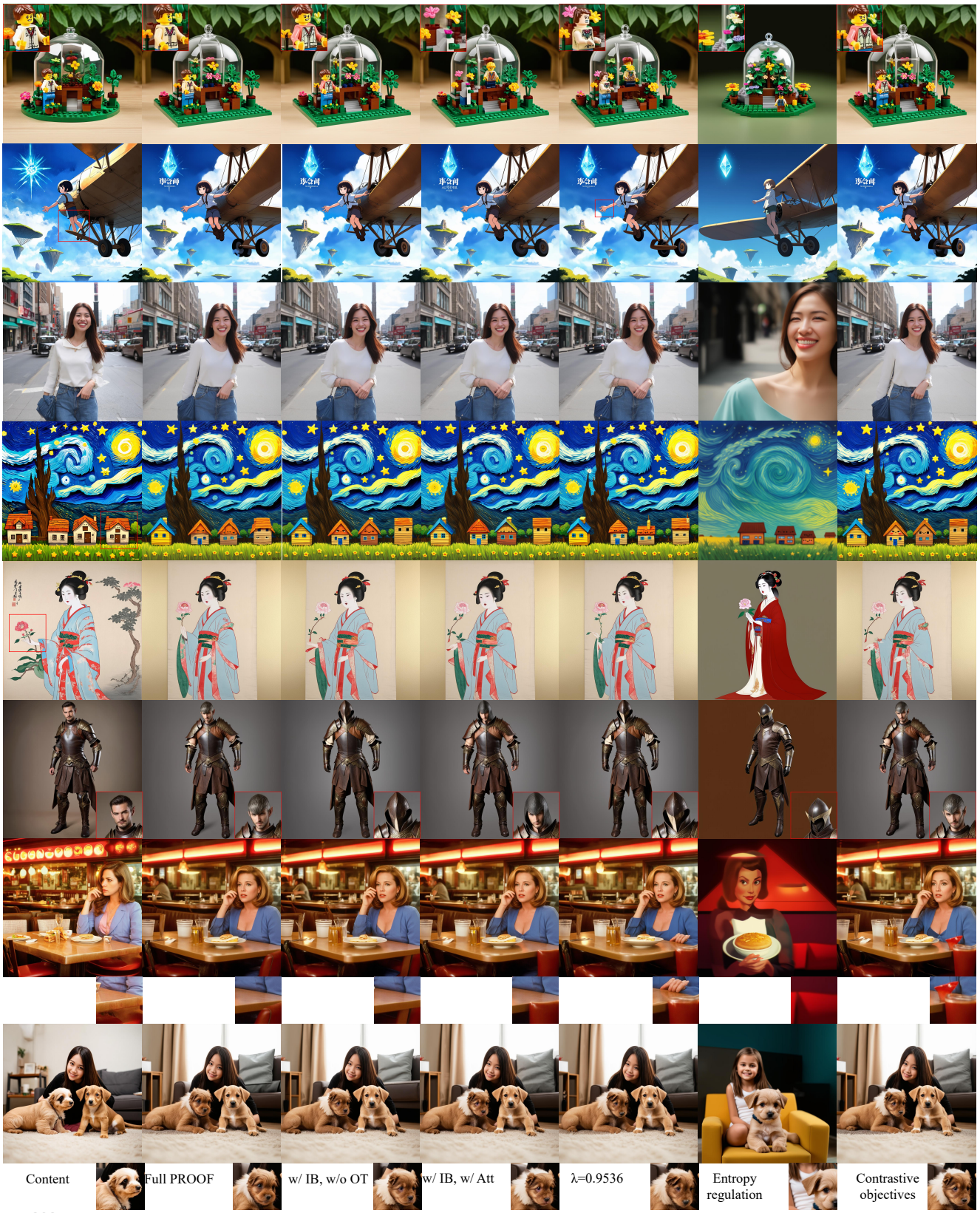

Figure 6: PROOF sufficiently preserves the global structure and appearance based on OTIB, while other variants show lower content fidelity. Zoom in for better observation.

