

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 PROOF: PERTURBATION-ROBUST NOISE FINETUNE VIA OPTIMAL TRANSPORT INFORMATION BOTTLE- NECK 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-**R**Obust **n**Oise **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. (2023a); 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. Therefore, it is supposed to distribute attention in a coordinated manner to eliminate excessive local atten-



Figure 1: Content-diversity tradeoff: given a noise latent of a content, naive noise blending with interpolation weight  $\lambda$  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.

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  $\lambda$  for original noise and  $(1-\lambda)$  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 compose different adapters independently Mou et al. (2024); Zhao et al. (2023), or conduct adaptively feature modulations Zhang et al. (2023a); 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.

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

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

125 **Diffusion seed.** Previous diffusion control methods only treat Gaussian noise as a flexible random  
 126 generation seed Zhang et al. (2023a); Zhao et al. (2023); Ye et al. (2023); Zheng et al. (2023); Wang  
 127 et al. (2024); Zhou et al. (2024); Ruiz et al. (2023); Gal et al. (2023); Avrahami et al. (2023a); Po et al.  
 128 (2024); Ruiz et al. (2024). They constrain the pre-trained diffusion model using external structure or  
 129 textural data. Nevertheless, some diffusion inversion works Yang et al. (2025); Song et al. (2020);  
 130 Mokady et al. (2023) show high-fidelity image reconstruction and editing. Seed searching Ma et al.  
 131 (2025) is beyond the denoising steps for high-quality image generation. These methods establish  
 132 the critical role of noise representation, which is demonstrated by Figure 1 as well. Therefore, we  
 133 explore the implicit structure and appearance alignment based on noise in this paper.

134 **Information bottleneck.** Information bottleneck (IB) Tishby & Zaslavsky (2015) plays a repre-  
 135 sentation trade-off between information compression and information preservation for neural learn-  
 136 ing tasks. Furthermore, VIB Alemi et al. (2017) leverages variational inference to facilitate the IB  
 137 neural compression. IBA Schulz et al. (2020); Gao et al. (2021) polishes the attribution information  
 138 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.

### 139 3 PRELIMINARIES

#### 140 3.1 PROBLEM SETTING

141 Given source noise  $N_{Orig}$  and injected noise  $N_{Div}$  are from a consistent distribution  $\mathcal{N}(\mu_G, \sigma_G^2)$ ,  
 142 where  $\mu_G$  and  $\sigma_G$  represent the means and standard deviations. Then, the modulated manifold of  
 143 2D/3D asset can be formulated as follows Schulz et al. (2020):

$$144 N_{Out} = \lambda N_{Orig} + (1 - \lambda) N_{Div}, \quad (1)$$

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

$$149 \tilde{z}_0 = \frac{z_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, c, t)}{\sqrt{\bar{\alpha}_t}}. \quad (2)$$

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

153 Naive noise interpolation based on a constant  $\lambda$  and other diversity-inducing methods (e.g., entropy  
 154 regularization, contrastive objective) are not robust to perturbation from  $N_{Div}$ . Our PROOF learns  
 155 the adaptive interpolation weight based on the closed-form solution of OTIB. We define our noise  
 156 finetuning as:

$$157 \theta^* = \operatorname{argmin}_\theta \mathbb{E}_{N_{Orig}, N_{Div}} [\mathcal{L}_{noise}(PROOF_\theta(N_{Orig}, N_{Div}), N_{Orig}) + \mathcal{L}_{info}(PROOF_\theta(N_{Orig}), \lambda)], \quad (3)$$

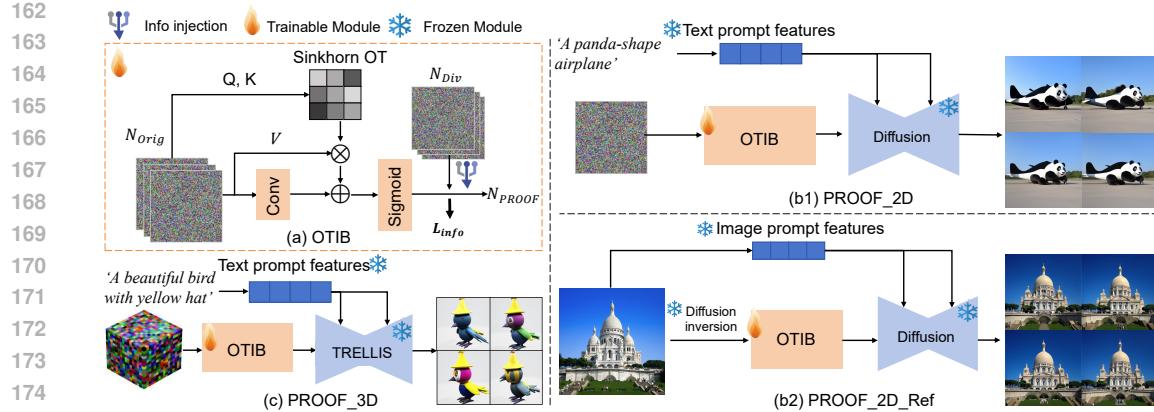


Figure 2: Method overview: as a plug-and-play content controller, PROOF can be employed for 2D/3D generation tasks, different architectures and model checkpoints. OTIB consists of a Sinkhorn Attention module and an information bottleneck module. We obtain  $N_{PROOF}$  by information compression of 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

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  $X$ :

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

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

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

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:

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

where  $\mu, \nu$  are discrete probability measures,  $\mu_i \geq 0, \nu_j \geq 0$  are probability masses ( $\sum_i \mu_i = \sum_j \nu_j = 1$ ),  $\delta_x$  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:

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

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  $\mathbf{C}_{ij} = d(x_i, y_j)$ ,  $\Pi(\mu, \nu) = \{\mathbf{T} \geq 0 \mid \mathbf{T} \mathbf{1}^N =$

$\mu, \mathbf{T}^\top \mathbf{1}^M = \nu\}$  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:

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

where  $G_\phi^{2D*}$  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 inverted noise to generate diverse images:

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

#### 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):

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

where  $G_\phi^{3D*}$  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:

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

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  $\lambda$  of OTIB.

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

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  $\lambda$  is 0, the whole manifold will be replaced by  $N_{Div}$ , which results in entire structure and appearance leakages. If  $\lambda$  is 1,  $Z$  excludes any form of diversity-inducing perturbations, which results in mode collapse. Qualitative analyses are illustrated in Sec. 5.

270 4.3 CLOSED-FORM SINKHORN-IB SOLUTION  
271272 We impose a Sinkhorn Attention module  $\mathcal{F}_{SA}$  in a spatial-OT view to improve contextual preserva-  
273 tion of PROOF. The Sinkhorn Attention algorithm is as follows:  
274275 **Algorithm 1** Sinkhorn-Attention Forward Pass

---

 276 1: **Input:** Feature map  $X \in \mathbb{R}^{B \times C \times H \times W}$   
 277 2:  $Q = \text{Conv\_Nd}(X)$ ,  $K = \text{Conv\_Nd}(X)$ ,  $V = \text{Conv\_Nd}(X)$  ▷ Learnable projections  
 278 3:  $A = QK^\top / \sqrt{C}$  ▷ Attention logits  
 279 4: **for**  $k = 1$  to  $n_{iters}$  **do**  
 280 5:    $A = A - \text{LogSumExp}(A, \text{dim} = 2)$  ▷ Row normalization  
 281 6:    $A = A - \text{LogSumExp}(A, \text{dim} = 1)$  ▷ Column normalization  
 282 7: **end for**  
 283 8:  $\mathbf{T} = \exp(A)$  ▷ Optimal attention weights  
 284 9: **return**  $\mathbf{T}V$  ▷ Transport applied to values  


---

285  
286 where  $Q, K, V \in \mathbb{R}^{B \times (HW) \times C}$  are Query, Key, Value tensors, respectively.  $A \in \mathbb{R}^{B \times (HW) \times (HW)}$   
287 is Attention logits matrix,  $\text{LogSumExp}(A)_i = \log \sum_j \exp(A_{ij})$ , and  $\mathbf{T}$  is Doubly-stochastic attention  
288 matrix. Our transport solution is established through:  
289

290 
$$\mathbf{T}_{ij} = \exp\left(\underbrace{\frac{q_i^\top k_j}{\sqrt{C}}}_{\text{Transport cost}} - \underbrace{\alpha_{OT}^i - \beta_{OT}^j}_{\text{Sinkhorn scalars}}\right) \quad (13)$$
  
291  
292

293 where  $\alpha_{OT}$  and  $\beta_{OT}$  are row and column normalization factors, respectively. The division by  $\sqrt{C}$   
294 stabilizes gradient flow. We consider the joint optimization objective of OTIB:  
295

296 
$$\min_{\lambda} \underbrace{I(R; Z)}_{\text{IB term}} + \gamma \underbrace{\langle A^*, \mathbf{C} \rangle}_{\text{Sinkhorn OT term}} + \epsilon H(A^*), \quad (14)$$
  
297

298 where  $Z = \lambda \odot N_{Orig} + (1 - \lambda) \odot N_{Div}$ ,  $A^* = \text{Sinkhorn}(\mathbf{C})$ , where  $\mathbf{C}_{ij} = \frac{\langle q_i, k_j \rangle}{\sqrt{d}}$ ,  $d = C$ .  
299300 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.  
301302 **Step 1:** Information Bottleneck Term Simplification. Under Gaussian assumptions, the mutual  
303 information and the gradient calculation are formulated as:  
304

