Generative Adversarial Diffusion

Supplementary Material

6. Convergence and Stability Analysis

In this section, we present a comprehensive analysis of the convergence behavior and stability of three generative frameworks to validate the effectiveness and generalizability of our proposed method: the baseline diffusion model, the diffusion model with GAN, and our proposed Generative Adversarial Diffusion (GAD). We employed Stable diffusion [38] as the baseline diffusion model, while we adopted Diffusion-GAN [52] as a diffusion model combined with GAN. Diffusion-GAN introduces a separate discriminator into the denoising process of the diffusion model [38] and alternates training between the diffusion model (*i.e.*, the U-Net) and the discriminator. In contrast, our proposed method employs a unified network for both the generator and discriminator, thus reducing the need for alternating updates and potentially shortening training time. For a fair comparison, both Diffusion-GAN and our method use the denoising process of Stable diffusion as the generator and are trained on the same subset of ImageNet [12] dataset.

Following related work [38, 61], we use the Fréchet Inception Distance (FID) [18] as a function of the training steps to measure the convergence speed and training stability. All models were trained under identical conditions (*i.e.*, with the same number of training steps and comparable network parameters) to allow a direct comparison.

Figure 6 shows the sample quality and changes in FID over the 2M training steps. Notably, after around 1.25M steps, Diffusion-GAN's FID begins to diverge, suggesting that its alternating generator-discriminator updates struggle to capture the complex and diverse distribution of the ImageNet dataset. These results highlight the structural limitations of GAN-based approaches, where the generator and discriminator are trained alternately in an unstable manner, often leading to issues such as mode collapse due to imbalanced or oscillatory updates between the two networks.

In contrast, the baseline diffusion method (*i.e.*, Stable diffusion) maintains stable convergence throughout the training, yet converges more slowly and finally achieves a higher FID than the proposed method. The proposed GAD efficiently leverages the U-Net as a unified generator and discriminator, applying adversarial constraints directly within the diffusion process. This key concept enables GAD to combine the stable convergence of diffusion models with high-resolution generation capabilities commonly associated with GANs, without resorting to alternating updates. Furthermore, the faster convergence of the GAD compared to the baselines indicates that the adversarial loss acts as a regularizer, guiding the diffusion process toward improved sample quality in fewer training steps.

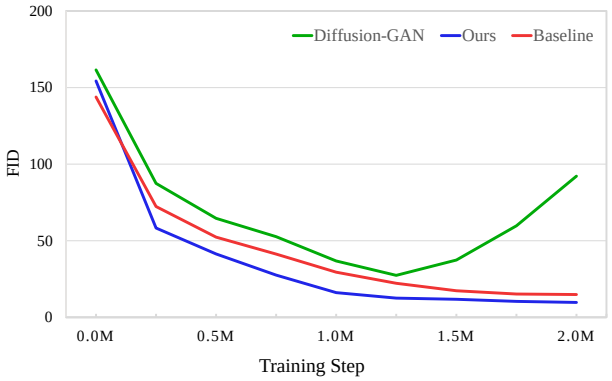


Figure 6. Convergence and stability analysis.

In general, these findings demonstrate the structural stability and robust performance of our proposed GAD framework on a large-scale and challenging dataset. By unifying the generator and discriminator in a single U-Net architecture, our method achieves high-fidelity, high-resolution image generation while maintaining stable training dynamics.

7. Discussion on Adversarial Margin m

In our experiments, we set the margin m by performing inference on the training set every 5 epochs and computing the average error. Furthermore, we explored the effect of varying m by adjusting it by $\pm 10\%$ to assess the sensitivity of our method to changes in the margin value. Our results indicated that these slight modifications did not lead to any statistically significant differences in performance. This outcome confirms that our adaptive margin setting strategy is both robust and well-calibrated, ensuring that the adversarial regularizer maintains a meaningful separation between the predicted noise of real and fake latents throughout training. Consequently, this design choice contributes to the stability and convergence of our GAD.

