$\mathbf{R^3}$-ADAPTER: PROGRESSIVE RESIDUAL REFINEMENT AND REPRESENTATIONAL ALIGNMENT FOR PERSONALIZED IMAGE GENERATION

Veddhanth Chakravarthy
Shiv Nadar University Chennai
veddhanth@gmail.com

Samir Kumar Das Mohapatra[1,2]
[1]Adobe Inc., India
[2]Shiv Nadar University Chennai
samir.dmp@gmail.com,dasmohap@adobe.com

Chandrakala Shanmuganathan
Shiv Nadar University Chennai
chandrakalas@snuchennai.edu.in

ABSTRACT

Personalized image generation with diffusion models has achieved remarkable success in single-subject scenarios, yet extending to multiple subjects remains challenging. We identify two critical limitations: the multi-granularity bottleneck, where single-level representations fail to capture semantic information from coarse categorical structure to fine-grained details, and the semantic alignment gap, where pixel-level optimization provides insufficient guarantees for maintaining subject identity during multi-subject composition. We propose **$\mathbf{R^3}$-Adapter**, a novel framework that addresses these challenges through Progressive Residual Refinement and Representation Alignment (REPA). Our method decomposes subject representations across four semantic levels through bounded residual corrections with timestep-adaptive routing, while REPA grounds the diffusion model's internal features using DINOv3 features via cross-attention and self-attention alignment. Comprehensive experiments demonstrate state-of-the-art performance: 6.24% improvement in CLIP-I and 12.24% in DINO on single-subject tasks, with even larger gains of 21.36% in DINO on challenging multi-subject scenarios. Ablations confirm that refinement and semantic alignment operate synergistically, achieving combined improvement over baselines.

1 INTRODUCTION

The rapid advancement of large-scale diffusion models has revolutionized image synthesis, enabling photorealistic generation from text descriptions with unprecedented quality and diversity Rombach et al. (2022), Podell et al. (2023). However, these models are trained on massive internet-scale datasets and excel primarily at generating generic instances of concepts—a "golden retriever," a "blue chair," or a "smiling person"—rather than specific individuals with unique identifying characteristics. This limitation fundamentally restricts their applicability in creative and commercial domains where users need to generate images of particular subjects: their own pet, a specific product, or themselves in novel contexts. While recent personalization methods have made significant progress in single-subject scenarios Wei et al. (2025), extending these approaches to multiple subjects simultaneously introduces profound new challenges Jin et al. (2025), the model must: (1) accurately preserve and maintain unique features of the different subjects; (2) prevent attribute leakage between subjects; and (3) compose them coherently according to the specified spatial relationships and interactions. Existing multi-subject methods extract subject-specific tokens through spatial grounding and disentangled attention mechanisms. While effective, these approaches suffer from two critical limitations that fundamentally constrain their quality ceiling:

Challenge 1: The Multi-Granularity Bottleneck. Subject identity is inherently hierarchical, encompassing information at multiple semantic levels. A specific subject's representation must simul-

taneously encode coarse categorical information, mid-level breed characteristics, part-based decomposition, and fine-grained instance-specific details. Current methods extract a single-level representation, forcing this fixed-capacity bottleneck to simultaneously encode all semantic granularities. This creates an unavoidable trade-off: tokens that capture broad structural understanding sacrifice fine-grained details, while tokens that encode specific markings lose compositional flexibility.

Challenge 2: The Semantic Alignment Gap in Multi-Subject Personalization. Diffusion models are trained exclusively on pixel-space reconstruction objectives—minimizing the error between predicted and actual noise in latent space. While effective for learning general image statistics, this objective provides no explicit guarantees about semantic preservation: two images with similar pixel distributions may contain entirely different semantic content, while semantically identical subjects may exhibit substantial pixel-level variation due to pose, lighting, expression, or background changes. This fundamental disconnect between pixel-level optimization and semantic consistency—what we term the semantic alignment gap—becomes critically pronounced in multi-subject personalization. When generating multiple specific subjects simultaneously, the model must satisfy numerous concurrent semantic constraints, each introducing independent requirements that interact in complex ways. The vastly expanded solution space admits many pixel-level optima that are semantically inconsistent: the model might minimize reconstruction error while inadvertently blending one subject's eye color with the second subject's, or generating a compositionally plausible scene where neither subject maintains their true identity.