305 
$$I(R; Z) = \frac{1}{2} \log \left( 1 + \frac{\lambda^2 \sigma_R^2}{(1 - \lambda)^2 \sigma_{N_{Div}}^2} \right), \frac{\partial I}{\partial \lambda} = \frac{\lambda \sigma_R^2 - (1 - \lambda) \sigma_{N_{Div}}^2}{\lambda^2 \sigma_R^2 + (1 - \lambda)^2 \sigma_{N_{Div}}^2} \quad (15)$$

306 **Step 2:** Sinkhorn Term Gradient. Using the Envelope Theorem and chain rule:  
307

308 
$$\frac{\partial \mathcal{L}_{OT}}{\partial \lambda} = \left\langle \frac{\partial A^*}{\partial \lambda}, \mathbf{C} \right\rangle + \left\langle A^*, \frac{\partial \mathbf{C}}{\partial \lambda} \right\rangle \approx \left\langle A^*, \frac{\partial \mathbf{C}}{\partial \lambda} \right\rangle, \quad (16)$$
  
309

310 where  $A^* = \text{diag}(u)K\text{diag}(v)$  with  $K = e^{-\mathbf{C}/\epsilon}$ .  $\frac{\partial \mathbf{C}_{ij}}{\partial \lambda} = \frac{\partial}{\partial \lambda} \left( \frac{\langle q_i, k_j \rangle}{\sqrt{d}} \right) = \frac{1}{\sqrt{d}} \langle q_i, \frac{\partial k_j}{\partial Z_j} \cdot \frac{\partial Z_j}{\partial \lambda} \rangle \approx$   
311  $\frac{1}{\sqrt{d}} \langle q_i, \frac{\partial k_j}{\partial N_{Orig}^j} \cdot \frac{\partial Z_j}{\partial \lambda} \rangle = \frac{1}{\sqrt{d}} \langle q_i, W_K(N_{Orig}^j - N_{Div}^j) \rangle$ .  
312313 **Step 3:** First-Order Optimality Condition Setting. The total gradient to zero:  
314

315 
$$\frac{\lambda \sigma_R^2 - (1 - \lambda) \sigma_{N_{Div}}^2}{\lambda^2 \sigma_R^2 + (1 - \lambda)^2 \sigma_{N_{Div}}^2} + \gamma \mathbb{E}_{A^*} \left[ \frac{\partial \mathbf{C}_{ij}}{\partial \lambda} \right] = 0 \quad (17)$$
  
316

317 **Step 4:** Closed-Form OTIB Solution. The optimal weights take the form (More details are in Ap-  
318 pendix A):  
319

320 
$$\lambda^* = \sigma \left( \frac{1}{\eta} \left( \gamma \cdot \text{Align}(N_{Orig}, N_{Div}) - \frac{\sigma_{N_{Div}}^2}{\sigma_R^2} \right) \right), \quad (18)$$
  
321

322 where  $\text{Align}(\cdot) = \mathbb{E}_{A^*} \left[ \frac{\partial \mathbf{C}_{ij}}{\partial \lambda} \right]$ ,  $\sigma(\cdot)$  is the sigmoid function, and  $\eta$  is a hyperparameter. The closed-  
323 form solution is aligned with Equ. 12, where  $\text{Conv}$  approximates  $\sigma^2$  ratio, and  $\mathcal{F}_{SA}$  approximates  
324  $\text{Align}$  term.  
325

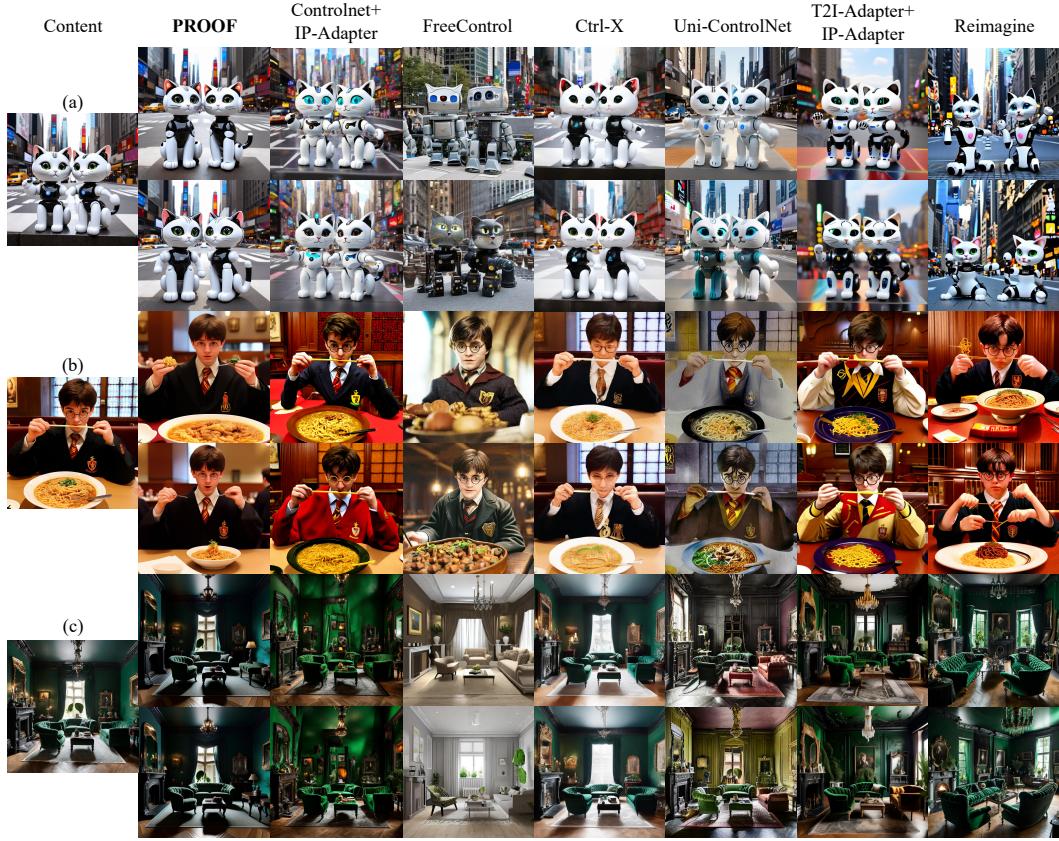


Figure 3: Qualitative results of PROOF\_2D, ControlNet + IP Adapter Zhang et al. (2023a); Ye et al. (2023), FreeControl Mo et al. (2024), Ctrl-X Lin et al. (2024), Uni-ControlNet Zhao et al. (2023), T2I-Adapter + IP Adapter Mou et al. (2024); Ye et al. (2023), and Reimagine AI (2023). Zoom in for better 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):

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

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

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

Our framework eliminates the need for feature mean/variance pre-calculation by leveraging the pre-defined properties of Gaussian noise ( $\mu_G=0$ ,  $\sigma_G=1$ ). 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  $\mu_G$  and  $\sigma_G$ , then the information compression metric of PROOF is:

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

Finally, the total loss of PROOF is formulated as:

$$\mathcal{L}_{PROOF} = \beta \mathcal{L}_{info} + \mathcal{L}_{noise}, \quad (22)$$

where  $\beta$  is the content-diversity tradeoff weight (Fig. 10a). Higher  $\beta$  usually intentionally relaxes contextual constraints but boosts the diversity (Fig. 13, Fig. 19).

378  
 379 Table 1: PROOF outperforms other SOTA methods in structure and appearance alignments and  
 380 robustness, measured by DINO ViT self-similarity and DINO-I. We report the inference time of  
 381 PROOF\_2D\_Ref, where diffusion inversion Mokady et al. (2023) is time-consuming. We assess  
 382 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. (2023a)	✓	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
RIVAL Zhang et al. (2023b)	✗	13.91	0.035	0.826	56.64	21.12	6.22	47.50	0.5431
Prompt-Free Diffusion Xu et al. (2024)	✓	10.91	0.025	0.824	22.35	19.92	6.21	40.36	0.4671
<b>PROOF (ours)</b>	✓	27.2	0.038	0.841	16.61	25.67	6.29	41.58	0.4342



436  
 437 Figure 4: Robust inference performance of PROOF across distinct latent resolutions. We set  $\beta$  of  
 438 PROOF as 0.2, which is aligned with  $\lambda = 0.8$ . It's efficient for Sinkhorn attention and information  
 439 bottleneck to finetune on low-resolution noise space while inferring on high-resolution latent.  
 440

## 5 EXPERIMENTS

441 Comprehensive qualitative and quantitative evaluations validate PROOF's dual capability in main-  
 442 taining content fidelity while enhancing generation diversity for digital asset creation. Training  
 443 protocol and baselines are presented in Appendix B and C. Additional results, e.g., golden noise  
 444 Zhou et al. (2025) finetune (Fig. 18), are shown in Appendix F.

