RADIANCE: Relative Adaptive Denoising with IP-Adapter for Novel Concept Enhancement

Zi-Xiang Ni[1], Bo-Lun Huang[1], Teng-Fang Hsiao[1], Bo-Kai Ruan[1], and Hong-Han Shuai[1]*

National Yang Ming Chiao Tung University, Hsinchu, Taiwan
{zaq312.ee12, kevin503.ee12, bluedyee.ee09, bkruan.ee11, hhshuai}@nycu.edu.tw

Abstract. Text-to-image (T2I) diffusion models have achieved striking progress but still struggle to synthesize rare concepts involving unusual attribute-object pairings, often resulting in concept omission or semantic drift where a dominant entity overwhelms the generation. Tracing these failures to a lack of compositional balance during the denoising trajectory, we propose RADIANCE, a training-free framework that treats inference as a closed-loop feedback process. RADIANCE augments pretrained backbones with three modular components: (1) a Compositional Similarity Monitor (CSM) that tracks the emergence of objects and attributes in intermediate latents via CLIP-based feedback; (2) a Bidirectional Scale Controller (BSC) that applies a reactive "restoring force" using positive and negative IP-Adapter scales to rebalance biased trajectories; and (3) a Feedback Guidance Scheduler (FGS) that coordinates these updates across timesteps without additional training. We further extend the framework to multi-object prompts via Delayed Adapter Activation (DAA) and Layer-wise Alternating Guidance (LAG) to prevent premature concept fusion. By overlapping monitoring and denoising through pipelined execution, RADIANCE maintains competitive latency while significantly enhancing the per-sample success rate and effective throughput. Experiments on RareBench and T2I-CompBench demonstrate that RADIANCE consistently enhances compositional alignment and perceptual quality over state-of-the-art baselines.

1 Introduction

Text-to-image (T2I) diffusion models have rapidly become a foundation for visual content creation, enabling users to synthesize images directly from natural language prompts [4–7,10,32,35,37]. By combining with powerful text encoders such as CLIP [29], modern systems [26, 28, 34] can produce diverse and photorealistic samples that follow user intent. These capabilities have unlocked applications in digital art and design, advertising, and data augmentation for downstream

* Corresponding author.

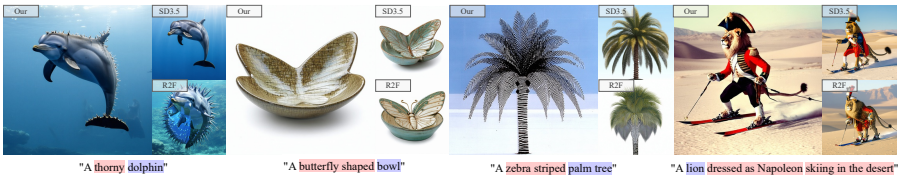

Fig. 1: Generated samples from RADIANCE across different rare concept prompts (attributes highlighted in red and objects in blue. Prior methods typically focus on either generating objects or capturing attributes, but struggle to merge both faithfully. For example, as shown in the first column, existing approaches [8, 27] either fail to generate thorns or lose the dolphin semantics. In contrast, our training-free approach, RADIANCE, yields coherent compositions that correctly combine rare attributes with their corresponding objects.

vision tasks [39]. Recent MM-DiT-based architectures *e.g.*, SD 3.5 [8], further improve scalability and cross-modal reasoning, narrowing the gap between model outputs and human imagination.

However, even the strongest T2I diffusion models still struggle when prompts describe *rare concepts*, such as objects with unusual attributes, shapes, or textures (*e.g.*, "a thorny dolphin" or "a zebra striped palm tree" in Fig. 1). In practice, training data follow a long-tailed distribution: frequent attribute-object combinations are well represented, while rare concepts appear only a handful of times [16, 27]. As a result, models tend to fall back to frequent co-occurrences, which leads to attribute omission, entangled attributes, or incorrect bindings. Prior work on compositional and attribute aware generation [2, 3, 11, 13, 16, 19, 21, 22, 41] improves how colors, shapes, and parts attach to a base object, but mainly for common concepts. More recent training-free strategies such as Rare to Frequent (R2F) [27] and prompt editing methods [33] replace rare tokens with semantically related frequent proxies and then switch back to the original prompt at selected timesteps, partially recovering the presence of rare concepts.

Despite the advances, existing approaches exhibit characteristic failure modes for rare concept generation. The rare attribute often materializes as disjoint entity in the image (*e.g.*, for the prompt *"a butterfly shaped bowl"* the model produces a normal bowl and a separate butterfly instead of a bowl whose silhouette forms a butterfly in Fig. 1, or the base object loses its defining characteristics and becomes distorted (*e.g.*, for *"a thorny dolphin"* the dolphin collapses into a spiky creature that no longer looks like a dolphin in Fig. 1. In both cases, the semantic balance between the rare attribute and the base object is broken. Prompt-level methods such as R2F [27] improves rare word expression but does not monitor object-attribute alignment, failing to correct step-wise drift when one component dominates. Other controllable generation methods [9, 31] use *static* guidance scales, making them less adaptive to prompt dynamics.