We propose a framework that addresses both limitations through the following key contributions:

1. We identify and formalize the multi-granularity bottleneck in subject representation learning, demonstrating how single-level encodings fundamentally limit personalization quality.
2. We propose Progressive Residual Refinement, a novel architecture that explicitly decomposes representations across four semantic levels through bounded residual corrections.
3. We introduce Representation Alignment (REPA) using our dual alignment strategy that provides complementary supervision at both subject-specific and global structural levels. By aligning the diffusion model's internal representations with DINOv3's semantic space, we provide auxiliary supervision that directly targets semantic consistency.
4. We conduct comprehensive experiments on multi-subject and single-subject benchmarks, achieving state-of-the-art quantitative results and qualitative results.

2 Related Work

Diffusion Models and Spatial Control. Denoising Diffusion Probabilistic Models (DDPMs) Ho et al. (2020); Song et al. (2021) have emerged as the dominant paradigm for image generation. Latent diffusion models Rombach et al. (2022) dramatically improved efficiency by operating in compressed VAE latent space, with subsequent works like SDXL Podell et al. (2023) refining architecture and quality. Spatial control methods including ControlNet Zhang et al. (2023), T2I-Adapter Mou et al. (2023), and GLIGEN Li et al. (2023b) enable precise conditioning through edge maps, poses, and grounded descriptions.

Subject Personalization. DreamBooth Ruiz et al. (2023) pioneered personalization via full model fine-tuning on 3-5 reference images, while Textual Inversion Gal et al. (2023) optimizes only text embeddings with lower fidelity. Custom Diffusion Kumari et al. (2023) achieves DreamBooth-quality with $6-10\times$ speedup by fine-tuning only cross-attention projections. Feed-forward methods eliminate test-time optimization: IP-Adapter Ye et al. (2023) decouples image and text conditioning through separate cross-attention, while BLIP-Diffusion Li et al. (2023a) aligns BLIP features with diffusion semantics. These methods achieve efficient personalization but produce lower fidelity than fine-tuning approaches, particularly for fine-grained details.

Multi-Subject Generation. Initial composition-based methods like Mix-of-Show Gu et al. (2023) merge separately trained LoRA Hu et al. (2021) adapters but suffer from attribute leakage. Break-A-Scene Avrahami et al. (2023) uses masked diffusion loss with subject-specific tokens, while MS-Diffusion Wang et al. (2025)—the current state-of-the-art—employs a grounding resampler with phrase-level conditioning and bounding boxes. However, these methods still use single-level representations and lack explicit semantic grounding during generation.

Self-Supervised Vision Features. Self-supervised methods like DINO Caron et al. (2021), DINOv2 Oquab et al. (2023), and DINOv3 Siméoni et al. (2025) learn robust semantic representations with emergent properties like spatial part structure and pose invariance. Recent work Yu et al. (2024) demonstrates that aligning diffusion features with such representations accelerates convergence and enhances fidelity. Related techniques include diffusion self-distillation Cai et al. (2025) for zero-shot customization and preference optimization Mo et al. (2025); Jin et al. (2025) for aesthetic alignment.

3 Proposed Method

We present a novel framework for multi-subject personalized image generation that advances with two key innovations as seen in Figure 1: (1) a Progressive Residual Refinement mechanism with timestep-adaptive weighting that progressively enriches subject-specific token representations across multiple granularity levels, and (2) a Representation Alignment (REPA) framework that aligns the diffusion model's internal representations with DINOv3 features. These contributions address fundamental limitations in existing approaches: the inability to capture fine-grained subject details at different semantic levels and the lack of semantic grounding during the denoising process.

3.1 Problem Formulation and Architecture Overview

Given a set of N reference images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, where each $I_j \in \mathbb{R}^{H\times W\times 3}$ depicts a distinct subject S_j, and a text prompt P describing desired composition and context, our goal is to synthesize a photorealistic image $\hat{I}$ that: (a) accurately depicts all subjects $S_1, \ldots, S_N$ with high identity fidelity, (b) composes them according to the spatial and semantic relationships specified in P, and (c) maintains photorealism and coherent scene structure. Our method builds upon the Stable Diffusion XL (SDXL) architecture and follows the MS-Diffusion framework for handling multiple subjects. The overall pipeline consists of four key components: (i) a perceiver-based grounding resampler that extracts subject-specific visual tokens, (ii) a residual refiner that progressively enriches these tokens across multiple semantic levels, (iii) a timestep-adaptive router that dynamically weights different refinement levels based on the denoising timestep, and (iv) a representation alignment module that grounds the diffusion model's features using DINOv3 representations.