### 5.1 QUANTITATIVE EVALUATION

445 Tab. 1 shows a quantitative comparison of natural images of datasets Lin et al. (2024). The content  
 446 alignment metrics include DINO ViT self-similarity Tumanyan et al. (2022), DINO-I Ruiz et al.  
 447 (2023) (details are explained in Appendix D). Note that PROOF shows consistent superiority on  
 448 self-sim and DINO-I scores. As for image quality, we utilize PickScore Kirstain et al. (2023), HPSv2  
 449 Wu et al. (2023), and Aesthetic Score (AES) Schuhmann (2023). We assess the diversity via LPIPS  
 450 Zhang et al. (2018) between the source image and the generated image. Note that PROOF\_2D using  
 451 SD-3 takes around 7s, which is more efficient for variant generation. Meanwhile, the subjective  
 452 metrics consist of quality, fidelity, and diversity without compromising fidelity. PROOF achieves  
 453 comparable user preference (Tab. 4).

### 5.2 QUALITATIVE RESULT

454 PROOF only learn noise representation supervised by itself based on OTIB. Visually comparable  
 455 results demonstrate that our implicit PROOF is a better workbench for highly correlated asset editing  
 456 (Fig. 3, Fig. 18, more examples in Appendix F), based on robust noise representation learning.  
 457

458 **Content transformation** Although PROOF applies intrinsic interpolation to manipulate noise, the  
 459 latent space compressed by VAE is already a high-dimensional manifold where nonlinear content  
 460 transformations are represented to some extent. That means the change of a specific noise point with  
 461 a certain channel and position is capable of imposing contextual transformation on several image-  
 462 level areas, therefore leading to visually object deformation or novel-view rendering (Fig. 14).



Figure 5: Robust local editing visualization. PROOF preserves local content layout and synthesizes controllable and diverse inpainting results with highly faithful details.



Figure 6: 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. 8.

**Train-Test resolution discrepancy** We conduct experiments concerning the latent resolution discrepancy between the fine-tuning and inference phases (Fig. 4). The overall contents of different finetuning models are consistent. However, the finetuning model employing 32-resolution data (Col 5) hardly captures local topological and textural details when dealing with 128-resolution inference.

**Local variation** PROOF can be employed by generation models equipped with the inpainting function to implement local content variation. As shown in Fig. 5, it's also important to provide uniform attention distribution based on optimal transport in the local editing scenario. PROOF synthesizes higher-fidelity and higher-quality human components. Furthermore, we evaluate PROOF on the edge controller (Fig. 11) and semantic editing (Fig. 12), which significantly strengthens PROOF's generalizability to broader applications.

**DiT-based model generalization** Whether PROOF can be applied to more advanced diffusion models featuring distinct architectural frameworks, e.g., Flux or SD3.5 based on Diffusion Transformer, has been further investigated. Conducting empirical validations on such state-of-the-art models substantially reinforces PROOF's ability to generalize and amplify its broader applicability across scenarios. Note that Figures 4, 5, 7 are all based on Flux Labs et al. (2025). Additionally, Fig. 17 shows robust variant results using SD-3.5 Esser et al. (2024). We report the computational complexity comparison of OTIB for different models in Tab. 3.

486

487  
488  
489  
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  $\lambda$ .

Configuration	self-sim ↓	DINO-I ↑	PickScore↑	HPSv2↑	AES↑	LPIPS
w/o IB $\triangleq$ Content	0	0.9999	20.95	34.64	5.50	0
Full PROOF $\beta=0.05$	<b>0.0314</b>	<b>0.9026</b>	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	<b>18.81</b>	<b>33.80</b>	5.37	0.4590
Naive interpolation $\lambda=0.9536$	0.0423	0.8650	14.83	33.15	5.39	0.4549
Entropy regularization $w=0.45$	0.0947	0.6299	12.30	31.20	<b>5.76</b>	<b>0.6790</b>
Contrast objective $w=0.085$	0.0320	0.9012	17.38	33.45	5.41	0.4565

490  
491  
492  
493  
494  
495  
Table 3: Computational complexity comparison of OTIB for different models.

Models	Spatial latent	MACs	Params	Inference time
SD-1.4, SD-1.5, SD-2	(B, 4, 64, 64)	134.64 MMac	100	0.1579s
SD-3, SD-3.5, Flux	(B, 16, 128, 128)	8.61 GMac	1.36 k	0.2185s