To resolve these compositional failures, we introduce **RADIANCE** (Rare Attribute Diffusion with Adaptive Novel Concept Enhancement), a training-free framework that reframes inference as a closed-loop feedback process. RADIANCE augments pretrained T2I backbones with three lightweight modules op-

erating entirely at inference time. First, the Compositional Similarity Monitor (CSM) utilizes CLIP to extract real-time similarity signals from intermediate latents to track the emergence of both the base object and its rare attributes. Second, the Bidirectional Scale Controller (BSC) translates these signals into a reactive "restoring force"; it applies positive or negative IP-Adapter scales to either reinforce lagging concepts or actively suppress dominant ones, thereby maintaining semantic equilibrium. Finally, the Feedback Guidance Scheduler (FGS) coordinates these scale updates across the denoising trajectory, ensuring the rare attribute is faithfully integrated while preserving the base object's identity. Together, these components form a robust controller that rebalances concept influence online without modifying the underlying diffusion model.

Furthermore, building on this single-object controller, RADIANCE extends to prompts with multiple rare objects: it first synthesizes attribute-faithful single-object images as references and then uses them as guidance for multi-object generation with *delayed adapter activation* and *layer-wise alternating guidance*, which prevents premature fusion and preserves object identities in the final composition. The contributions of this work can be summarized as follows:

- We propose RADIANCE, a training-free method that combines similarity monitoring, bidirectional scaling, and an adaptive guidance scheduler to maintain balance between rare attributes and base objects along the denoising trajectory for single object prompt.
- We further extend RADIANCE to multi-object rare compositions by first synthesizing attribute faithful single object reference images and then using delayed and layer-wise guidance, which prevents premature fusion and preserves distinct object identities in the final composition.
- Experiments on RareBench and T2I-CompBench manifest that RADIANCE significantly improves compositional alignment and visual fidelity over SOTA baselines. These gains are rigorously validated through automated scoring and extensive human preference studies.

2 Preliminary

2.1 Problem Formulation

Let c denote a text prompt. We decompose c into one object term o and m attribute terms $A = \{a_i\}_{i=1}^{m}$, where each term is a word or a short phrase and A is treated as an unordered set and $m \in \mathbb{N}_{\geq 0}$. The goal of compositional text-to-image generation is to produce an image x_0 that depicts the object o while satisfying the attributes $\{a_i\}_{i=1}^{m}$ as faithfully as possible. For example, o may be *"monkfish"* and the attributes may include *"horned"* and *"bearded"*.

Rare Concepts. A concept c is *rare* if its probability under the training distribution is near zero:

$$p_{\text{data}}(c) \approx 0, \qquad c \in \mathcal{C}_R \subseteq \mathcal{C}.$$

While a model may possess sufficient priors for individual components (**atomic concepts**), the **joint distribution** of their pairing is often poorly represented, defined here as **compositional rarity** where $P_{data}(c) \approx 0$. This data sparsity induces a "prior dominance" effect during sampling: the high-frequency prior of the base object (e.g., a standard dolphin) effectively suppresses the lower-frequency rare attribute (e.g., thorns). Consequently, the denoising trajectory gravitates toward frequent co-occurrences, leading to attribute omission or semantic entanglement even when the model is capable of synthesizing the individual components in isolation [14, 27].

2.2 Related Work

Text-to-Image Models Diffusion models have become the dominant framework for T2I generation, achieving remarkable visual fidelity and strong semantic alignment [4–7, 10, 32, 35, 37]. For example, Latent Diffusion Models (LDMs) [4, 28, 32] pioneered the integration of CLIP-based semantic conditioning with the efficiency of a VAE latent space [17]. More recent architectures based on MM-DiT, such as Stable Diffusion 3.5 [8] and FLUX [18], extend scalability and cross-modal reasoning by incorporating larger language models like T5 [30], thereby narrowing the gap between model outputs and human imagination. Despite the advancements, generating coherent and faithful images for rare or highly compositional prompts remains challenging, as models tend to overfit to frequent co-occurrences and deviate from the intended composition [3, 25, 27].

Rare Concept and Compositional Generation Growing research focuses on attribute binding and compositional control in diffusion models [2, 3, 13, 16, 19, 21, 22, 41], aiming to better associate attribute with a base object. However, these methods perform poorly on rare concepts due to the long-tailed nature of web-scale data, which biases models toward frequent patterns and underrepresents rare ones [27]. Training-free approaches such as R2F [27] address this issue by substituting rare tokens with semantically related proxies and reverting to the original prompt at specific timesteps. **Yet, they do not explicitly correct the balance between the rare attribute and its base object during denoising.** In contrast, our method treats inference as a feedback process—continuously measuring object–attribute alignment and adaptively rebalancing their influence for more coherent and accurate generation.