C.1 CONTRASTIVE OBJECTIVE

Given flat_Z = flatten(Z) $\in \mathbb{R}^{N\times d}$, flat_h = flatten($R$) $\in \mathbb{R}^{N\times d}$, flat_l = flatten($N_{Div}$) $\in \mathbb{R}^{N\times d}$, we calculate the cross-modal cosine similarity explicitly as sim_zh = $\cos$(flat_Z, flat_h), sim_zl = $\cos$(flat_Z, flat_l), sim_hl = $\cos$(flat_h, flat_l). Then the contrastive objective loss is indicated as:

$$\mathcal{L}_{\text{contrast}} = w * (\text{MSE}(\text{sim\_zh}, \text{sim\_hl}) + \text{MSE}(\text{sim\_zl}, 1 - \text{sim\_hl})), \tag{37}$$

where w is the loss weight.

Note that the contrastive objective has some limitations as follows:

1. Exhibits significantly weaker robustness compared to PROOF under strong perturbations.

2. Fails to perform effective representation learning at the manifold distribution level.

3. Demonstrates notable scalability constraints in real-world applications.

4. Generates structural and appearance artifacts (Fig. 5, Fig. 6).

C.2 ENTROPY REGULARIZATION

1. Input tensor flattening (flatten the i-th sample of Z from multi-dimensional to a vector)

$$Z_i^{\text{flat}} = \text{view}(Z_i, -1) \tag{38}$$

(i.e., flattened into a $1 \times \text{D}$ vector, where D is the flattened dimension)

2. Softmax probability calculation of the j-th class for the i-th sample (compute class probabilities for each flattened sample)

$$p_{i,j} = \text{Softmax}(Z_i^{\text{flat}})_j = \frac{\exp((Z_i^{\text{flat}})_j)}{\sum_{k=1}^{D} \exp((Z_i^{\text{flat}})_k)} \tag{39}$$

3. Entropy calculation for a single sample (ϵ is added to avoid meaningless logarithm)

$$H(Z_i) = -\sum_{j=1}^{D} p_{i,j} \cdot \log(p_{i,j} + \epsilon) \tag{40}$$

Therefore, the function definition of Entropy regularization is:

$$\text{Entropy}(Z, w, \epsilon = 10^{-8}) = -w \cdot \frac{1}{N} \sum_{i=1}^{N} H(Z_i) \tag{41}$$

Note that entropy regularization has some limitations as follows (Fig. 5, Fig. 6):

1. Complete loss of background information.

2. Fails to ensure a minimal sufficient representation learning.

3. Poor robustness in structure and appearance preservation.

D ADDITIONAL RESULTS

In this section, we provide additional qualitative results of 2D (Figure 13, 15, 16, 17) or 3D asset (Figure 14) creation based on PROOF. Figure 12 indicates the workable function of OTIB to conduct controllable diversity implicitly. Note that the detailed differences for small β are not obvious. Please zoom in sufficiently and observe patiently.

Model select As for PROOF_2D_Ref, we use Realistic_Vision_ V4.0_noVAE for diffusion inversion and denoising, with ip-adapter-plus_sd15 for appearance transfer. The VAE module is from stabilityai-stable-diffusion-2-1-base. In Figure 17, iRFDS+Instantx uses the checkpoint of InstantX-SD3.5-Large-IP-Adapter. In Figure 7, images of PROOF_2D are synthesized based on the checkpoint of Stable Diffusion v2-1_512-ema-pruned. In Figure 11, we use stabilityai-stable-diffusion-xl-base-1.0.

Note that because of the strong constraints from the image condition of TRELLIS Xiang et al. (2025), there is little diverse space for direct PROOF_3D_Img. Therefore, we first synthesize the image variants based on PROOF_2D and then conduct 3D modeling based on the trellis-image-large model. Text-based PROOF_3D uses the trellis-text-xlarge model, as shown in Figure 14.

User Study We invite 100 domain experts to conduct the user study. First, we briefly explain the highly correlated asset creation task. We suggest that users carefully observe the original content and generated image variants obtained by 6 state-of-the-art methods and our proposed PROOF. Each observed algorithm has 20 samples. These observers need to select the better image variant set from 3 aspects: (a) overall quality, (b) overall fidelity considering structure and appearance, (c) controllable diversity subject to the fidelity. The interface of our user study is shown in Figure 18.