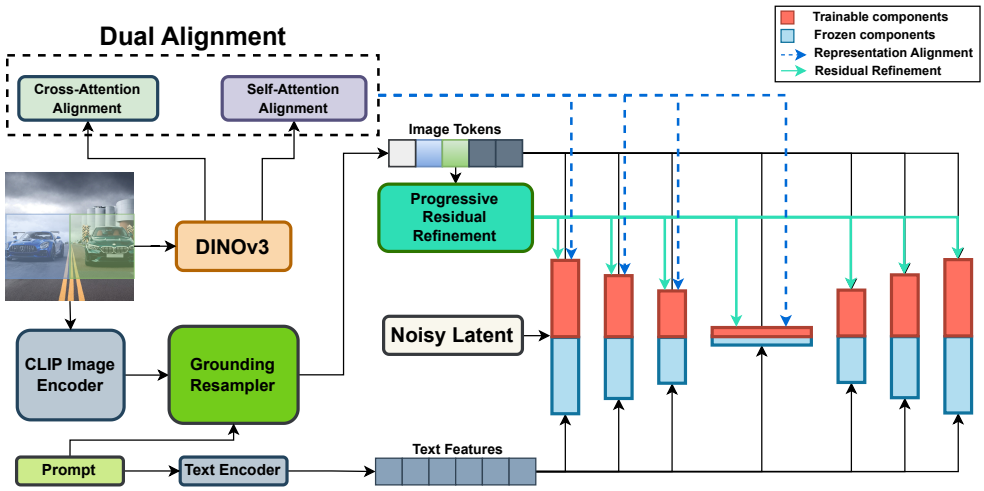

Figure 1: Overview diagram of R^3-Adapter: The pipeline processes reference images and bounding boxes through a perceiver-based grounding resampler. Our Progressive Residual Refinement distills the output of the resampler into 3 levels of refined tokens through its mechanisms. Cross-Attention alignment matches subject-specific diffusion features with corresponding DINOv3 CLS tokens within spatial regions, and Self-Attention alignment ensures global coherence by aligning full-image features with DINOv3 patch representations.

3.2 PROGRESSIVE RESIDUAL REFINEMENT

Subject identity is inherently hierarchical, encompassing information at multiple semantic granularities. Consider generating an image of a specific golden retriever: the representation must capture coarse categorical information ("dog", "canine"), mid-level breed characteristics ("golden retriever", "long fur"), and fine-grained instance-specific details (unique facial markings, specific fur patterns). Existing personalization methods extract a single-level representation from reference images. This creates a fundamental bottleneck: a fixed-capacity token set of dimension $M \times D$ must simultaneously encode information across all semantic levels. This single-level approach manifests in several failure modes: (1) Detail loss: fine-grained features like specific markings or textures are often omitted or hallucinated incorrectly; (2) Attribute bleeding: when generating multiple subjects, instance-specific details of one subject leak into others because the model lacks explicit hierarchical separation; (3) Compositional failures: the model struggles to compose subjects in novel poses or contexts because coarse structural understanding is entangled with instance-specific appearance. We address these through a progressive residual refinement architecture that progressively enriches base subject tokens across four semantic levels: L_0 (base categorical representation), L_1 (local part-based features), L_2 (part-level structural details), and L_3 (fine-grained instance-specific attributes). Critically, we formulate refinement as bounded residual corrections rather than independent representations. Each refinement level ℓ produces a small correction ΔT_ℓ that is added to the cumulative representation from previous levels ensuring (a) smooth progressive refinement without distribution shift, and (b) stable training dynamics through gradual warmup.

3.2.1 ARCHITECTURE

Given base subject tokens $T_{\text{base}} \in \mathbb{R}^{B\times N\times D}$ extracted from the resampler (where B is batch size, N is the number of tokens per subject, and $D = 1024$ is the embedding dimension), we apply three progressive refinement transformations:

Level 1: Local Part Attention. We employ windowed local attention with window size $w = 4$ to enforce part-based decomposition. This mechanism constrains each token to attend only to tokens within its local window, naturally discovering spatial parts (e.g., head, body, legs):

$$T_1 = T_{\text{base}} + \alpha \cdot s_1 \cdot \rho_1(\text{LocalAttn}(T_{\text{base}} + e_1)) \tag{1}$$

where $\alpha \in [0, 1]$ is a warmup coefficient , s_1 is a learnable residual scale initialized to 0.1, e_1 is a level-specific positional embedding, and ρ_1 is a residual projection network comprising LayerNorm followed by two linear layers with Tanh activation to bound corrections. Windowed attention provides a strong inductive bias: by limiting receptive fields, we encourage the emergence of compositional part-based representations. This is analogous to convolutional layers discovering local features in CNNs, but applied in the token space of subject representations.

Level 2: Part-Level Sparse Attention. We apply instance-specific sparse attention where each token dynamically attends to its top-$k = 8$ most relevant tokens, enabling cross-part interactions:

$$T_2 = T_1 + \alpha \cdot s_2 \cdot \rho_2(\text{SparseAttn}(T_1 + e_2 + \text{Skip}_1(T_{\text{base}}))) \tag{2}$$

where Skip_1 is a learnable linear projection providing direct connections from base tokens to higher refinement levels.