3 RADIANCE

3.1 Overview

The composition of rare attribute-object pairs frequently fails as the sampling process gravitates toward familiar correlations, leading to either attribute omission or the loss of base object identity. To resolve this imbalance, we propose RADIANCE, which reframes inference as a feedback-driven process that continuously monitors and adjusts compositional alignment. Equipped with a flow-based

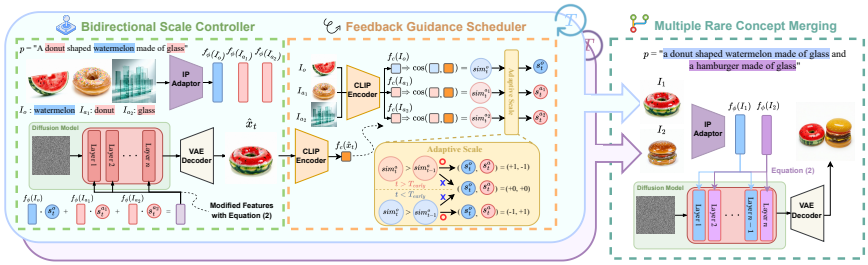

Fig. 2: Overview of the RADIANCE pipeline. At each denoising step, the *Feedback Guidance Scheduler* (Sec. 3.4) computes adaptive guidance. The resulting scale is then applied by the *Bidirectional Scale Controller* (Sec. 3.3) to inject reference information into the diffusion process. Finally, multiple rare-concept outputs are merged via the *Multiple Rare Concepts Merging* (Sec. 3.5) that composed of DAA and LAG. Together, these components ensure coherent and accurate generation of rare concepts.

backbone, the framework utilizes a Compositional Similarity Monitor (CSM) to decode intermediate latents and evaluate the emergence of each concept via CLIP-based similarity signals. Based on this real-time feedback, the Bidirectional Scale Controller (BSC) applies positive or negative scaling to the IP-Adapter pathway, providing a reactive "restoring force" that reinforces lagging components or suppresses dominant ones. This capability for negative scaling is especially critical, as it allows the model to actively counter excessive attribute influence rather than relying on additive guidance. Finally, the Feedback Guidance Scheduler (FGS) orchestrates the temporal policy, prioritizing the initial emergence of both concepts before shifting focus toward preserving object structural identity as the generation converges.

This feedback loop operates without additional training or gradients and requires only a single decode per step, ensuring both efficiency and practicality. For multi-object prompts (Sec. 3.5), we extend the same principle by activating multiple IP-Adapter branches layer-wise. Generated reference images of rare concepts are injected at different layers, allowing each object to first form independently before gradually interacting in later stages. By unifying measurement, actuation, and scheduling within a single controller, RADIANCE preserves semantic balance and prompt fidelity throughout sampling, enabling coherent generation for both single- and multi-object rare compositions without retraining. Fig. 2 shows the model architecture of the proposed RADIANCE.

3.2 Compositional Similarity Monitor (CSM)

To track compositional alignment during inference, we introduce the *Compositional Similarity Monitor (CSM)*, which measures how object and attribute contributions evolve at each denoising step. CSM requires a signal that is both lightweight, *i.e.*, allowing stepwise computation, and reliable enough to indicate whether the current generation is dominated by the object or the at-

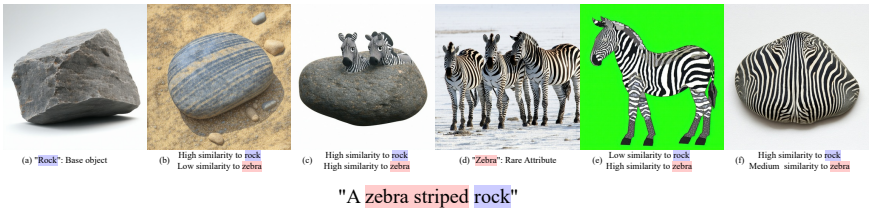

Fig. 3: Visualization of reference images and compositional outcomes for the prompt "A zebra striped rock". The left column shows reference images for the object (rock) and attribute (zebra striped), while the right columns present three common failure cases and one correctly composed result.

tribute. Instead of employing time-consuming VQA-style evaluation [24], we adopt CLIP [29] as the backbone for similarity measurement, providing an efficient and scalable foundation for real-time compositional feedback.

Specifically, for each prompt, we prepare one reference image I_o for the base object and one reference image I_a per attribute.[1] These reference images serve two roles. (1) they define the visual targets used for per-step similarity measurement. (2) they can serve as conditions for multiple image prompts in later stages. During sampling, at each denoising step t, the diffusion backbone produces an intermediate latent $\hat{\boldsymbol{z}}_t$. We estimate the latent $\hat{\mathbf{z}}_t$ at time $t = 0$ based on the flow formulation [23], then decode it into the current estimated image $\hat{\mathbf{x}}_t = \text{Decode}(\hat{\mathbf{z}}_t)$. At each step t, we respectively compute the CLIP-based cosine similarity between $\hat{\mathbf{x}}_t$ and the corresponding reference images I_{a_i} (for attributes) and I_o (for the object) as follows:

$$\begin{aligned} \text{sim}_t^{a_i} &= \cos(f_c(\hat{\mathbf{x}}_t), f_c(I_{a_i})), \\ \text{sim}_t^{o} &= \cos(f_c(\hat{\mathbf{x}}_t), f_c(I_o)), \end{aligned} \tag{1}$$

where f_c denotes the CLIP image encoder. This monitoring process requires only a single decoding operation per step and introduces no modification or fine-tuning to the backbone model.