500  
501  
502  
503  
504  
505  
506  
507  
Figure 7: Comparison of naive blending and PROOF over a wide parameter range. Naive blending  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
7510  
7511  
7512  
7513  
7514  
7515  
7516  
7517  
7518  
7519  
7520  
7521  
7522  
7523  
7524  
7525  
7526  
7527  
7528  
7529  
7530  
7531  
7532  
7533  
7534  
7535  
7536  
7537  
7538  
7539  
7540  
7541  
7542  
7543  
7544  
7545  
7546  
7547  
7548  
7549  
7550  
7551  
7552  
7553  
7554  
7555  
7556  
7557  
7558  
7559  
7560  
7561  
7562  
7563  
7564  
7565  
7566  
7567  
7568  
7569  
7570  
7571  
7572  
7573  
7574  
7575  
7576  
7577  
7578  
7579  
7580  
7581  
7582  
7583  
7584  
7585  
7586  
7587  
7588  
7589  
7590  
7591  
7592  
7593  
7594  
7595  
7596  
7597  
7598  
7599  
75100  
75101  
75102  
75103  
75104  
75105  
75106  
75107  
75108  
75109  
75110  
75111  
75112  
75113  
75114  
75115  
75116  
75117  
75118  
75119  
75120  
75121  
75122  
75123  
75124  
75125  
75126  
75127  
75128  
75129  
75130  
75131  
75132  
75133  
75134  
75135  
75136  
75137  
75138  
75139  
75140  
75141  
75142  
75143  
75144  
75145  
75146  
75147  
75148  
75149  
75150  
75151  
75152  
75153  
75154  
75155  
75156  
75157  
75158  
75159  
75160  
75161  
75162  
75163  
75164  
75165  
75166  
75167  
75168  
75169  
75170  
75171  
75172  
75173  
75174  
75175  
75176  
75177  
75178  
75179  
75180  
75181  
75182  
75183  
75184  
75185  
75186  
75187  
75188  
75189  
75190  
75191  
75192  
75193  
75194  
75195  
75196  
75197  
75198  
75199  
75200  
75201  
75202  
75203  
75204  
75205  
75206  
75207  
75208  
75209  
75210  
75211  
75212  
75213  
75214  
75215  
75216  
75217  
75218  
75219  
75220  
75221  
75222  
75223  
75224  
75225  
75226  
75227  
75228  
75229  
75230  
75231  
75232  
75233  
75234  
75235  
75236  
75237  
75238  
75239  
75240  
75241  
75242  
75243  
75244  
75245  
75246  
75247  
75248  
75249  
75250  
75251  
75252  
75253  
75254  
75255  
75256  
75257  
75258  
75259  
75260  
75261  
75262  
75263  
75264  
75265  
75266  
75267  
75268  
75269  
75270  
75271  
75272  
75273  
75274  
75275  
75276  
75277  
75278  
75279  
75280  
75281  
75282  
75283  
75284  
75285  
75286  
75287  
75288  
75289  
75290  
75291  
75292  
75293  
75294  
75295  
75296  
75297  
75298  
75299  
75300  
75301  
75302  
75303  
75304  
75305  
75306  
75307  
75308  
75309  
75310  
75311  
75312  
75313  
75314  
75315  
75316  
75317  
75318  
75319  
75320  
75321  
75322  
75323  
75324  
75325  
75326  
75327  
75328  
75329  
75330  
75331  
75332  
75333  
75334  
75335  
75336  
75337  
75338  
75339  
75340  
75341  
75342  
75343  
75344  
75345  
75346  
75347  
75348  
75349  
75350  
75351  
75352  
75353  
75354  
75355  
75356  
75357  
75358  
75359  
75360  
75361  
75362  
75363  
75364  
75365  
75366  
75367  
75368  
75369  
75370  
75371  
75372  
75373  
75374  
75375  
75376  
75377  
75378  
75379  
75380  
75381  
75382  
75383  
75384  
75385  
75386  
75387  
75388  
75389  
75390  
75391  
75392  
75393  
75394  
75395  
75396  
75397  
75398  
75399  
75400  
75401  
75402  
75403  
75404  
75405  
75406  
75407  
75408  
75409  
75410  
75411  
75412  
75413  
75414  
75415  
75416  
75417  
75418  
75419  
75420  
75421  
75422  
75423  
75424  
75425  
75426  
75427  
75428  
75429  
75430  
75431  
75432  
75433  
75434  
75435  
75436  
75437  
75438  
75439  
75440  
75441  
75442  
75443  
75444  
75445  
75446  
75447  
75448  
75449  
75450  
75451  
75452  
75453  
75454  
75455  
75456  
75457  
75458  
75459  
75460  
75461  
75462  
75463  
75464  
75465  
75466  
75467  
75468  
75469  
75470  
75471  
75472  
75473  
75474  
75475  
75476  
75477  
75478  
75479  
75480  
75481  
75482  
75483  
75484  
75485  
75486  
75487  
75488  
75489  
75490  
75491  
75492  
75493  
75494  
75495  
75496  
75497  
75498  
75499  
75500  
75501  
75502  
75503  
75504  
75505  
75506  
75507  
75508  
75509  
75510  
75511  
75512  
75513  
75514  
75515  
75516  
75517  
75518  
75519  
75520  
75521  
75522  
75523  
75524  
75525  
75526  
75527  
75528  
75529  
75530  
75531  
75532  
75533  
75534  
75535  
75536  
75537  
75538  
75539  
75540  
75541  
75542  
75543  
75544  
75545  
75546  
75547  
75548  
75549  
75550  
75551  
75552  
75553  
75554  
75555  
75556  
75557  
75558  
75559  
75560  
75561  
75562  
75563  
75564  
75565  
75566  
75567  
75568  
75569  
75570  
75571  
75572  
75573  
75574  
75575  
75576  
75577  
75578  
75579  
75580  
75581  
75582  
75583  
75584  
75585  
75586  
75587  
75588  
75589  
75590  
75591  
75592  
75593  
75594  
75595  
75596  
75597  
75598  
75599  
75600  
75601  
75602  
75603  
75604  
75605  
75606  
75607  
75608  
75609  
75610  
75611  
75612  
75613  
75614  
75615  
75616  
75617  
75618  
75619  
75620  
75621  
75622  
75623  
75624  
75625  
75626  
75627  
75628  
75629  
75630  
75631  
75632  
75633  
75634  
75635  
75636  
75637  
75638  
75639  
75640  
75641  
75642  
75643  
75644  
75645  
75646  
75647  
75648  
75649  
75650  
75651  
75652  
75653  
75654  
75655  
75656  
75657  
75658  
75659  
75660  
75661  
75662  
75663  
75664  
75665  
75666  
75667  
75668  
75669  
75670  
75671  
75672  
75673  
75674  
75675  
75676  
75677  
75678  
75679  
75680  
75681  
75682  
75683  
75684  
75685  
75686  
75687  
75688  
75689  
75690  
75691  
75692  
75693  
75694  
75695  
75696  
75697  
75698  
75699  
75700  
75701  
75702  
75703  
75704  
75705  
75706  
75707  
75708  
75709  
75710  
75711  
75712  
75713  
75714  
75715  
75716  
75717  
75718  
75719  
75720  
75721  
75722  
75723  
75724  
75725  
75726  
75727  
75728  
75729  
75730  
75731  
75732  
75733  
75734  
75735  
75736  
75737  
75738  
75739  
75740  
75741  
75742  
75743  
75744  
75745  
75746  
75747  
75748  
75749  
75750  
75751  
75752  
75753  
75754  
75755  
75756  
75757  
75758  
75759  
75760  
75761  
75762  
75763  
75764  
75765  
75766  
75767  
75768  
75769  
75770  
75771  
75772  
75773  
75774  
75775  
75776  
75777  
75778  
75779  
75780  
75781  
75782  
75783  
75784  
75785  
75786  
75787  
75788  
75789  
75790  
75791  
75792  
75793  
75794  
75795  
75796  
75797  
75798  
75799  
75800  
75801  
75802  
75803  
75804  
75805  
75806  
75807  
75808  
75809  
75810  
75811  
75812  
75813  
75814  
75815  
75816  
75817  
75818  
75819  
75820  
75821  
75822  
75823  
75824  
75825  
75826  
75827  
75828  
75829  
75830  
75831  
75832  
75833  
75834  
75835  
75836  
75837  
75838  
75839  
75840  
75841  
75842  
75843  
75844  
75845  
75846  
75847  
75848  
75849  
75850  
75851  
75852  
75853  
75854  
75855  
75856  
75857  
75858  
75859  
75860  
75861  
75862  
75863  
75864  
75865  
75866  
75867  
75868  
75869  
75870  
75871  
75872  
75873  
75874  
75875  
75876  
75877  
75878  
75879  
75880  
75881  
75882  
75883  
75884  
75885  
75886  
75887  
75888  
75889  
75890  
75891  
75892  
75893  
75894  
75895  
75896  
75897  
75898  
75899  
75900  
75901  
75902  
75903  
75904  
75905  
75906  
75907  
75908  
75909  
75910  
75911  
75912  
75913  
75914  
75915  
75916  
75917  
75918  
75919  
75920  
75921  
75922  
75923  
75924  
75925  
75926  
75927  
75928  
75929  
75930  
75931  
75932  
75933  
75934  
75935  
75936  
75937  
75938  
75939  
75940  
75941  
75942  
75943  
75944  
75945  
75946  
75947  
75948  
75949  
75950  
75951  
75952  
75953  
75954  
75955  
75956  
75957  
75958  
75959  
75960  
75961  
75962  
75963  
75964  
75965  
75966  
75967  
75968  
75969  
75970  
75971  
75972  
75973  
75974  
75975  
75976  
75977  
75978  
75979  
75980  
75981  
75982  
75983  
75984  
75985  
75986  
75987  
75988  
75989  
75990  
75991  
75992  
75993  
75994  
75995  
75996  
75997  
75998  
75999  
75100  
75101  
75102  
75103  
75104  
75105  
75106  
75107  
75108  
75109  
75110  
75111  
75112  
75113  
75114  
75115  
75116  
75117  
75118  
75119  
75120  
75121  
75122  
75123  
75124  
75125  
75126  
75127  
75128  
75129  
75130  
75131  
75132  
75133  
75134  
75135  
75136  
75137  
75138  
75139  
75140  
75141  
75142  
75143  
75144  
75145  
75146  
75147  
75148  
75149  
75150  
75151  
75152  
75153  
75154  
75155  
75156  
75157  
75158  
75159  
75160  
75161  
75162  
75163  
75164  
75165  
75166  
75167  
75168  
75169  
75170  
75171  
75172  
75173  
75174  
75175  
75176  
75177  
75178  
75179  
75180  
75181  
75182  
75183  
75184  
75185  
75186  
75187  
75188  
75189  
75190  
75191  
75192  
75193  
75194  
75195  
75196  
75197  
75198  
75199  
75200  
75201  
75202  
75203  
75204  
75205  
75206  
75207  
75208  
75209  
75210  
75211  
75212  
75213  
75214  
75215  
75216  
75217  
75218  
75219  
75220  
75221  
75222  
75223  
75224  
75225  
75226  
75227  
75228  
75229  
75230  
75231  
75232  
75233  
75234  
75235  
75236  
75237  
75238  
75239  
75240  
75241  
75242  
75243  
75244  
75245  
75246  
75247  
75248  
75249  
75250  
75251  
75252  
75253  
75254  
75255  
75256  
75257  
75258  
75259  
75260  
75261  
75262  
75263  
75264  
75265  
75266  
75267  
75268  
75269  
75270  
75271  
75272  
75273  
75274  
75275  
75276  
75277  
75278  
75279  
75280  
75281  
75282  
75283  
75284  
75285  
75286  
75287  
75288  
75289  
7

540 REFERENCES  
541

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

543

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

546

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

550

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

555

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

558

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

561

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

564

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

568

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

572

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

576

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

579

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

582

583 Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Otseg: Multi-prompt sinkhorn attention for zero-  
584 shot semantic segmentation. In *European Conference on Computer Vision*, pp. 200–217. Springer,  
585 2024.

586

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

589

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

593

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

594 Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril  
 595 Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey,  
 596 Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini,  
 597 Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and  
 598 editing in latent space, 2025. URL <https://arxiv.org/abs/2506.15742>.

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

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

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

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

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

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

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

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

622 Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular  
 623 customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer*  
 624 *Vision and Pattern Recognition*, pp. 7964–7973, 2024.

625 Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe  
 626 Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image  
 627 synthesis. In *International Conference on Learning Representations*, 2024.

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

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

632 Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman.  
 633 Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Pro-  
 634 ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–  
 635 22510, 2023.

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

640 Christoph Schuhmann. Improved aesthetic predictor. 2023.

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

650

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

653 Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In  
 654 *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

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

658

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

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

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

668

669 Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free  
 670 diffusion: Taking “text” out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF*  
 671 *conference on computer vision and pattern recognition*, pp. 8682–8692, 2024.

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

675 Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng  
 676 Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Pro-  
 677 ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–  
 678 14255, 2023.

679

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

682

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

685

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

689

690 Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion  
 691 inversion chain. *Advances in Neural Information Processing Systems*, 36:30641–30661, 2023b.