Level 3: Instance-Specific Refinement. The final level employs even sparser attention ($k = 4$) to capture the most discriminative instance-specific features:

$$T_3 = T_2 + \alpha \cdot s_3 \cdot \rho_3(\text{SparseAttn}(T_2 + e_3 + \text{Skip}_2(T_{\text{base}}))) \tag{3}$$

The final output $T_{\text{refined}} = [T_{\text{base}}; T_1; T_2; T_3]$ concatenates all levels, providing the UNet with access to representations at all semantic granularities simultaneously. Each level's residual projection ρ_ℓ follows an identical architecture designed to produce bounded corrections:

$$\rho_\ell(x) = \text{Linear}_2(\text{Tanh}(\text{Linear}_1(\text{LayerNorm}(x)))) \tag{4}$$

where both Linear_1 and Linear_2 are $\mathbb{R}^D \to \mathbb{R}^D$ projections. The Tanh activation bounds intermediate values to $[-1, 1]$, and when multiplied by the small scale $s_\ell \approx 0.1$ and warmup coefficient $\alpha \leq 1$, ensures residual corrections start small ($\approx \pm 10\%$ of token magnitude) and grow gradually during training. This is critical for preventing distribution shift.

3.3 TIMESTEP-ADAPTIVE ROUTING

The denoising process in diffusion models exhibits an inherent hierarchical structure that has been extensively documented in prior work Ho et al. (2020). Early timesteps (high noise levels, $t \approx 1000$) establish global composition, object placement, and coarse structure. Mid timesteps ($t \approx 500$) refine shapes, poses, and spatial relationships. Late timesteps ($t \approx 0$) add fine details, textures, and instance-specific attributes.

This temporal hierarchy naturally aligns with our semantic hierarchy: base tokens (Level 0) encode coarse categorical information useful at early timesteps, while fine-grained tokens (Level 3) encode details relevant at late timesteps. However, naively concatenating all levels treats them equally throughout generation, potentially confusing the model by providing irrelevant information at inappropriate stages. We leverage this observation through a timestep-adaptive router that dynamically weights the contribution of each refinement level based on the current denoising timestep.

Given timestep $t \in [0, 1000]$ and refined tokens, we compute timestep-dependent gating weights using Gaussian Radial Basis Function (RBF) interpolation between three anchor distributions:

$$w(t) = \sum_{k=1}^{3} \phi_k(t) \cdot \text{softmax}(\theta_k) \tag{5}$$

where $\phi_k(t) = \exp(-\|t - c_k\|^2/\tau^2)$ are RBF kernels centered at $c_1 = 0.95$ (early), $c_2 = 0.65$ (mid), $c_3 = 0.2$ (late) with temperature $\tau = 0.3$, and θ_k are learnable anchor logits. The anchor logits are initialized to favor base tokens at early timesteps, balanced distribution at mid timesteps, and detailed tokens at late timesteps. The final gated tokens are computed as:

$$T_{\text{gated}} = \text{Concat}_{i=0}^{3}(w_i(t) \cdot T_i) \tag{6}$$

This routing mechanism allows the model to naturally focus on structure during early denoising and progressively shift attention to finer details as synthesis progresses, mirroring the nature of the diffusion process itself.

3.4 REPRESENTATION ALIGNMENT WITH DINOV3 FEATURES (REPA)

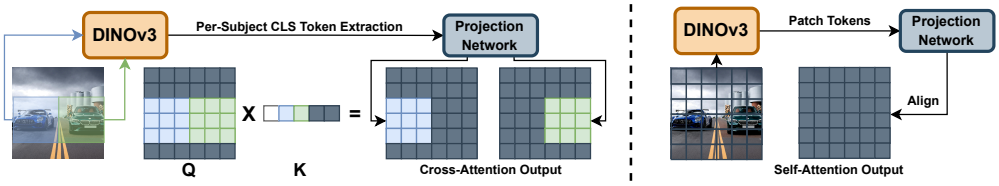

Figure 2: Detailed Diagram of Cross-Attention and Self-Attention alignment

While refinement enriches subject tokens, the UNet's attention mechanisms must learn to effectively utilize these representations. However, diffusion models are notoriously difficult to fine-tune: the denoising objective alone provides sparse, indirect supervision on the internal representations learned by attention layers. We address this through Representation Alignment (REPA), which grounds the UNet's internal features using frozen DINOv3 representations as auxiliary supervision.