As shown in Fig. 3, leveraging CLIP-based similarity allows us to effectively distinguish different types of generation outcomes. In Fig. 3(a) and (c), the generated images are dominated by a single aspect, representing common failure modes. In contrast, Fig. 3(b) exhibits high sim_0^o and $\text{sim}_0^{a_1}$ scores, yet the concepts merely co-exist rather than integrate, resulting in an implausible composition. Meanwhile, Fig. 3(d) shows both similarities at moderate, balanced levels, indicating that the attribute is coherently expressed on the object. **These patterns underscore the core challenge: maintaining a proper balance between the rare attribute and the base object is more crucial than simply maximizing both similarity scores.**

By tracking $\text{sim}_t^{a_i}$ and sim_t^o throughout the denoising trajectory, CSM can detect when the generation becomes dominated by one component (*e.g.*, a rare

[1] In practice, each object and attribute reference (for example, *zebra stripped*, *stone*) is obtained by a brief synthesized by the original backbone.

attribute) or when another lags behind (*e.g.*, the object), thereby providing the signal for adaptive guidance adjustment.

3.3 Bidirectional Scale Controller (BSC)

The Compositional Similarity Monitor tells us when the denoising trajectory becomes biased. **Instead of forcing the similarities in Eq. (1) to match some preset numeric targets, which would vary with the prompt, timestep, and sampling noise, we only require that their trajectory remains balanced,** *i.e.*, both the object and attribute similarities stay within a reasonable range and neither collapses nor dominates for steps. When the signals indicate this balance is broken at step t, we adjust the image-prompt pathway.

To this end, we propose *Bidirectional Scale Controller (BSC)* based on IP-Adapter [40], which integrates image prompts into a pre-trained text-to-image diffusion model θ through a decoupled cross-attention mechanism. Given a text prompt c, a reference image I, an image encoder $f_\phi(\cdot)$, and a text encoder $f_\mathcal{E}(\cdot)$, IP-Adapter modifies the intermediate feature propagation of the model θ as:

$$\mathbf{h}_t^{l+1} = F_A^l\big(\mathbf{h}_t^l, f_\mathcal{E}(c)\big) + \sum_{i=1}^{n} s_i \, F_A^l\big(\mathbf{h}_t^l, f_\phi(I_{a_i})\big)\,, \tag{2}$$

where $\mathbf{h}_t^l$ denotes the intermediate feature of latent $\mathbf{z}_t$ at timestep t and layer l, and $F_A^l(\cdot,\cdot)$ represents the attention block at layer l [36]. The first argument provides the queries, while the second provides keys and values. In our setting, we attach one IP-Adapter branch for the base object and each attribute. For a concept with multiple reference images $I_{a_1}, \ldots, I_{a_n}$, we weight the contribution from each reference image by a corresponding scalar s_i and sum them, as shown in the second term, which can also be referred in the left part of Fig. 2.

One key empirical observation is that **IP-Adapter guidance remains effective even when its scale is negative**. Positive scales strengthen the influence of a reference image, while negative scales actively suppress it, as shown in 4. This bidirectional effect enables a simple bias–correction rule: when the generation becomes overly influenced by one component, we assign a negative scale to that component and a positive scale to the weaker one. The result is a reactive "restoring force" that pulls the trajectory back toward balanced composition.

3.4 Feedback Guidance Scheduler (FGS)

Building on the idea of negative-scale correction, we now integrate detection, decision, and correction into a single adaptive guidance pipeline, namely, Feedback Guidance Scheduler (FGS). As shown in Fig. 2, FGS requires no additional training or gradient computation, and adjusts the IP-Adapter scales online according to compositional feedback provided by the CLIP-based detector.

Adaptive Scale Assignment. At each denoising step t, we obtain the current synthesized image $\hat{x}_t$ and compute the similarities sim_t^a and sim_t^o with respect

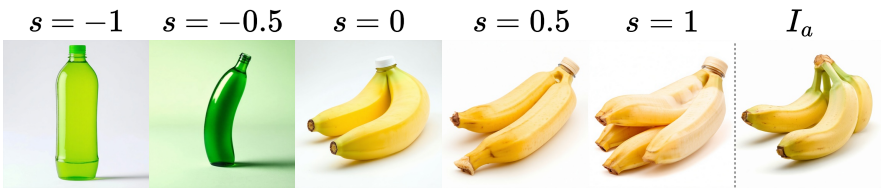

Fig. 4: Illustration of IP-Adapter guidance with different scales for the prompt "A banana shaped bottle". Here, the scalar s controls the guidance of the IP-Adapter for the specific concept "banana shaped " (I_a). In this example, when generating without guidance from I_a ($s = 0$), the output deviates from the "bottle". Interestingly, when setting s to negative values (-0.5 and -1), the concept of "banana shaped" (I_a) is suppressed, leading to a more accurate depiction of the desired concept.

to the rare attribute and the base object, respectively. sim_t^a and sim_t^o are then compared with the similarities from the previous step, sim_{t-1}^a and sim_{t-1}^o to check whether either component is becoming over- or under-represented. If an imbalance is detected, we update the IP-Adapter scales at the next step so that the latent trajectory is nudged back toward a balanced composition. Fig. 5 provides a stepwise visualization of this process.