From a decision-theoretic standpoint, the optimal estimator is the conditional mean under a squared error loss function [22]. In other words, the mean minimizes the expected squared error (*i.e.*, it is the Bayes estimator under squared loss). By choosing the average error computed from the training set as the margin m, we effectively set a threshold that reflects the central tendency of the error distribution observed during training. This adaptive choice is theoretically justified because it provides a balanced and statistically grounded criterion that neither overestimates

Table 4. Comparisons with respect to λ_{adv} values.

λ_{adv}	FID (↓)	CLIP score (↑)
0.05	11.01	0.3142
0.01	**9.71**	**0.3468**
0.005	12.98	0.3195

Table 5. Comparisons with GAN loss variants.

Loss Type	FID (↓)
Vanilla GAN	9.11
Non-saturated GAN	8.57
Wasserstein GAN	7.91
Energy-based GAN (ours)	**7.08**

nor underestimates the typical error. Consequently, this method of selecting m aligns with classical results in estimation theory and further supports our approach to calibrating the adversarial regularizer.

8. Sensitivity Analysis of Parameter λ_{adv}

In this section, we analyze the sensitivity of our method with respect to the adversarial regularization weight λ_{adv} on the text-to-image generation task using the ImageNet [12] dataset. We evaluated generation quality by measuring both the FID [18] and CLIP score [36], following the experimental setup described in Sec. 4.2.1.

We conducted experiments with three different values of λ_{adv}: 0.05, 0.01, and 0.005. In particular, our quantitative results, summarized in Table 4, show that $\lambda_{adv} = 0.01$ gives the best empirical performance on the text-to-image generation task. Furthermore, similar performance trends were observed in experiments on conditional text-to-image generation and 2D-to-3D generation, indicating that the optimal value of λ_{adv} remains consistent across different generation tasks. While additional values of λ_{adv} were also investigated during preliminary experiments, these three settings clearly illustrate that $\lambda_{adv} = 0.01$ is the most suitable choice.

In our framework, the adversarial regularization weight λ_{adv} balances the standard denoising objective L_{ldm} with the adversarial regularizer L_{adv}, which enforces a margin between the predicted noise for real and fake latents. If λ_{adv} is set too high, the adversarial term can dominate the overall objective, forcing the predicted noise to deviate excessively from the true noise in order to maintain the prescribed margin, and thereby degrading the model's denoising performance. Conversely, if λ_{adv} is too low, the adversarial regularizer has minimal effect, and the model behaves almost like a pure diffusion approach, missing out on the benefits of adversarial regularization.

As shown in Table 4, setting $\lambda_{adv} = 0.01$ provides an optimal balance between these extremes, ensuring both stable convergence and strong generative performance for our GAD framework. This balanced setting maintains a meaningful separation between the predicted noise of real and fake latents, ultimately contributing to robust training and improved generation quality.

9. Comparison with GAN Loss Variants

In this section, we analyze the impact of the proposed energy-based adversarial regularizer by replacing it with different GAN loss formulations. Specifically, we compare Vanilla GAN [54], Non-saturating GAN [40], Wasserstein GAN [4], and our Energy-based GAN [63] loss on the text-to-image generation task using the ImageNet [12] dataset. To evaluate generation quality,we measure the FID [18], following the experimental setup described in Sec. 4.2.1.

Since our framework is based on noise prediction in latent space, we adapt scalar-based GAN losses to this setting by computing a similarity signal between the predicted and ground-truth noise vectors. In particular, we apply a sigmoid function for BCE-style objectives, while for Wasserstein GAN we use the cosine similarity between noise vectors as a surrogate signal.

As shown in Table 5, our margin-based energy formulation achieves the best performance. These results demonstrate that the energy-based adversarial term in the proposed method enforces directional and scale-aware separation between predicted and ground-truth noise.

10. Additional Qualitative Comparisons

Due to page limitations of the *main manuscript*, we provide additional qualitative comparisons in this *supplemental material* to further validate the effectiveness of our proposed method across various generative tasks. These results expand upon the main comparisons presented in Sec. 4.2 of the main paper.

10.1. Text-to-Image Generation

Figure 7 presents additional results for the text-to-image generation task, comparing our method with Stable diffusion [38], as discussed in Sec. 4.2.1. The results consistently demonstrate that our method generates images with significantly higher quality and improved detail fidelity.