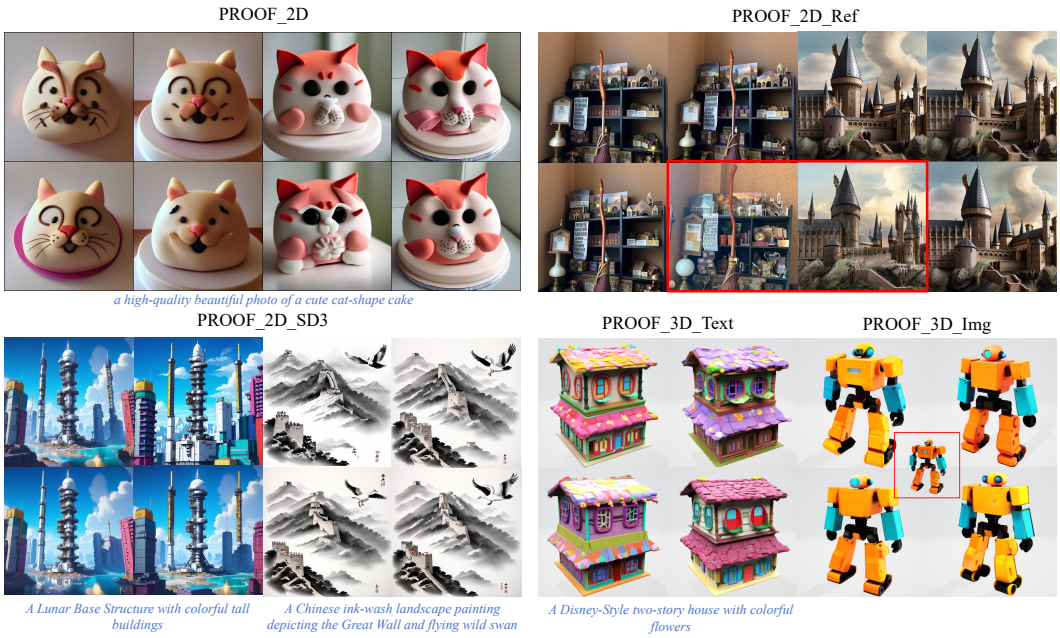

Figure 7: Our proposed PROOF is an effective learning framework to synthesize highly correlated assets where variants exhibit consistent structure and appearance. Test-time PROOF facilitates high-quality 2D assets Esser et al. (2024) and 3D assets Xiang et al. (2025) with high contextual fidelity and controllable diversity, under any text or image condition (red boxes).

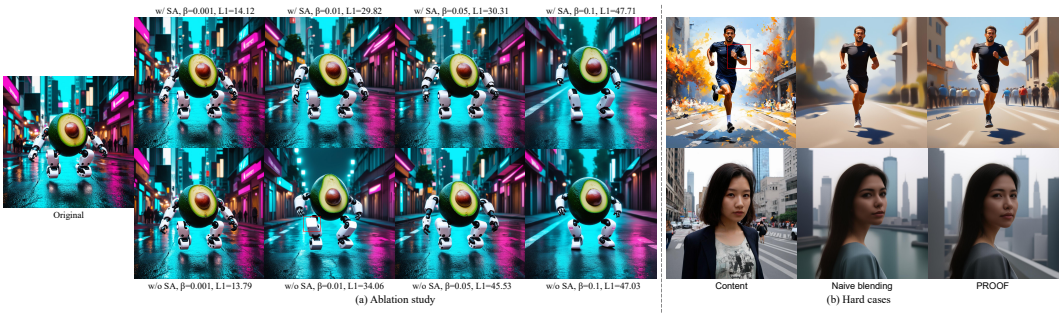

Figure 8: (a) PROOF variants show that methods w/ SA preserve better appearance statistics than those w/o SA. Higher β usually intentionally relaxes contextual constraints but boosts the diversity. (b) The background lacks abundant details for large-scale information compression (e.g., λ=0.8), while the human identity and pose are maintained well.

Comparision with DSG While achieving similar editing effects to DSG Epstein et al. (2023) in Figure 10, our *PROOF* doesn't require any explicit guidance, e.g., position, size, shape.

Comparison with Golden Noise Task Differentiation of Golden Noise Zhou et al. (2025) and PROOF: Golden Noise focuses on text-embedding alignment in noise space and embeds semantic information into noise for semantic fidelity. PROOF targets content-aligned variation generation by

Table 3: PROOF exhibits competitive human preference percentages. Preference consistency is 87%, std. deviation is ±3.0%, and the p-value of Wilcoxon is 0.016, which demonstrates the results are statistically significant.