For clarity, we first describe the policy for a single rare attribute a. Based on the similarity trends between two consecutive steps, we set the IP-Adapter guidance scales for the next step, $\left(s_{t-1}^a, s_{t-1}^o\right)$, as:

$$(s_{t-1}^a, s_{t-1}^o) = \begin{cases} (1,-1), & \text{if } t \geq T_{\text{early}} \text{ and } \text{sim}_t^a < \text{sim}_{t+1}^a \\ (-1,1), & \text{else if } \text{sim}_t^o < \text{sim}_{t+1}^o \\ (0,0), & \text{otherwise.} \end{cases} \tag{3}$$

Here, T_{early} splits the trajectory into an early stage, with both the rare attribute and base object still forming, and a later stage, where the attribute is already stabilized and preserving object identity becomes the priority. When $t \geq T_{\text{early}}$ and the attribute similarity keeps decreasing, indicating that the object becomes dominant, we set $(s_a, s_o) = (1, -1)$ to slightly suppress the object and reinforce the attribute. Conversely, if the object similarity is the one that continues to decrease, we flip the signs to preserve the object instead. When neither similarity exhibits a concerning trend, the scales remain $(0, 0)$, and the IP-Adapter functions as a neutral controller. Detailed comparison rules and the multi-attribute scheduler are provided in Appendix **??**. By continuously monitoring and correcting compositional drift in real-time, this ensures the faithful coexistence of rare and frequent concepts at zero additional training cost.

3.5 Multiple Rare Concepts Merging

Building on our single rare concept framework, we further extend RADIANCE to prompts that contain multiple rare objects with their own attributes. Given a

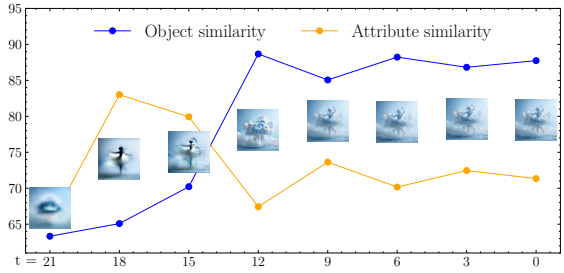

Fig. 5: Step-wise visualization of adaptive guidance. We shows each $\hat{x}_t$ and corresponding object, attribute similarity.

prompt $\mathcal{T} = \{o_1, o_2, \ldots, o_n\}$ with n rare objects, we first generate single-object images I_i for each object o_i. We then reuse them as image prompts to steer a joint multi-object synthesis via IP-Adapter. **A major challenge arises if we simply feed all reference images into the IP-Adapter in parallel**: the model tends to mix up distinct objects. For example, a prompt such as "a horned elephant and a hairy frog" may produce a single object that mixes features of both the elephant and the frog, instead of two clearly separated objects. To mitigate premature fusion of multiple rare objects, we introduce two complementary strategies.

Delayed Adapter Activation (DAA). We disable IP Adapter guidance for the first T_{delay} steps, allowing the base model to form a coherent layout before object-specific prompts refine regions toward their rare concepts.

Layer-Wise Alternating Guidance (LAG). To further reduce early mixing between rare objects, we assign reference images to the diffusion model in a cyclic manner. Let $\{I_1, \ldots, I_n\}$ denote the reference images for n rare objects. At each denoising step t, the hidden state of layer l is updated as:

$$\mathbf{h}_t^{l+1} = F_A^l\Big(\mathbf{h}_t^l, f_{\mathcal{E}}(c)\Big) + F_A^l\Big(\mathbf{h}_t^l, f_\phi(I_{(l \bmod n)+1})\Big), \tag{4}$$

where $f_\phi(I_i)$ encodes the reference image of rare concepts c_i.

By cycling objects across layers, each rare concept is allocated dedicated capacity within distinct functional regions of the network, ensuring that their feature representations contribute independently to the denoising process. This layer-wise allocation effectively preserves distinct object identities while enabling the model to capture their complex spatial and semantic inter-relationships. As visualized in Fig. 6, the synergy between DAA and LAG (formulated in Eq. (4)) provides a robust architectural defense against premature concept fusion and semantic "bleeding". Together, these mechanisms orchestrate a faithful multi-object rare concept synthesis, maintaining rigorous compositional integrity and visual coherence across diverse and challenging prompts.

Fig. 6: Visualization of the effect of DAA and LAG.