The key insight is that DINOv3, pretrained via self-supervised learning on massive image datasets, has learned robust semantic representations of visual concepts. By aligning the UNet's attention features with these frozen representations, we provide auxiliary semantic guidance that helps the model maintain semantic coherence during fine-tuning. Critically, we use frozen orthogonal projections to map DINOv3 features to the UNet's embedding space, preventing gradient-based cheating where the alignment network could simply learn trivial mappings.

3.4.1 ORTHOGONAL PROJECTION ARCHITECTURE

DINOv3 extracts features of dimensionality 1024, while UNet hidden states vary across layers (typically 320–1280). Direct cosine similarity between mismatched dimensions is meaningless. Previous

work might use learnable linear projections, but this creates a critical problem: during backpropagation, the projection network can learn to map DINOv3 features to trivial targets that artificially minimize alignment loss without improving semantic quality.

We solve this through a two-layer frozen orthogonal projection:

$$P_{\text{frozen}} = W_2 W_1, \quad \text{where } W_1 \in \mathbb{R}^{512\times 1024},\ W_2 \in \mathbb{R}^{D\times 512} \tag{7}$$

Both W_1 and W_2 are initialized with orthogonal matrices and completely frozen during training. Orthogonal matrices preserve the semantic structure of DINOv3 features while enabling dimension compatibility. By the Johnson-Lindenstrauss Lemma Johnson & Lindenstrauss (1984), random projections to sufficiently high dimensions preserve pairwise distances with high probability, meaning our frozen orthogonal initialization preserves semantic relationships between concepts despite being randomly initialized. Crucially, no gradients flow through P_{frozen}, forcing the UNet to adapt its representations rather than the projection network learning trivial shortcuts.

3.4.2 Dual Alignment: Cross-Attention and Self-Attention

As seen in Figure 2, we apply representation alignment at two complementary components:

Cross-Attention Alignment. Cross-attention layers attend from spatial latent features to subject tokens, allowing the model to "look up" subject-specific information when generating different image regions. We align these attended features with subject-specific DINOv3 CLS tokens to ensure the model correctly interprets subject identity. For each subject j, we extract DINOv3 CLS tokens from the cropped reference image region corresponding to subject j's bounding box. During forward pass, we compute cross-attention outputs $h_{\text{cross},j}$ for region j, pool these features via mean-pooling, project the DINOv3 CLS token through P_{frozen}, and minimize:

$$\mathcal{L}_{\text{cross}} = \frac{1}{N}\sum_{j=1}^{N}\left[1 - \cos(\text{mean}(h_{\text{cross},j}), P_{\text{frozen}}(f^{\text{CLS}}_{\text{DINOv3},j}))\right] \tag{8}$$

This ensures cross-attention features attending to subject tokens align with that subject's semantic representation in DINOv3 space.

Self-Attention Alignment. While cross-attention focuses on subject-specific identity, self-attention layers in the UNet establish spatial relationships and global scene structure. We align self-attention features with whole-image DINOv3 patch tokens to ensure the model maintains coherent spatial layout and compositional understanding. We process the full reference image through DINOv3, extracting patch tokens $f^{\text{patch}}_{\text{DINOv3}}$. These patch tokens encode spatial structure: nearby patches have similar embeddings, and patches from the same object part cluster together. During the forward pass through the UNet, self-attention layers produce hidden states h_{self}, which we align with DINOv3 patch tokens:

$$\mathcal{L}_{\text{self}} = 1 - \cos(\text{mean}(h_{\text{self}}), P_{\text{frozen}}(\text{mean}(f^{\text{patch}}_{\text{DINOv3}}))) \tag{9}$$

We apply alignment selectively to mid-block and early down-block layers (layers closest to the latent bottleneck), where semantic representations are most abstract. The combined alignment loss is:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{self}} \tag{10}$$

Our complete training objective combines the standard diffusion loss with the alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diffusion}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} \tag{11}$$

where $\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,x,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2]$ is the standard noise prediction objective.

4 Experiments

4.1 Datasets

We utilize a curated dataset of 10,000 multi-subject compositions called Subject Dataset 10k from Huang et al. (2024). Each image contains 1-4 subjects with corresponding bounding box annotations, textual descriptions, and segmentation masks. The dataset covers diverse categories including