692

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

696

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

700

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

704

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

## 702 A APPENDIX A: DETAILED DERIVATION OF CLOSED-FORM SOLUTION

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

$$706 \quad \frac{\lambda\sigma_R^2 - (1 - \lambda)\sigma_{N_{Div}}^2}{\lambda^2\sigma_R^2 + (1 - \lambda)^2\sigma_{N_{Div}}^2} + \gamma\text{Align} = 0 \quad (23)$$

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

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

$$712 \quad \lambda\sigma_R^2 - (1 - \lambda)\sigma_{N_{Div}}^2 = -\gamma\text{Align}(\lambda^2\sigma_R^2 + (1 - \lambda)^2\sigma_{N_{Div}}^2) \quad (24)$$

713 This form removes the denominator but introduces quadratic terms in  $\lambda$ .

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

$$716 \quad f(\lambda) = \lambda\sigma_R^2 - (1 - \lambda)\sigma_{N_{Div}}^2 + \gamma\text{Align} [\lambda^2\sigma_R^2 + (1 - \lambda)^2\sigma_{N_{Div}}^2] \quad (25)$$

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

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

- 721 • It's the midpoint of possible  $\lambda$  values
- 722 • The function is most linear in this region
- 723 • Higher-order terms are minimized here

### 725 4.1. Function value at $\lambda = 0.5$ :

$$726 \quad f(0.5) = 0.5(\sigma_R^2 - \sigma_{N_{Div}}^2) + 0.25\gamma\text{Align}(\sigma_R^2 + \sigma_{N_{Div}}^2) \quad (26)$$

728 This combines the linear difference and quadratic alignment terms.

### 729 4.2. First derivative:

$$731 \quad f'(\lambda) = \sigma_R^2 + \sigma_{N_{Div}}^2 + \gamma\text{Align} [2\lambda\sigma_R^2 - 2(1 - \lambda)\sigma_{N_{Div}}^2] \quad (27)$$

$$732 \quad f'(0.5) = \sigma_R^2 + \sigma_{N_{Div}}^2 + \gamma\text{Align}(\sigma_R^2 - \sigma_{N_{Div}}^2) \quad (28)$$

734 The derivative shows how sensitive the function is to  $\lambda$  changes.

735 **4.3. Linear approximation solution:** Using Taylor expansion:

$$737 \quad \lambda \approx 0.5 - \frac{f(0.5)}{f'(0.5)} = 0.5 - \frac{0.5(\sigma_R^2 - \sigma_{N_{Div}}^2) + 0.25\gamma\text{Align}(\sigma_R^2 + \sigma_{N_{Div}}^2)}{\sigma_R^2 + \sigma_{N_{Div}}^2 + \gamma\text{Align}(\sigma_R^2 - \sigma_{N_{Div}}^2)} \quad (29)$$

739 This gives us a first-order approximation of the optimal  $\lambda$ .

740 **4.4. Simplified linear expression:** When  $\gamma\text{Align}$  is relatively small compared to the variance terms:

$$743 \quad \lambda \approx \underbrace{\frac{\sigma_{N_{Div}}^2}{\sigma_R^2 + \sigma_{N_{Div}}^2}}_C + \underbrace{0.25\gamma \cdot \text{Align}}_K, \quad (30)$$

746 where  $C$  represents the baseline compression ratio, and  $K$  determines how strongly alignment af-  
 747 fects the result.

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

- 750 • When Align is too large,  $\lambda$  may exceed 1
- 751 • When Align is too small,  $\lambda$  may be less than 0

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

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

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

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

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

764     Using the inverse of the sigmoid function (logit function)  $\sigma^{-1}(y) = \ln\left(\frac{y}{1-y}\right)$ , we get:  
 765

$$\sigma^{-1}(C) = \ln\left(\frac{\sigma_{N_{Div}}^2}{\sigma_R^2}\right) \quad (32)$$

766     This ensures the sigmoid preserves the baseline behavior when Align = 0.  
 767

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

$$x = \frac{1}{\eta} \left( \gamma \cdot \text{Align} - \frac{\sigma_{N_{Div}}^2}{\sigma_R^2} \right), \quad (33)$$

774     where  $\eta > 0$  controls the steepness of the transition. The final solution becomes:  
 775

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

778     This closed-form solution is presented in step 4 of Section 4.3, and satisfies all our requirements:  
 779

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

## 785     B TRAINING PROTOCOL

788     We train our PROOF on Gaussian noise tensors with corresponding dimension shape of different  
 789     architectures, e.g., 4\*64\*64 Rombach et al. (2022), 16\*128\*128 Esser et al. (2024), 8\*16\*16\*16  
 790     Xiang et al. (2025).  $N_{Orig}$  and  $N_{Div}$  are random noises in each training step. As for PROOF\_3D,  
 791     we utilize 3D convolutions for SA and IB modules. We train PROOF for 20k iterations with one  
 792     NVIDIA RTX 4090 GPU. The training batch size is set to 1. During training, we employ Adam  
 793     Kingma & Ba (2014) with  $2 * 10^{-3}$  learning rate. We set  $\beta = 0.01$  for mild diversity (Figure  
 794     3a),  $\beta = 0.1$  for substantial diversity (Fig. 3b, Fig. 13), and  $\beta = 1$  for diversity with reference  
 795     constraints (Fig. 3c).

## 796     C BASELINES

797     There are several state-of-the-art controllable synthesis methods based on diffusion models. ControlNet Zhang et al. (2023a) and T2I-Adapter Mou et al. (2024) align diffusion priors to the external  
 798     control structures. We further apply IP-Adapter Ye et al. (2023) to them for better textural transfer.  
 799     These methods present low topological flexibility with restriction by the explicit structure alignment,  
 800     and limited textural fidelity with global appearance control. FreeControl Mo et al. (2024)  
 801     has large-scale content variance due to imprecise structure and appearance representations (col 4  
 802     in Fig. 3). Ctrl-X Lin et al. (2024) provides too-strict structure and appearance alignments, and  
 803     there are texture distortions. Uni-ControlNet Zhao et al. (2023) also suffers from the global appear-  
 804     ance representation (col 6 in Fig. 3). Reimagine AI (2023) produces uncontrollable content layout,  
 805     despite high image quality and diversity (col 8 in Fig. 3). RIVAL Zhang et al. (2023b) conducts  
 806     distribution alignment between the generative and inverse paths to realize semantic and structural  
 807     fidelity. Prompt-free Diffusion Xu et al. (2024) discards the text encoder and text prompts, which  
 808     may bring about semantic degradation. We evaluate all methods on SDXL v1.0 Podell et al. (2024)  
 809     when workable and on their pre-configured base models otherwise.

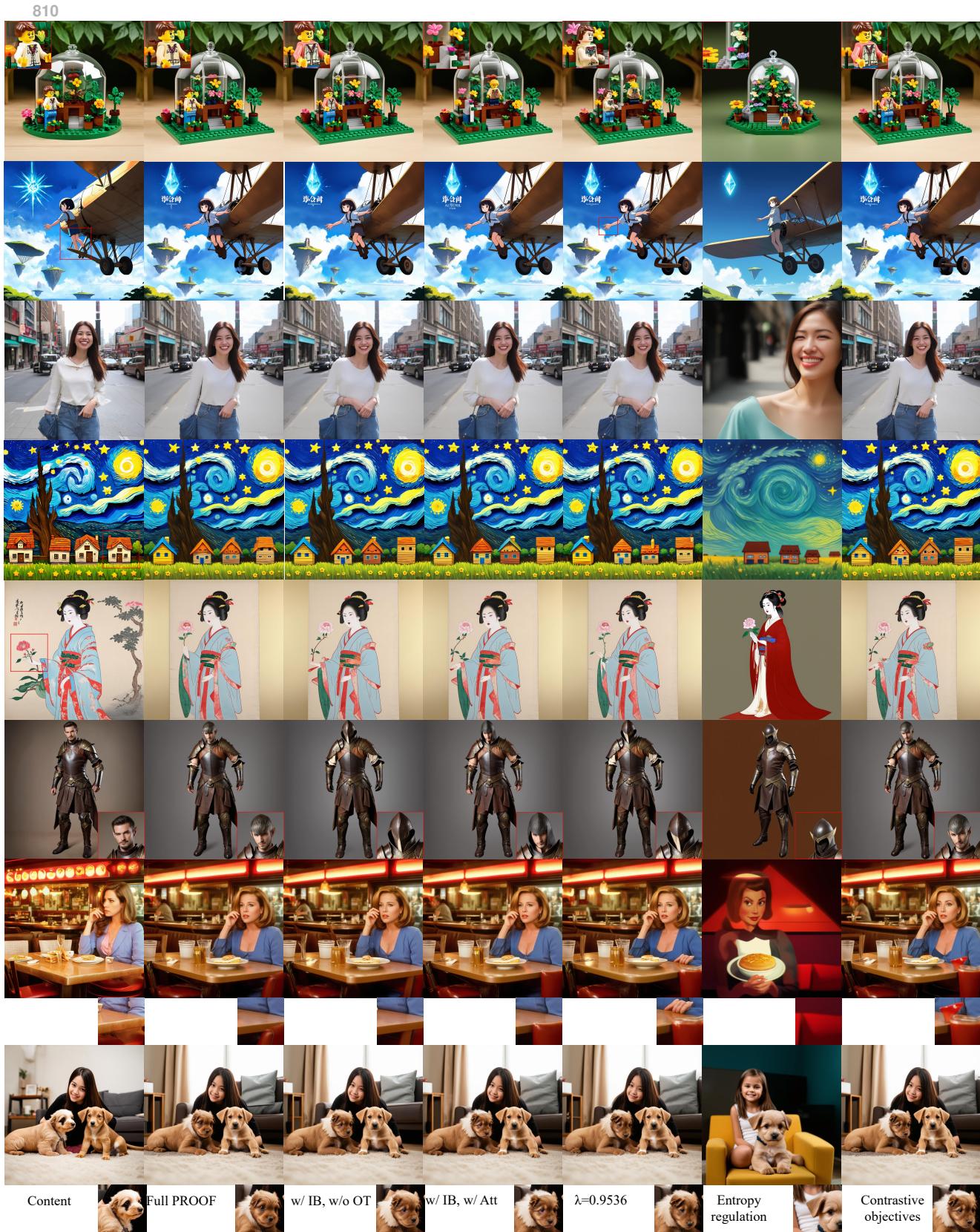


Figure 8: PROOF sufficiently preserves the global structure and appearance based on OTIB, while other variants show lower content fidelity. Zoom in for better observation.