In particular, in Figs. 7 (a), (b), (e), (g), (l), and (m), the proposed method depicts human body shapes and facial features with more natural and detailed high-frequency information than the Stable diffusion baseline. Moreover, our method demonstrates superior capability in faithfully generating images based on complex textual descriptions. As shown in Figs. 7 (b), (d), and (o), the proposed method

accurately captures and integrates multiple elements from the text prompt. For example, in Fig. 7 (b), the terms "red chair" and "a book" are clearly represented in the generated image, while in Fig. 7 (d), elements such as "three men" and "a shack" are distinctly incorporated. In Fig. 7 (o), the generated image more accurately represents "a dog wearing a red number 6" in contrast to the baseline, where these details are either missing or visually ambiguous.

Additionally, our method consistently produces sharper and more vivid images across various artistic styles, including watercolor paintings, animations, and digital paintings. As demonstrated in Figs. 7 (b), (c), (i), (k), and (n), the proposed method enhances the clarity of the image and the overall visual appeal. Notably, in Fig. 7 (c), the fine details of the "mushroom" and the "soft rays of sunlight filtering through the trees" are captured more vividly compared to the baseline. Similarly, in Fig. 7 (n), the depiction of "snowflakes" falling around the fox is significantly more refined, reinforcing the ability of our model to generate high-quality outputs with a better representation of scene elements.

The improvements become particularly prominent when dealing with complex captions, where multiple objects and intricate relationships need to be rendered accurately. As evidenced in Figs. 7 (c) and (n), our method preserves essential elements such as "mushroom" and "sunlight" in (c) and "snowflakes" in (n), which the baseline often does not represent accurately. These findings confirm that incorporating GAD into the diffusion model not only enhances the overall quality and realism of generated images, but also improves text fidelity by capturing fine details more effectively. The combination of adversarial regularization and diffusion preserves stable denoising performance while simultaneously increasing image sharpness and consistency with textual descriptions.

10.2. Conditional Text-to-Image Generation

Figure 8 extends the comparison with GLIGEN [26] for the conditional text-to-image generation task, complementing the results shown in Sec. 4.2.2. Our proposed method demonstrates improved performance in accurately generating images that align with the given bounding box constraints while maintaining high visual quality.

Specifically, in Fig. 8 (a), the proposed method ensures that the objects corresponding to "towering buildings" and "crowded street" are not only correctly positioned within the bounding boxes but also generated with greater structural fidelity. Similarly, in Fig. 8 (b), our method correctly preserves the spatial arrangement of "a lighthouse" and "a small boat," ensuring that both elements are placed appropriately while maintaining a realistic representation of their forms. These results highlight the improved spatial consistency and semantic alignment achieved by incorporating GAD. Moreover, in Figs. 8 (d) and (e), the proposed method effectively integrates textual descriptions with bounding box constraints, producing images that appear more natural and visually coherent.

While both methods, including the standard GLIGEN, successfully incorporate the caption details, our approach further ensures that the generated content within the bounding boxes is not only spatially accurate but also naturally structured. This is particularly evident in the case of "a person" in Figs. 8 (d), where the proposed method generates an image with realistic proportions and natural posture, unlike the baseline, which often struggles with distorted body structures. These findings indicate that our method significantly enhances the model's ability to simultaneously learn and enforce textual and spatial constraints, thereby improving the realism and usability of conditional text-to-image generation.

Figure 9 presents additional qualitative comparisons between our method and the standard Textual Inversion [16]. Our proposed method demonstrates a stronger ability to learn new concepts efficiently from a limited dataset while maintaining flexibility in combining various caption styles.

In particular, when generating objects with intricate patterns or geometric structures, our method consistently captures fine details without overfitting to the training samples. For instance, in Figs. 9 (d), (e), and (f), the proposed method effectively integrates complex teapot patterns with diverse caption prompts, producing visually distinct images that remain faithful to the given text. In contrast, the standard Textual Inversion method exhibits overfitting, generating nearly identical teapot-like images regardless of whether the captions specify "mug cup" or "T-shirt," as shown in Figs. 9 (d) and (f). This demonstrates that our approach prevents excessive reliance on the initial training patterns and instead generalizes better to diverse styles.