living entities, objects, vehicles, and scenes. Secondly, face image generation is particularly arduous, due to the complexities that arise when dealing with the high concentration of fine details in face images along with the human visual system's exceptional sensitivity to facial inconsistencies. So, we also train on a 30,000 single-subject face image dataset, CelebAMask-HQ from Lee et al. (2020). We utilize two complementary benchmarks to conduct comprehensive evaluation of our method: **MSBench** - A systematic benchmark framework adapted from Wang et al. (2025), which provides a thorough assessment of compositional generation capabilities across varying subject complexities. The benchmark encompasses 7 categories with a total of 830 unique subject combinations. We evaluate single-subject personalized generation using the CelebA-HQ Facial Identity Recognition Dataset from Na et al. (2022), which presents one of the most challenging scenarios for personalization methods, face image generation. For each combination, we generate images using 5 prompt variations with random seeds, allowing us to assess across diverse multi-subject scenarios. For evaluation, we take 20 real subject images, across 10 prompts, and 5 random seeds, giving us a total of 1000 generations per model tested. Our evaluation employs three metrics that provide comprehensive assessment of model performance: **CLIP-I** - Measures visual similarity between generated and reference images using CLIP's visual encoder. **CLIP-T** - Evaluates semantic alignment between generated images and text prompts using CLIP's joint vision-language embedding space. **DINO** - Leverages DINO's robust self-supervised representations to evaluate semantic consistency.

4.2 Quantitative Results

Table 1 presents comprehensive quantitative results comparing our method against state-of-the-art multi-subject personalization methods on both single-subject and multi-subject benchmarks.

Table 1: Performance comparison of different models on single and multi-subject tasks

	Single Subject			Multi Subject		
Model	**CLIP-I**	**CLIP-T**	**DINO**	**CLIP-I**	**CLIP-T**	**DINO**
λ-Eclipse	0.8006	0.2757	0.6201	0.7083	0.3035	0.3909
OmniGen	0.7940	0.2747	0.6239	0.7088	0.2814	0.4539
MS-Diffusion	0.7976	0.2730	0.6297	0.7115	0.3302	0.4648
R^3-Adapter (Ours)						
Progressive Residual Refinement	0.8023	0.2792	0.6369	0.7221	0.3391	0.4782
+Cross-Attention Alignment	0.8211	0.2845	0.6687	0.7347	0.3429	0.4973
+Self Attention Alignment	**0.8472**	**0.2918**	**0.7010**	**0.7458**	**0.3533**	**0.5267**

Our full model achieves state-of-the-art results on both single-subject and multi-subject benchmarks. On single-subject tasks, we obtain 6.24% improvement in CLIP-I, 6.31% CLIP-T improvement, and a DINO score improvement of 12.24% averaged over all the baselines. The particularly strong DINO improvement demonstrates our method's superior semantic anchoring. On the more challenging multi-subject scenarios, improvements are even more pronounced: CLIP-I reaches 5.11%, CLIP-T 16.31%, and DINO 21.36% improvement over the baselines. The larger relative gains in multi-subject settings validate our core hypothesis that the semantic alignment gap and multi-granularity bottleneck become critically limiting when handling multiple concurrent identity constraints.

Component-wise Ablation. Progressive Residual Refinement alone yields 2.9% improvement over MS-Diffusion, demonstrating that hierarchical semantic decomposition provides measurable benefits even without explicit alignment. Incorporating Cross-Attention Alignment provides an additional 4.0% gain (0.4973), confirming that subject-specific DINOv3 grounding substantially improves identity preservation. Finally, adding Self-Attention Alignment contributes another 5.9%, showing that global structural coherence and subject-specific identity are complementary objectives that must both be satisfied. The total improvement of +13.3% over baseline demonstrates true synergy: indicating that hierarchical refinement and dual alignment enable each other's effectiveness. More extensive ablation studies have been presented in A.2