864 **D EVALUATION METRIC**  
865866 Below is the explicit explanation of how DINO ViT self-similarity and DINO-I are calculated:  
867868 1. The structural consistency is quantified as:  
869

870 
$$\text{Self-sim} = \frac{1}{N} \sum_{i=1}^N \|\phi_{\text{DINO}}(I_{\text{Ref}})_i - \phi_{\text{DINO}}(I_{\text{Out}})_i\|_2^2, \quad (35)$$
  
871

872 where  $\phi_{\text{DINO}}$ : DINO-ViT base model (patch size=8) feature extractor,  $I_{\text{Ref}}$ : Reference input image,  
873  $I_{\text{Out}}$ : Generated output image,  $N$ : Number of feature vectors (layer\_num=11).  
874875 2. The appearance similarity is computed as:  
876

877 
$$\text{DINO-I} = \frac{\mathbf{v}_{\text{ref}} \cdot \mathbf{v}_{\text{out}}}{\|\mathbf{v}_{\text{ref}}\|_2 \|\mathbf{v}_{\text{out}}\|_2}, \quad (36)$$

878 where  $\mathbf{v}_{\text{ref}} = \phi_{\text{DINO}}^{[\text{CLS}]}(I_{\text{ref}})$ : DINO-ViT [CLS] token embedding of reference image,  $\mathbf{v}_{\text{out}} =$   
879  $\phi_{\text{DINO}}^{[\text{CLS}]}(I_{\text{out}})$ : DINO-ViT [CLS] token embedding of output image,  $\phi_{\text{DINO}}$ : DINO-ViT small model  
880 (patch size=16) feature extractor,  $\cdot$  denotes dot product.  
881882 **E DIVERSITY-BOOSTING METHODS**  
883884 In these diversity-inducing settings, we maintain the  $\mathcal{L}_{\text{noise}}$  of Equ. 22 to conduct global content  
885 preservation.  
886887 **E.1 CONTRASTIVE OBJECTIVE**  
888889 Given  $\text{flat\_Z} = \text{flatten}(Z) \in \mathbb{R}^{N \times d}$ ,  $\text{flat\_h} = \text{flatten}(R) \in \mathbb{R}^{N \times d}$ ,  $\text{flat\_l} = \text{flatten}(N_{\text{Div}}) \in \mathbb{R}^{N \times d}$ ,  
890 we calculate the cross-modal cosine similarity explicitly as  $\text{sim\_zh} = \cos(\text{flat\_Z}, \text{flat\_h})$ ,  $\text{sim\_zl} =$   
891  $\cos(\text{flat\_Z}, \text{flat\_l})$ ,  $\text{sim\_hl} = \cos(\text{flat\_h}, \text{flat\_l})$ . Then the contrastive objective loss is indicated as:  
892

893 
$$\mathcal{L}_{\text{contrast}} = w * (\text{MSE}(\text{sim\_zh}, \text{sim\_hl}) + \text{MSE}(\text{sim\_zl}, 1 - \text{sim\_hl})), \quad (37)$$

894 where  $w$  is the loss weight.  
895896 Note that the contrastive objective has some limitations as follows:  
897898 1. Exhibits significantly weaker robustness compared to PROOF under strong perturbations.  
899 2. Fails to perform effective representation learning at the manifold distribution level.  
900 3. Demonstrates notable scalability constraints in real-world applications.  
901 4. Generates structural and appearance artifacts (Fig. 6, Fig. 8).  
902903 **E.2 ENTROPY REGULARIZATION**  
904905 1. Input tensor flattening (flatten the  $i$ -th sample of  $Z$  from multi-dimensional to a vector)  
906

907 
$$Z_i^{\text{flat}} = \text{view}(Z_i, -1) \quad (38)$$

908 (i.e., flattened into a  $1 \times D$  vector, where  $D$  is the flattened dimension)  
909910 2. Softmax probability calculation of the  $j$ -th class for the  $i$ -th sample (compute class probabilities  
911 for each flattened sample)  
912

913 
$$p_{i,j} = \text{Softmax}(Z_i^{\text{flat}})_j = \frac{\exp((Z_i^{\text{flat}})_j)}{\sum_{k=1}^D \exp((Z_i^{\text{flat}})_k)} \quad (39)$$
  
914

915 3. Entropy calculation for a single sample ( $\epsilon$  is added to avoid meaningless logarithm)  
916

917 
$$H(Z_i) = - \sum_{j=1}^D p_{i,j} \cdot \log(p_{i,j} + \epsilon) \quad (40)$$

918  
 919 Table 4: PROOF exhibits competitive human preference percentages. Preference consistency is  
 920 87%, std. deviation is  $\pm 3.0\%$ , and the p-value of Wilcoxon is 0.016, which demonstrates the results  
 921 are statistically significant.

Methods	Quality $\uparrow$	Fidelity $\uparrow$	Diversity (subject to Fidelity) $\uparrow$
Uni-ControlNet Zhao et al. (2023)	78%	71%	75%
ControlNet + IP Adapter Zhang et al. (2023a); 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%
<b>PROOF (ours)</b>	89%	89%	<b>90%</b>

922  
 923 Therefore, the function definition of Entropy regularization is:  
 924  
 925

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

926 Note that entropy regularization has some limitations as follows (Fig. 6, Fig. 8):  
 927  
 928

1. Complete loss of background information.
2. Fails to ensure a minimal sufficient representation learning.
3. Poor robustness in structure and appearance preservation.

## 929 F ADDITIONAL RESULTS

930 In this section, we provide additional qualitative results of 2D (Figure 20, 22) or 3D asset (Figure  
 931 21) creation based on PROOF. Figure 19 indicates the workable function of OTIB to conduct  
 932 controllable diversity implicitly. Note that the detailed differences for small  $\beta$  are not obvious. Please  
 933 zoom in sufficiently and observe patiently.

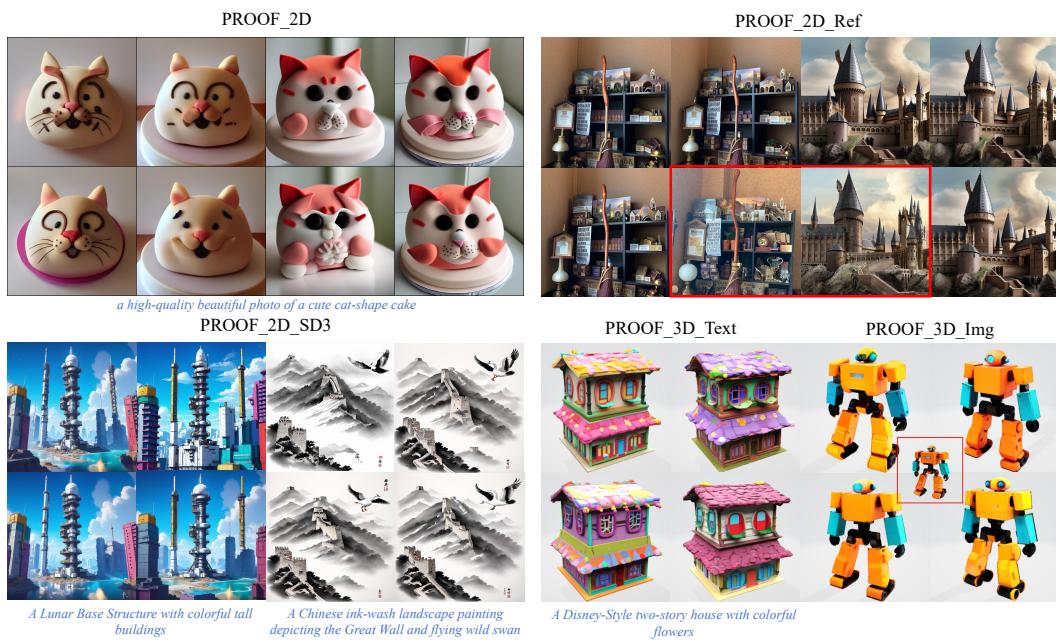
934 **Model select** As for PROOF\_2D\_Ref, we use Realistic\_Vision\_V4.0\_noVAE for diffusion  
 935 inversion and denoising, with ip-adapter-plus\_sd15 for appearance transfer. The VAE module is from  
 936 stabilityai-stable-diffusion-2-1-base. In Figure 22, iRFDS+Instantx uses the checkpoint of InstantX-  
 937 SD3.5-Large-IP-Adapter. In Figure 9, images of PROOF\_2D are synthesized based on the check-  
 938 point of Stable Diffusion v2-1\_512-ema-pruned. In Figure 18, we use stabilityai-stable-diffusion-xl-  
 939 base-1.0.

940 Note that because of the strong constraints from the image condition of TRELLIS Xiang et al.  
 941 (2025), there is little diverse space for direct PROOF\_3D\_Img. Therefore, we first synthesize the  
 942 image variants based on PROOF\_2D and then conduct 3D modeling based on the trellis-image-large  
 943 model. Text-based PROOF\_3D uses the trellis-text-xlarge model, as shown in Figure 21.