Methods	Quality ↑	Fidelity ↑	Diversity (subject to Fidelity)↑
Uni-ControlNet Zhao et al. (2023)	78%	71%	75%
ControlNet + IP Adapter Zhang et al. (2023); Ye et al. (2023)	57%	64%	75%
T2I-Adapter + IP Adapter Mou et al. (2024); Ye et al. (2023)	67%	65%	78%
Ctrl-X Lin et al. (2024)	81%	**90%**	74%
FreeControl Mo et al. (2024)	76%	51%	66%
Reimagine AI (2023)	**91%**	37%	51%
PROOF (ours)	89%	89%	**90%**

Table 4: GENEVAL Ghosh et al. (2023) scores of different models. Robust PROOF preserves the semantic content well and exhibits higher text-image correctness v.s. naive noise interpolation.

Model	Overall	Single object	Two object	Counting	Colors	Position	Color attribution
CLIP retrieval	0.35	0.89	0.22	0.37	0.62	0.03	0
minDALL-E	0.23	0.73	0.11	0.12	0.37	0.02	0.01
Stable Diffusion v1.5	0.43	0.97	0.38	0.35	0.76	0.04	0.06
Stable Diffusion v2.1	0.5	0.98	0.51	0.44	0.85	0.07	0.17
Stable Diffusion XL	0.55	0.98	0.74	0.39	0.85	0.15	0.23
IF-XL	0.61	0.97	0.74	0.66	0.81	0.13	0.35
Naive λ=0.85	0.61	0.96	0.67	0.54	0.80	0.23	0.46
PROOF β=0.1	0.70	0.98	0.80	0.65	0.91	0.32	0.55
PROOF β=0.01	0.72	0.98	0.83	0.67	0.92	0.35	0.57

Figure 9: More comparative results of naive blending and content-robust PROOF.

Figure 10: Feature workbench provided by DSG Epstein et al. (2023) is fine-grained but cumbersome. Our PROOF gives another efficient and diverse workbench to change the properties of objects.

modifying local structure and appearance distributions for contextual fidelity with diversity. We provide some comparative results in Fig. 11, which demonstrates that PROOF is powerful to synthesize high-fidelity and high-quality assets.

Specifically, given standard noise as N_{Orig}, we obtain golden noise $N_{Gold} = NPNet(N_{Orig}, c)$. Moreover, standard PROOF and golden PROOF are implemented based on N_{Orig} and N_{Gold}, where the same N_{Div} is adaptively interpolated via OTIB. Note that both NPNet and PROOF leverage SDXL as the pretrained base model.

Figure 11: Standard PROOF and Golden PROOF are based on the standard noise and golden noise, respectively. PROOF seems to produce more high-fidelity golden noise (col 3), and Zhou et al. (2025) exhibits low perturbation robustness (col 4).

Figure 12: PROOF effectively controls the structure and appearance of the content. Smaller tradeoff weight β puts content on a slight adjustment workbench, while larger β changes the content more obviously, but maintains the scene layout.

Figure 13: Image variants of the teaser figure 7 under magnified observation.

Figure 14: More qualitative results of PROOF_3D based on TRELLIS Xiang et al. (2025).

A hidden bedroom, suspended among ancient trees, where moss carpets the floor and fireflies glow instead of lamps

Birds eye view of inupiat whale hunters launching umiak boats on arctic ice

This dreamlike digital art captures a vibrant, kaleidoscopic bird in a lush rainforest

A palace blossoming like a sacred lotus, its petals carved in marble, glows under the moonlight

Figure 15: Additional visual results of PROOF_2D based on SD3 Esser et al. (2024).

A delicate blue-and-white porcelain plate, its surface painted with an intricate castle that seems to float between clouds and waves, where the kiln's fire has turned cobalt into dream

A futuristic robot and an ancient hourglass, contrasting technology and the passage of time

Figure 16: Additional visual results of PROOF_2D based on SD3 Esser et al. (2024).

Figure 17: Qualitative results of PROOF_2D_Ref, ControlNet Zhang et al. (2023); Ye et al. (2023), FreeControl Mo et al. (2024), Ctrl-X Lin et al. (2024), Uni-ControlNet Zhao et al. (2023), T2I-Adapter Mou et al. (2024); Ye et al. (2023), Reimagine AI (2023) and iRFDS Yang et al. (2025) on the wild images.