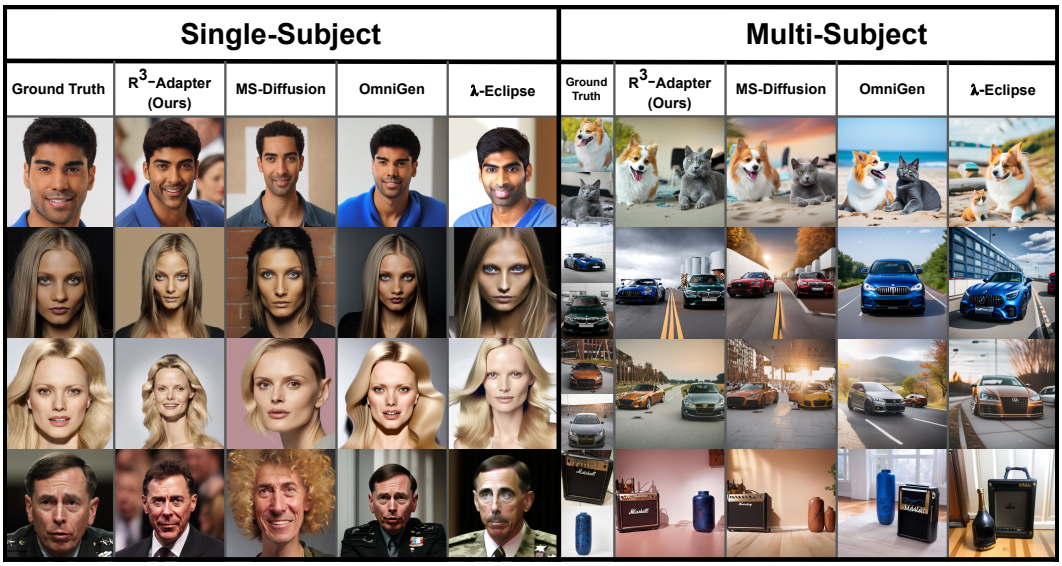

Figure 3: Qualitative results and comparisons against baselines

4.3 Qualitative Results

Figure 3 presents qualitative comparisons across single-subject and multi-subject scenarios, revealing critical differences in generation quality and identity preservation. OmniGen exhibits a characteristic overfitting pathology: while initially appearing to produce high-fidelity results, closer inspection reveals it essentially reproduces reference images nearly one-to-one rather than genuinely synthesizing novel poses or contexts. Generated images display telltale artifacts of synthesis—unnatural smoothness, inconsistent lighting, and the characteristic artifacts. λ-Eclipse suffers similar issues while additionally producing noticeably lower overall image quality, with visible artifacts, reduced photorealism, and inconsistent detail rendering. Critically, both OmniGen and λ-Eclipse exhibit severe feature bleeding in multi-subject scenarios: when generating two subjects simultaneously, their distinctive attributes blend together, producing chimeric subjects that average features rather than maintaining individual identities. MS-Diffusion demonstrates substantially improved compositional quality and reduced feature bleeding through its spatial grounding mechanisms, successfully generating distinct subjects in most cases. However, it consistently fails to capture fine-grained identity details: specific facial features like eye shape and expression, unique fur patterns on animals, distinctive textures on objects, and characteristic markings are either omitted entirely or incorrectly hallucinated. In contrast, R^3-Adapter generates photorealistic images with genuine novel synthesis without exhibiting memorization artifacts. Most importantly, our method preserves fine details across both single and multi-subject scenarios and unique attributes are faithfully reproduced.

5 Conclusion

We presented **$\mathbf{R^3}$-Adapter**, a principled framework for multi-subject personalized image generation that addresses fundamental limitations in existing methods. By identifying the multi-granularity bottleneck and semantic alignment gap, we introduced two innovations: Progressive Residual Refinement decomposes subject identity across four hierarchical semantic levels through bounded residual corrections, while Representation Alignment (REPA) grounds diffusion features using DINOv3 representations. Our timestep-adaptive routing mechanism weights semantic granularities throughout the denoising process, and our dual alignment strategy ensures both subject-specific identity preservation and global structural coherence. Extensive experiments validate our approach, achieving state-of-the-art results with pronounced improvements in challenging multi-subject scenarios where semantic constraints interact complexly. This work establishes that explicit semantic decomposition and representational grounding are critical for advancing personalization methods. Future work may explore extending our framework to video generation.

REFERENCES

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*, SA '23, pp. 1–12. ACM, December 2023. doi: 10.1145/3610548.3618154. URL http://dx.doi.org/10.1145/3610548.3618154.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18434–18443, June 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

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. *International Conference on Learning Representations (ICLR)*, 2023.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models, 2023. URL https://arxiv.org/abs/2305.18292.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation, 2024. URL https://arxiv.org/abs/2409.17920.

Qiaoqiao Jin, Siming Fu, Dong She, Weinan Jia, Hualiang Wang, Mu Liu, and Jidong Jiang. Focusdpo: Dynamic preference optimization for multi-subject personalized image generation via adaptive focus. *arXiv preprint arXiv:2509.01181*, 2025.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. URL https://arxiv.org/abs/2212.04488.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, 2023a. URL https://arxiv.org/abs/2305.14720.

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, 2023b. URL https://arxiv.org/abs/2301.07093.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Ligong Han, Ying Ba, and Dimitris N. Metaxas. Prefgen: Multimodal preference learning for preference-conditioned image generation. *arXiv preprint arXiv:2512.06020*, 2025.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, 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, 2023. URL https://arxiv.org/abs/2302.08453.

Dongbin Na, Sangwoo Ji, and Jong Kim. Unrestricted black-box adversarial attack using gan with limited queries. In *European Conference on Computer Vision*, pp. 467–482. Springer, 2022.