Methods	BackBone	Venue	Single					Multi			Overall
			Property	Shape	Texture	Action	Complex	Concat	Relation	Complex	
-	SD 1.5 [32]	CVPR'22	55.0†	38.8†	33.8†	23.1†	36.9†	23.1†	24.4†	36.3†	33.9†
-	SDXL [28]	ICLR'24	60.0†	56.9†	71.3†	47.5†	58.1†	39.4†	35.0†	47.5†	52.0†
-	SD 3.0 [8]	ICML'24	49.4†	76.3†	53.1†	71.9†	65.0†	55.0†	51.2†	70.0†	61.5†
-	FLUX [18]	-	69.4	78.1	52.5	65.6	66.9	63.1	61.3	81.3	67.3
-	SD 3.5 [8]	ICML'24	77.5	80.6	75.0	83.1	80.0	71.9	58.8	80.6	75.9
SynGen [31]	SD 1.4 [32]	NeurIPS'23	61.3	54.4	50.6	45.6	51.2	32.5	39.4	40.0	46.9
InitNO [9]	SD 1.4 [32]	CVPR'24	48.8	46.9	36.9	33.8	55.6	38.8	28.7	29.4	39.9
InitNO [9]	SD 3.5 [8]	CVPR'24	56.9	62.5	51.2	67.5	76.2	56.3	58.1	75.6	63.0
LMD [22]	SD 1.5 [32]	TMLR'24	23.8†	35.6†	27.5†	23.8†	35.6†	33.1†	33.4†	33.1†	30.7†
RPG [38]	SDXL [28]	ICML'24	33.8†	54.4†	66.3†	31.9†	37.5†	21.9†	15.6†	29.4†	36.4†
ELLA [12]	SD 1.5 [32]	arXiv'24	33.1	51.2	65.6	46.9	60.6	41.3	48.8	62.5	51.3
R2F [27]	SD 3.0 [8]	ICLR'25	89.4†	79.4†	81.9†	80.0†	72.5†	70.0†	58.8†	73.8†	75.7†
R2F [27]	SD 3.5 [8]	ICLR'25	<u>90.0</u>	<u>82.5</u>	**89.4**	<u>86.9</u>	<u>80.6</u>	<u>78.1</u>	**64.4**	<u>81.3</u>	<u>81.7</u>
Our	SD 1.5 [32]	-	63.1	55.0	47.5	42.5	37.5	33.1	30.6	43.1	44.0
Our	SD 3.5 [8]	-	**97.5**	**89.4**	**89.4**	**87.5**	**85.6**	**80.0**	<u>63.1</u>	**85.0**	**84.7**

Table 1: Quantitative results on RareBench across various methods. Bold values indicate the best, <u>underlined</u> values the second best, and † denotes numbers from R2F [27]. This comparison including different T2I metohds: T2I backbones, attention-based T2I enhancement, LLM-grounded compositional generation, and prompt-switching rare-concept generation.

4 Experiments

4.1 Setup

Datasets. We evaluate RADIANCE on **RareBench** [27], a benchmark designed for rare concept. RareBench provides 40 prompts for each rare concept type and covers five single-object concept categories. We also evaluate RADIANCE on **T2I-CompBench** [15], which offers a more diverse set of general compositional prompts. In particular, each attribute class (*color*, *shape*, *texture*) contains 300 *multi-object* prompts. For more detailed analysis, we extract the first object from each prompt to construct a corresponding *single-object* subset, which allows us to evaluate under both single-object and multi-object settings.

Implementation Details. We implement RADIANCE using SD 3.5 [8] as the backbone diffusion model and employ publicly available IP-Adapters [40] for compositional guidance. All experiments are conducted with 24 denoising steps during inference, and all other hyperparameters follow the default settings of SD 3.5. In our method, we set $T_{\text{early}} = 15$ and $T_{\text{late}} = 20$ for general cases. Also we set $T_{\text{early}} = 18$ for prompts involving relational descriptions. For CLIP-based similarity evaluation, we adopt the CLIP ViT-B/32 model.

Evaluation Metrics. Following R2F [27], we evaluate the generated images using GPT-4o [1]-based scoring, which measures the compositional consistency between images and prompts. We also conduct a user study to complement the automatic evaluation. For the T2I-CompBench [14], we follow the official setting, which adopts BLIP [20] instead of GPT-4o as the evaluator.

4.2 Quantitative Results

Performance on RareBench. Tab. 1 shows that stronger backbones such as SD 3.5 [8] and FLUX [18] establish higher baselines for rare concept synthesis, confirming the impact of foundational model capacity. While LLM-grounded methods like ELLA [12] or layout-based planners like RPG [38] provide structural cues, they often introduce semantic drift or over-constrain the generation, leading to lower stability compared to their original backbones [22]. In contrast, RADIANCE achieves an overall improvement of approximately 3 points over the current SOTA, R2F [27]. Notably, our method excels in *property* and *shape* categories by maintaining a feedback-driven equilibrium. Unlike R2F, which often sacrifices object identity for attribute expression, RADIANCE preserves structural integrity via its reactive "restoring force". Comparable *texture* results suggest backbone saturation in material binding, limiting potential for further gain.
Multi-Object and Spatial Reasoning. In multi-object scenarios, RADIANCE maintains dominance in *Concat* and *Complex* settings, validating the efficacy of our DAA and LAG modules in mitigating premature concept fusion. However, a slight performance ceiling in *Relation* indicates that while our design effectively rebalances semantic weights, high-order spatial arrangements (e.g., *left of*, *behind*) remain largely dependent on the backbone's intrinsic positional priors established in early denoising stages.
Architecture Decoupling and Generalization. We further validate the portability of RADIANCE by applying it to heterogeneous backbones (SD 1.5 and SD 3.5) without re-tuning. While attention-based methods like InitNO [9] are sensitive to model-specific internal maps which can distort generations during cross-architecture transfer, our framework's reliance on image-level feedback enables more stable transfer. This robustness is further evidenced on T2I-CompBench [15]; unlike R2F, which may neglect base concepts in frequent-prompt regimes, RADIANCE consistently improves compositional accuracy by balancing both base and rare semantic elements.