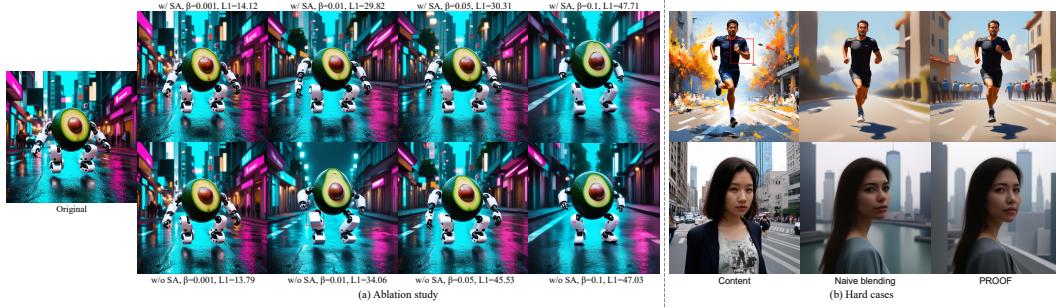
944 **User Study** We invite 100 domain experts to conduct the user study. First, we briefly explain the  
 945 highly correlated asset creation task. We suggest that users carefully observe the original content  
 946 and generated image variants obtained by 6 state-of-the-art methods and our proposed PROOF.  
 947 Each observed algorithm has 20 samples. These observers need to select the better image variant  
 948 set from 3 aspects: (a) overall quality, (b) overall fidelity considering structure and appearance,  
 949 (c) controllable diversity subject to the fidelity. The interface of our user study is shown in Figure 23.

950  
 951 Table 5: GENEVAL Ghosh et al. (2023) scores of different models. Robust PROOF preserves the  
 952 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 $\lambda=0.85$	0.61	0.96	0.67	0.54	0.80	0.23	0.46
PROOF $\beta=0.1$	0.70	0.98	0.80	0.65	0.91	0.32	0.55
PROOF $\beta=0.01$	0.72	0.98	0.83	0.67	0.92	0.35	0.57



993 Figure 9: Our proposed PROOF is an effective learning framework to synthesize highly correlated  
994 assets where variants exhibit consistent structure and appearance. Test-time PROOF facilitates high-  
995 quality 2D assets Esser et al. (2024) and 3D assets Xiang et al. (2025) with high contextual fidelity  
996 and controllable diversity, under any text or image condition (red boxes).



1007 Figure 10: (a) PROOF variants show that methods w/ SA preserve better appearance statistics than  
1008 those w/o SA. Higher  $\beta$  usually intentionally relaxes contextual constraints but boosts the diversity.  
1009 (b) The background lacks abundant details for large-scale information compression (e.g.,  $\lambda=0.8$ ),  
1010 while the human identity and pose are maintained well.



Figure 11: Integration of PROOF and a structure-guided controller. Despite being constrained by edge conditions, PROOF maintains structure and texture fidelity in local areas while still generating diverse variations. Under large-scale perturbation, PROOF performs robust variant generation.



Figure 12: PROOF with semantic editing Mokady et al. (2023) produces high-quality editing results considering structure and appearance preservation.

**Comparision with DSG** While achieving similar editing effects to DSG Epstein et al. (2023) in Figure 13, our *PROOF* doesn't require any explicit guidance, e.g., position, size, shape.



Figure 13: 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.

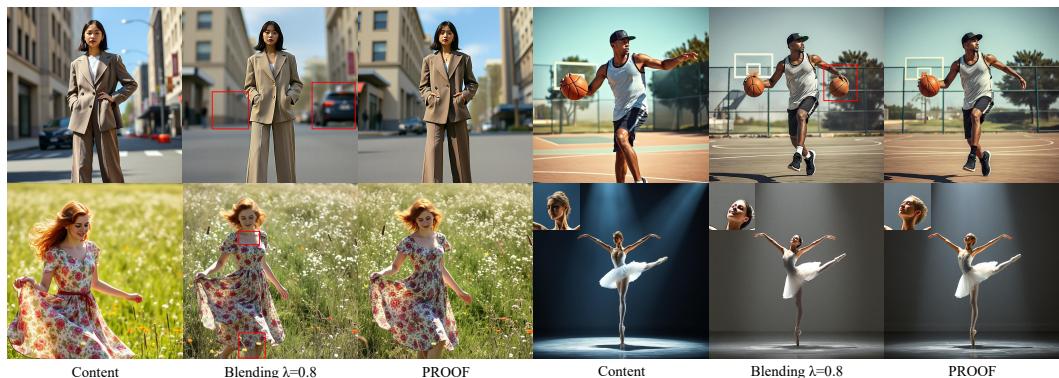


Figure 14: Content transformations based on FLUX.1-schnell with PROOF. We show some examples of human deformation with different poses and novel perspectives, which demonstrate that intrinsic interpolation to manipulate noise is efficient to model complex nonlinear transformation patterns. PROOF outperforms naive blending, as the latter often leads to noticeable content distortion and undesirable artifacts.



Figure 15: 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  $\lambda$ . PROOF preserves fine-grained structure and appearance features.



Figure 16: Variant comparison of PROOF and other image variation works. Zhang et al. (2023b) maintains alignment between the latent distributions of the generative and inverse paths to improve semantic and structural fidelity. Xu et al. (2024) eliminates the text encoder and text prompts, which may result in semantic degradation (e.g., face and teddy bear in Row 3). PROOF leveraging robust manifold manipulation preserves fine-grained structure and appearance features (Row 2). Moreover, adaptive interpolation via OTIB efficiently produces diverse high-fidelity image variants.

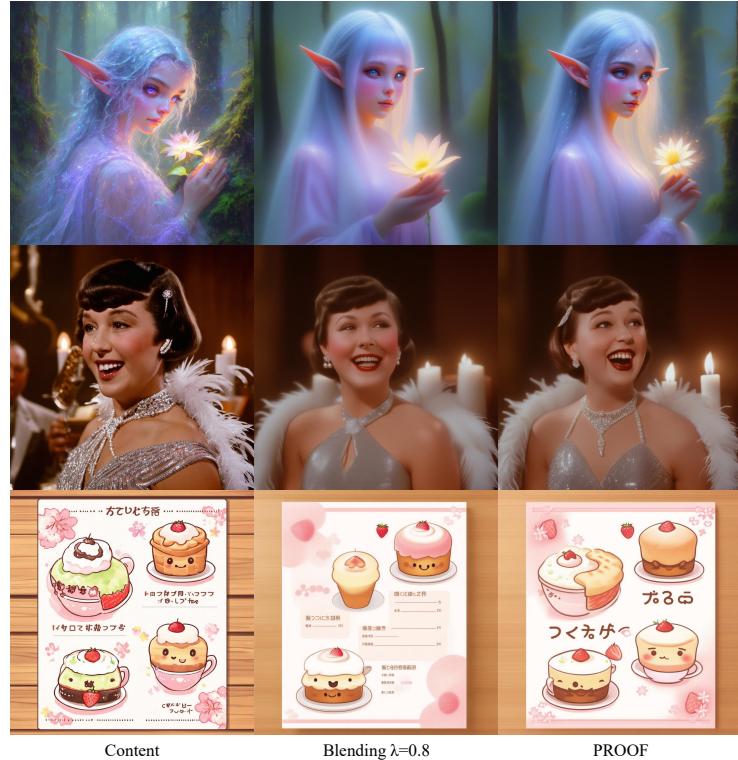


Figure 17: Additional visual results of PROOF\_2D based on Stability AI SD3.5 Medium. PROOF is more robust to defend against noise perturbation.

1134  
 1135 **Comparison with Golden Noise** Task Differentiation of Golden Noise Zhou et al. (2025) and  
 1136 PROOF: Golden Noise focuses on text-embedding alignment in noise space and embeds semantic  
 1137 information into noise for semantic fidelity. PROOF targets content-aligned variation generation by  
 1138 modifying local structure and appearance distributions for contextual fidelity with diversity. We pro-  
 1139 vide some comparative results in Fig. 18, which demonstrates that PROOF is powerful to synthesize  
 1140 high-fidelity and high-quality assets.

1140 Specifically, given standard noise as  $N_{Orig}$ , we obtain golden noise  $N_{Gold} = NPNet(N_{Orig}, c)$ .  
 1141 Moreover, standard PROOF and golden PROOF are implemented based on  $N_{Orig}$  and  $N_{Gold}$ , where  
 1142 the same  $N_{Div}$  is adaptively interpolated via OTIB. Note that both NPNet and PROOF leverage  
 1143 SDXL as the pretrained base model.



1185 Figure 18: Standard PROOF and Golden PROOF are based on the standard noise and golden noise,  
 1186 respectively. PROOF seems to produce more high-fidelity golden noise (col 3), and Zhou et al.  
 1187 (2025) exhibits low perturbation robustness (col 4).



Figure 19: PROOF effectively controls the structure and appearance of the content. Smaller tradeoff weight  $\beta$  puts content on a slight adjustment workbench, while larger  $\beta$  changes the content more obviously, but maintains the scene layout.