Maxime Oquab, Théo Darcet, Theo Moutakanni, Pierre Fernandez, Daniel Haziza, Francisco Massa, Marc Szafraniec, Maxim Berman, Joseph Salmon, Oron Ashual, Nicolas Ballas, Mahmoud Assran, Ishan Misra, Mike Rabbat, Diane Larlus, Piotr Bojanowski, and Armand Joulin. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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, 2023. URL https://arxiv.org/abs/2307.01952.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=PJqP0wyQek. Poster.

Yuxiang Wei, Yiheng Zheng, Yabo Zhang, Ming Liu, Zhilong Ji, Lei Zhang, and Wangmeng Zuo. Personalized image generation with deep generative models: A decade survey. *arXiv preprint arXiv:2502.13081*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Lvmin Zhang, Aditya Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

A Appendix

A.1 Training Strategy and Implementation Details

Our implementation utilizes PyTorch with mixed-precision training on a single NVIDIA L40S Tensor Core GPU with 48GB VRAM. Training procedure builds upon Stable Diffusion XL as our base generative model which we freeze and incorporate our adapter on top of.

We employ a two-phase optimization strategy:

Phase 1: Warmup with frozen refinement and alignment. Only base tokens and UNet attention layers are trained. The refiner produces zero corrections ($\alpha = 0$), and the alignment model is inactive ($\lambda = 0$), allowing the model to adapt to the multi-subject setting without distribution shift. **Phase 2:** The warmup coefficient α linearly increases from 0 to 1, gradually introducing corrections. We also introduce alignment loss with $\lambda = 0.01$ (empirically chosen), enabling the Dual Alignment Framework. We find that this setup allows the adapter to grasp image generation fundamentals while also aligning the generations to be more identity preserving and closer to the reference subjects.

A.2 Ablation Studies

Ablation 1: Hierarchical Refinement Depth. The progressive addition of hierarchical levels demonstrates systematic improvements across all metrics, with each level contributing distinct semantic granularities essential for comprehensive subject encoding. The 2-level hierarchy establishes categorical structure and part-based decomposition, providing a foundational representation. The 3-level configuration maintains similar performance while introducing cross-part structural relationships. The full 4-level hierarchy incorporates instance-specific details that capture fine-grained identity features, achieving +1.3% DINO improvement over the 2-level baseline. This progression reveals that four distinct semantic granularities are necessary to fully encode subject identity without creating representational bottlenecks. Attempting to compress this hierarchy into fewer levels loses critical information at intermediate abstraction scales, as evidenced by the consistent performance gaps between configurations.

Table 2: Ablation studies on multi-subject personalization tasks

Configuration	CLIP-I	CLIP-T	DINO
Ablation 1: Hierarchical Refinement Depth			
2-Level (L0+L1)	0.7159	0.3345	0.4723
3-Level (L0+L1+L2)	0.7164	0.3378	0.4721
4-Level (Full)	**0.7221**	**0.3391**	**0.4782**
Ablation 2: Alignment Layer Selection			
Down-Blocks	0.7378	0.3441	0.5012
Mid-Block Only	0.7412	0.3467	0.5089
Up-Blocks	0.7098	0.3334	0.4579
Mid + Down Blocks	**0.7458**	**0.3533**	**0.5267**
Ablation 3: Routing Strategy			
Static (25% each level)	0.7312	0.3401	0.4891
Timestep-Adaptive	**0.7458**	**0.3533**	**0.5267**

Ablation 2: Strategic Layer Selection for Alignment. The hierarchical nature of U-Net architectures imposes critical constraints on effective alignment strategies. Down-blocks provide solid baseline performance by balancing feature abstraction with spatial information (DINO: 0.5012). Mid-block alignment achieves stronger performance by operating at the highest abstraction level where subject identity is most concentrated (DINO: 0.5089). Critically, aligning with up-blocks consistently degrades performance across all metrics (DINO: 0.4579), as external supervision disrupts the carefully learned reconstruction pathways that synthesize spatial details from abstract mid-block representations. Our combined Mid + Down Blocks strategy achieves superior results through architectural synergy, with the mid-block preserving core identity semantics while down-blocks refine structural details, yielding +3.5% DINO improvement over mid-block-only alignment.

Ablation 3: Timestep-Adaptive vs. Static Routing. Static routing with uniform weights across all hierarchical levels fails to exploit the temporal structure of the diffusion process, where different denoising stages require different semantic granularities. Early timesteps benefit from coarse categorical structure, while later timesteps require fine-grained instance details. Our timestep-adaptive strategy employs RBF kernel interpolation between learned anchor weight distributions, enabling

smooth transitions that align hierarchical refinement with diffusion dynamics. This temporal adaptation yields substantial improvements of +7.7% in DINO similarity, demonstrating that when to apply each semantic granularity is as critical as what granularities to encode. This approach particularly excels in preserving intricate identity features during the detail-synthesis phase of generation.