4.3 Qualitative Comparison

We qualitatively compare RADIANCE with representative diffusion methods and backbones (Fig. 7). Original backbones often fail to generate rare attributes at all (*e.g.*, "a horned elephant" in SD 3.5 and Flux). Other methods may strengthen rare attributes but let them dominate the base object (*e.g.*, "a giant mushroom-shaped building" in R2F and SynGen), or fail to fuse attributes correctly with the underlying geometry (*e.g.*, "an ax-shaped chocolate" in R2F and SD 3.5, where metallic regions appear on the ax head). Methods relying on

Methods	Single			Multi			Overall
	Color	Shape	Texture	Color	Shape	Texture	
SD 1.5	81.8	75.2	74.0	34.6	32.8	38.4	56.1
FLUX	82.4	**82.6**	**80.3**	75.1	58.7	70.4	<u>74.9</u>
SD 3.5	87.5	77.6	73.0	<u>76.2</u>	<u>60.0</u>	<u>71.1</u>	74.2
SynGen	86.0	78.5	<u>79.5</u>	67.1	42.1	61.1	69.1
ELLA	86.0	73.5	76.8	71.9	47.2	62.6	69.7
R2F	**89.1**	76.9	73.2	75.1	51.6	68.3	72.4
Our	<u>88.6</u>	<u>79.6</u>	77.0	**77.2**	**61.8**	**71.3**	**75.9**

Table 2: Quantitative results on T2I-CompBench across various methods. The best values are highlighted in **bold** and the second best values are <u>underlined</u>.

older backbones perform even worse on multi-attribute or multi-object prompts. In contrast, RADIANCE preserves both object identity and rare attributes, producing coherent results even under challenging multi-object compositions.

4.4 User Preference and Inference Latency

We conducted a user study with 52 participants and collected 1,248 valid responses. Each response corresponds to a pair of images generated by our method and a baseline. We evaluate the results using: **Pick** (%), the fraction of total responses that prefer our method, and **Win** (%), the fraction of prompt category in which our method receives $> 50\%$ of the votes (Tab. 3). We group the prompts into two categories: (1) *Basic* (Property, Shape, Texture) and (2) *Advanced* (Action, Complex, Multi-Concat, Multi-Relation, Multi-Complex). Our method shows strong dominance in the *Basic* category, with small gaps between Pick and Win, indicating that users are highly sensitive to object–attribute alignment. For *Advanced* prompts, although our method still leads, the larger gap between Pick and Win suggests that user attention is more dispersed across multiple elements, resulting in less uniform preferences.

We also analyze the overhead of RADIANCE, which adds per-step VAE decoding and CLIP monitoring. In practice, the wall-clock impact is modest (Table 3) because SD 3.5 latency is dominated by Transformer denoising, and we further overlap feedback at step t with denoising at step $t{+}1$ via pipelining. More importantly, deployment should consider *time-to-acceptable-image*: misaligned samples are typically discarded and trigger resampling. By improving compositional alignment, RADIANCE increases the per-sample success rate and reduces the expected resampling budget, improving effective throughput (usable images per unit time) while keeping inference speed competitive.

4.5 Ablation Study

To better understand the contributions of each component in RADIANCE, we conduct ablation experiments for single-object and multi-object generation. For single-object, we evaluate the impact of the Bidirectional Scale Controller (BSC) and the Feedback Guidance Scheduler (FGS) on RareBench single-object

Fig. 7: Qualitative comparison on RareBench. RADIANCE produces more accurate and coherent results compared to baselines.

Comparison	Basic		Advanced		Overall	
	Pick	Win	Pick	Win	Pick	Win
Our vs. SD 3.5	95.5%	100%	61.5%	60%	74.3%	75.0%
Our vs. R2F	84.6%	100%	60.0%	80%	69.2%	87.5%
Our vs. FLUX	79.5%	66.7%	71.5%	100%	74.5%	87.5%

Method	Time (sec)
SD3.5	7.36
R2F	7.38
Ours (w/o parallel)	16.48
Ours (parallel)*	**9.34**

Table 3: User preference and inference latency. Left: user study results. Right: average wall-clock time per image. * denotes our default pipelined execution.

prompts. As shown in Tab. 4, we consider three variants with fixed scales and a **Negative Prompt baseline**. While Negative Prompts provide a static guidance for attribute exclusion, they fail to dynamically adapt to the evolving denoising process, resulting in lower scores compared to our feedback-driven approach. Furthermore, we investigate the sensitivity of the feedback starting timestep T_{early}. As T_{early} aligns with the critical "sweet spot" where the diffusion process transitions from low-frequency structure to high-frequency details, we observe that performance peaks at T_{early}=15 (marked as *). The results show stability across neighboring steps, validating that our design captures the intrinsic generative priors of diffusion models rather than relying on narrow hyperparameter tuning. Overall, these results confirm that dynamic, feedback-driven scaling is essential for faithful single-object generation.

Methods	Single					Overall
	Property	Shape	Texture	Action	Complex	
Fixed s_o	46.3	65.6	71.9	35.0	63.1	56.4
Fixed s_a	66.3	<u>84.4</u>	57.5	55.0	71.9	67.0
Fixed both	83.8	51.2	82.5	78.8	68.8	73.0
w/o BSC	<u>93.8</u>	83.8	<u>88.8</u>	<u>86.3</u>	<u>81.3</u>	<u>86.8</u>
Negative prompt	87.5	84.4	87.5	83.1	74.4	83.4
T_{early}=13	94.4	88.1	88.1	**88.1**	86.9	89.1
T_{early}=15*	**97.5**	**89.4**	**89.4**	87.5	85.6	**89.9**
T_{early}=17	95.6	87.5	86.3	83.8	**87.5**	88.1

Table 4: Ablation results on RareBench. We evaluate the Bidirectional Scale Controller (BSC) against fixed-scale variants and a negative prompt baseline. The impact of the feedback start time T_{early} is also shown, with 15^* indicating our default setting used in the final model.