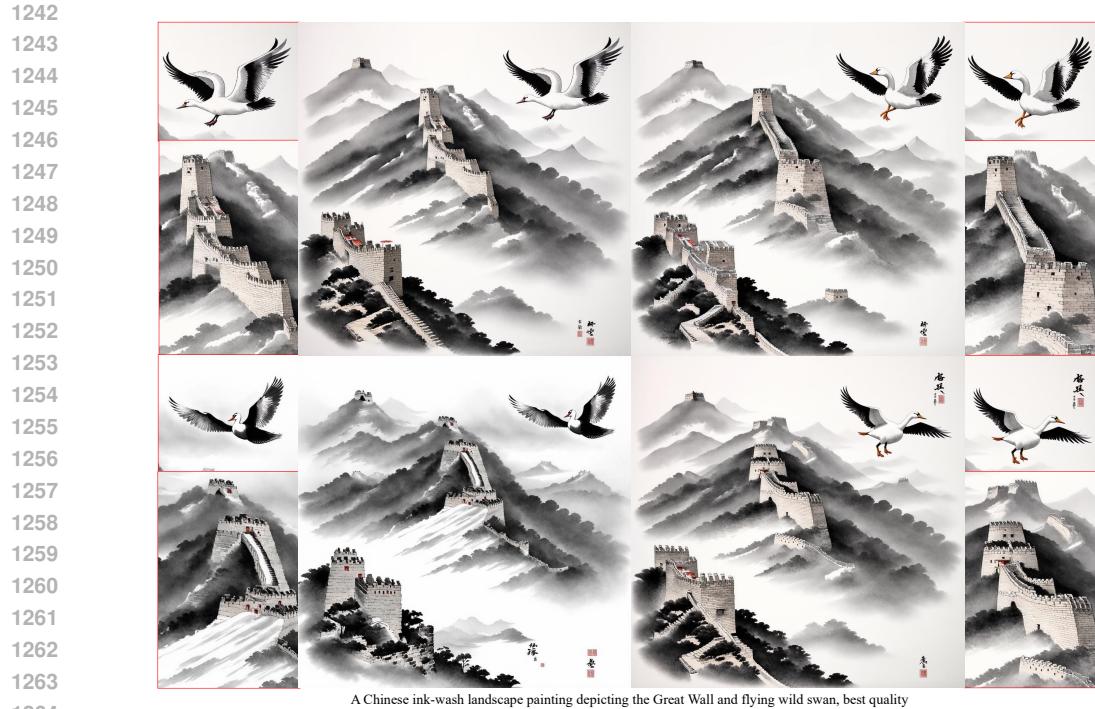


Figure 20: Image variants of the teaser figure 9 under magnified observation.

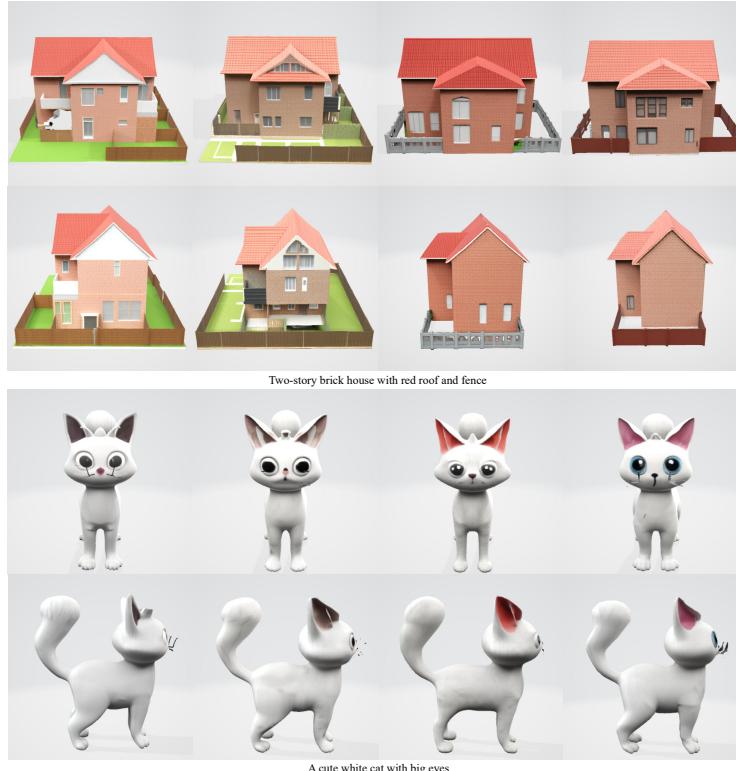


Figure 21: More qualitative results of PROOF\_3D based on TRELLIS Xiang et al. (2025).

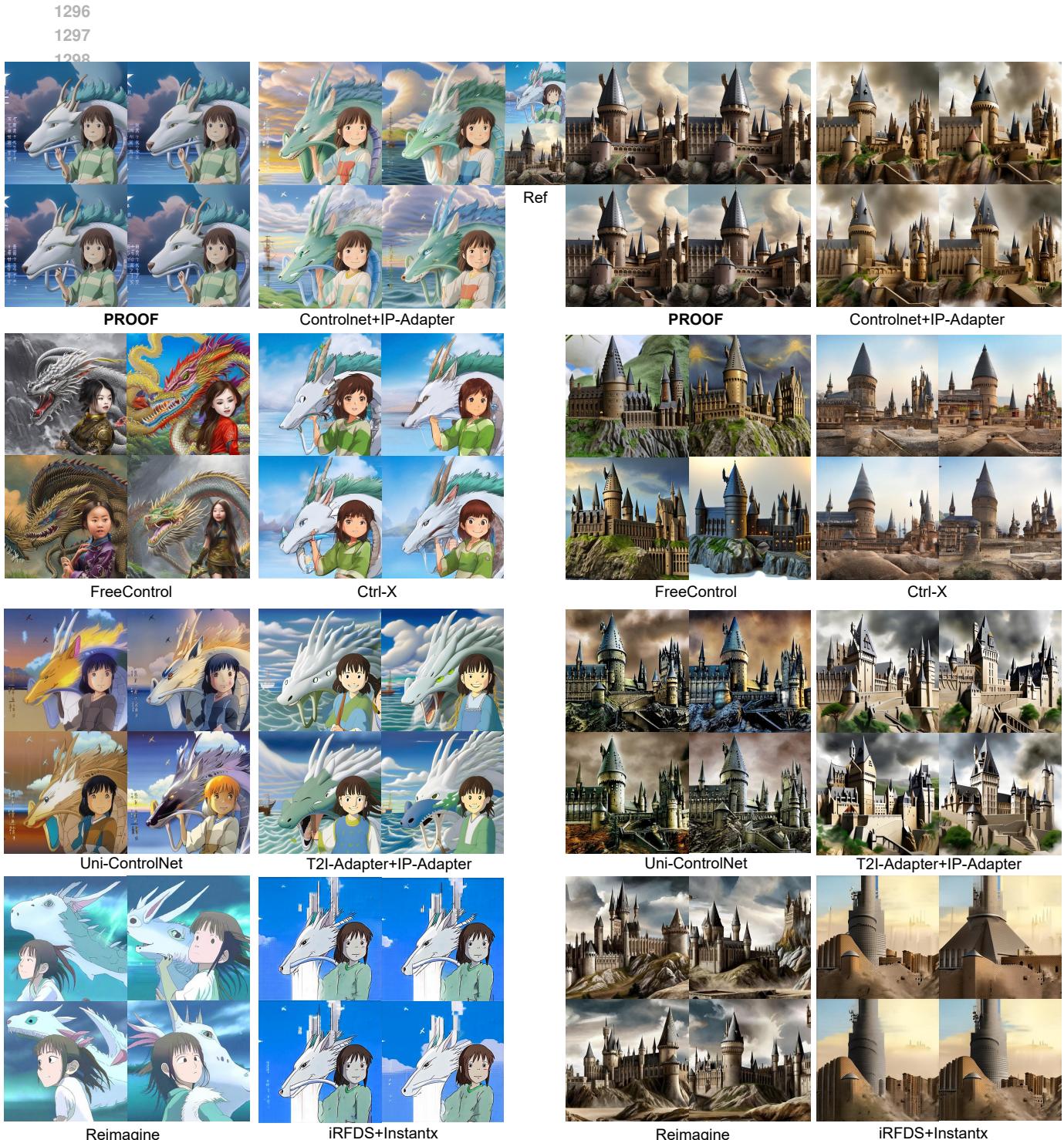


Figure 22: Qualitative results of PROOF\_2D\_Ref, ControlNet Zhang et al. (2023a); Ye et al. (2023), FreeControl Mo et al. (2024), Ctrl-X Lin et al. (2024), Uni-ControlNet Zhao et al. (2023), T2I-Adapter Mou et al. (2024); Ye et al. (2023), Reimagine AI (2023) and iRFDS Yang et al. (2025) on the wild images.

1350  
1351

Content



PROOF

Controlnet+  
IP-Adapter

FreeControl



Ctrl-X

Uni-  
ControlNetT2I-Adapter+  
IP-Adapter

Reimagine

1398  
1399  
1400  
1401  
1402  
1403

Figure 23: (a) Additional qualitative results of PROOF\_2D\_Ref, ControlNet Zhang et al. (2023a); 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.

Back

1/100

Next