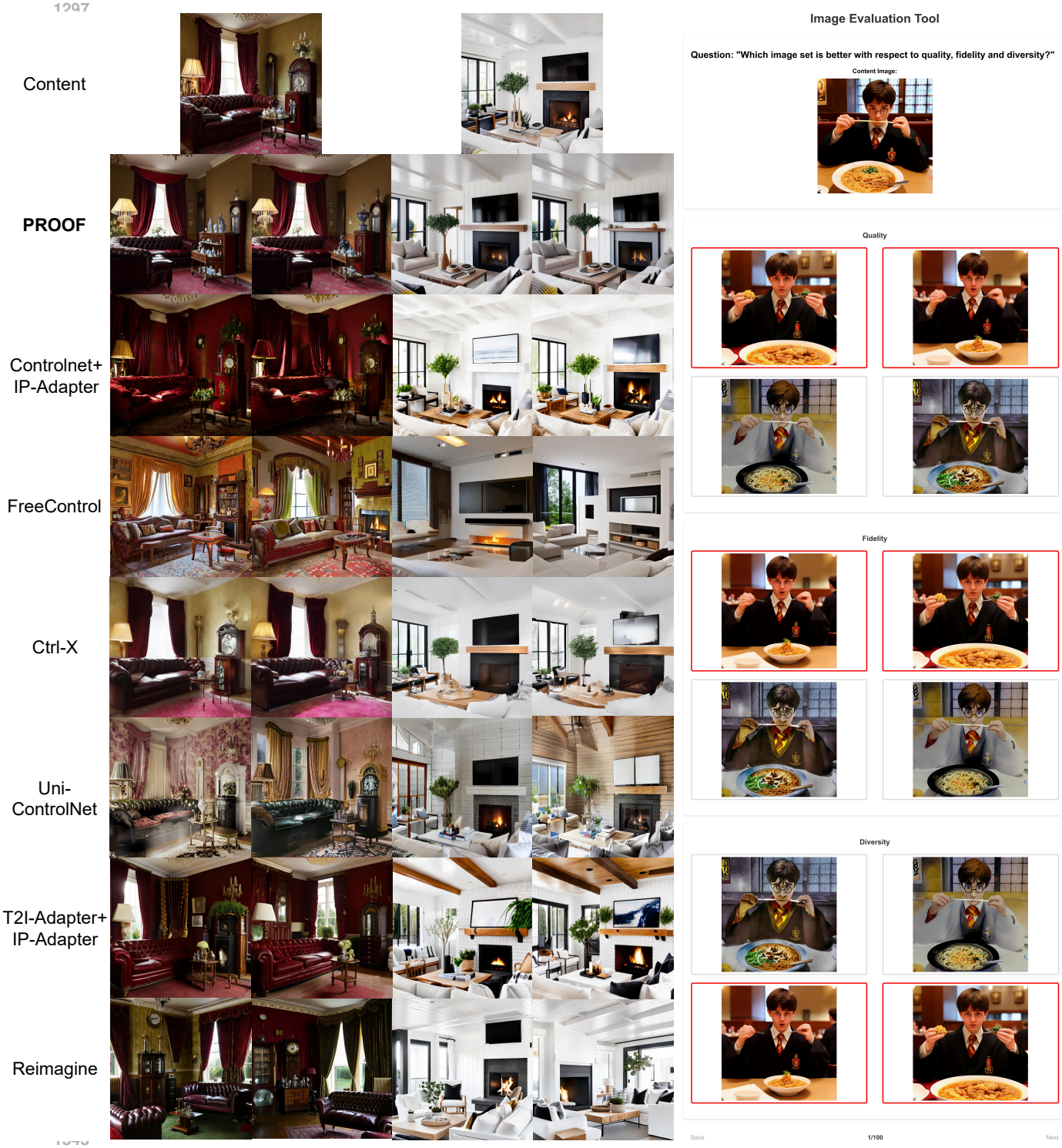

Figure 18: (a) Additional qualitative results of PROOF_2D_Ref, ControlNet Zhang et al. (2023); Ye et al. (2023), FreeControl Mo et al. (2024), Ctrl-X Lin et al. (2024), Uni-ControlNet Zhao et al. (2023), T2I-Adapter Mou et al. (2024); Ye et al. (2023), and Reimagine AI (2023). (b) The interface of our user study.