Methods	Multi			Overall
	Concat	Relation	Complex	
w/o DAA	65.6	40.6	<u>81.9</u>	62.7
w/o LAG	<u>76.2</u>	<u>58.1</u>	78.8	<u>71.0</u>
Our	**80.0**	**63.1**	**85.0**	**76.0**

Table 5: Ablation study for multi-object prompts. Each variant disables one of the two strategies to prevent premature fusion of multiple reference images.

For multi-object prompts, we examine the effect of our two strategies for preventing premature fusion: *Delayed Adapter Activation (DAA)* and *Layer-wise Alternating Guidance (LAG)*. As shown in Tab. 5, disabling either component causes clear performance drops. In *Concat* and *Relation* settings with fewer objects, DAA is more critical for avoiding early merging of concepts. As the number of objects grows and scenes become more complex, LAG brings larger gains by strengthening the representation of each object. Together, these results show that both mechanisms are necessary to maintain object identities and accurate attribute–object bindings in multi-object synthesis.

5 Conclusion

We propose RADIANCE, a training-free, feedback-driven framework designed to resolve the inherent compositional imbalances in rare concept generation. By monitoring similarity and applying bidirectional scaling, we can dynamically balances object and attribute guidance during sampling, which leads to more semantically coherent images across diverse prompts. Extensive experiments on RareBench and T2I-CompBench show that RADIANCE consistently improves compositional alignment and rare attribute fidelity over strong baselines, while preserving the visual quality of the underlying diffusion models.

Limitations and Future Work. While RADIANCE significantly improves the semantic fidelity of rare attribute binding , its current design prioritizes conceptual alignment over global spatial arrangement. Consequently, in scenes requiring precise spatial reasoning among numerous objects, the final layout remains primarily constrained by the backbone's intrinsic geometric priors. We view this as a strategic trade-off: RADIANCE provides a robust semantic foundation that

is a prerequisite for complex composition. Future work will investigate the integration of our closed-loop semantic controller with spatially-aware guidance and cross-object interaction modeling to further enhance compositional controllability in densely populated or spatially complex environments.

Acknowledgements

This work was partially supported by the National Science and Technology Council, Taiwan (Grants: NSTC-112-2221-E-A49-094-MY3 and NSTC-112-2221-E-A49-059-MY3, NSTC-114-2640-E-A49-011).

References

1. Gpt-4o api documentation. https://platform.openai.com/docs/models/gpt-4o, accessed: 2025-11-12
2. Butt, M.A., Wang, K., Vazquez-Corral, J., van de Weijer, J.: Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In: European Conference on Computer Vision. pp. 456–472. Springer (2024)
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
5. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in neural information processing systems **34**, 19822–19835 (2021)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
9. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9380–9389 (2024)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hsiao, T.F., Ruan, B.K., Wu, Y.L., Lin, T.L., Shuai, H.H.: Tf-ti2i: Training-free text-and-image-to-image generation via multi-modal implicit-context learning in text-to-image models. arXiv preprint arXiv:2503.15283 (2025)
12. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
13. Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
14. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=weHBzTLXpH

15. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
16. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. International Conference on Learning Representations (ICLR) (2014). `https://doi.org/10.48550/arXiv.1312.6114`
18. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
19. Laria, H., Gomez-Villa, A., Qin, J., Butt, M.A., Raducanu, B., Vazquez-Corral, J., van de Weijer, J., Wang, K.: Leveraging semantic attribute binding for free-lunch color control in diffusion models. arXiv preprint arXiv:2503.09864 (2025)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
21. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
22. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023)
23. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (2023)
24. Mañas, O., Krojer, B., Agrawal, A.: Improving automatic vqa evaluation using large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4171–4179 (2024)
25. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
26. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
27. Park, D., Kim, S., Moon, T., Kim, M., Lee, K., Cho, J.: Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. In: The Thirteenth International Conference on Learning Representations
28. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
30. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)

31. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. Advances in Neural Information Processing Systems **36**, 3536–3559 (2023)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
33. Ruan, B.K., Ni, Z.X., Huang, B.L., Hsiao, T.F., Shuai, H.H.: Not all thats rare is lost: Causal paths to rare concept synthesis. arXiv preprint arXiv:2505.20808 (2025)
34. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
35. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14398–14409 (2024)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: The Thirteenth International Conference on Learning Representations
38. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: Forty-first International Conference on Machine Learning (2024)
39. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM Comput. Surv. **56**(4) (Nov 2023). https://doi.org/10.1145/3626235, https://doi.org/10.1145/3626235
40. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023)
41. Zarei, A., Rezaei, K., Basu, S., Saberi, M., Moayeri, M., Kattakinda, P., Feizi, S.: Improving compositional attribute binding in text-to-image generative models via enhanced text embeddings. arXiv preprint arXiv:2406.07844 (2024)