Furthermore, Figs. 9 (j), (k), and (l) further highlight these differences when dealing with structurally intricate objects. Our method effectively learns and applies detailed patterns, such as the intricate textures of an "elephant", allowing it to synthesize a high-quality "dragon" in Fig. 9 (l) that accurately reflects the caption while maintaining the learned visual characteristics. The standard Textual Inversion method, on the other hand, struggles to incorporate the new semantic elements and tends to generate less diversified outputs.

In addition, as shown in Figs. 9 (b), (g), and (h), our method achieves more stable learning, leading to image generations that align more accurately with the intended textual descriptions. Specifically, Fig. 9 (b) demonstrates a precise representation of "green clothes," Fig. 9 (g) correctly places the object "in the forest," and Fig. 9 (h) faithfully illustrates "on stacks of paper," all of which are better captured by our method compared to the standard approach.

These results collectively demonstrate that integrating GAD into conditional diffusion models mitigates overfitting while enhancing the model's ability to capture both textual and visual conditioning. By efficiently balancing generative flexibility and concept preservation, the proposed method produces high-quality images that remain faithful to both the textual prompts and the learned representations.

10.3. 2D-to-3D Generation

For the 2D-to-3D generation task, Figure 10 presents supplementary results that compare our method with SyncDreamer [28], as detailed in Sec. 4.2.3. The results further show that our method not only synthesizes high-quality novel views but also maintains improved multi-view consistency compared to the baseline method.

For instance, in the first row of Fig. 10, our method accurately captures fine details such as bullet casings while preserving the correct gun barrel orientation and quantity. In contrast, the standard SyncDreamer method produces artifacts and struggles to maintain the structural integrity of these components across views.

This performance gap becomes even more apparent when dealing with complex objects. In the second row, which depicts a gauntlet, our approach significantly outperforms the standard SyncDreamer by preserving the detailed shape of the fingers and maintaining material consistency across all views. The baseline method, on the other hand, struggles to maintain structural coherence, often distorting the shape of the hand from certain perspectives.

Moreover, as observed in the third row, the baseline method exhibits difficulties in correctly distinguishing the front and rear of a car, leading to shape inconsistencies in side views. In contrast, the proposed method effectively learns the object's geometric structure, ensuring coherent and well-preserved forms across all generated views.

Furthermore, Figure 11 shows additional qualitative comparisons with Era3D [24], further validating the effectiveness of our approach. The results highlight that our proposed method consistently captures higher-frequency details and preserves structural integrity across multiple views.

For example, in the first row of Fig. 11, our method accurately learns the complex object structure and multi-colored blocks at various spatial positions, ensuring consistent novel view synthesis across all angles. Additionally, our approach effectively refines object boundaries and color transitions during multi-view transformations, reducing artifacts commonly observed in the baseline method.

In particular, in the second column of Fig. 11, our method demonstrates greater robustness in preserving planar objects, such as teapots, ensuring structural consistency between different viewpoints. The baseline method, in contrast, exhibits distortions and inconsistencies when generating certain angles.

Similarly, in the third column of Fig. 11, the proposed method eliminates color boundary artifacts that frequently arise in the baseline method when synthesizing multiple views. This improvement suggests that our adversarial training strategy efficiently mitigates multi-view inconsistencies, leading to more stable and coherent multi-view synthesis.

These additional qualitative results reinforce the findings presented in the main manuscript by further demonstrating the robustness and generalizability of our proposed method across diverse generative tasks.

11. Theoretical Foundations of GAD

In this section, we outline a simplified theoretical argument showing how the proposed loss $L_{gad}(\theta) = L_{ldm}(\theta) + L_{adv}(\theta)$ can act as a regularizer to encourage a margin-based separation between real and fake samples in the latent diffusion framework.

11.1. Problem Setup and Notation

Let θ be the parameters of our shared U-Net $\epsilon_\theta(\cdot, t)$. We define the following two terms:

$$L_{ldm}(\theta) = \mathbb{E}_{z_0,\,t,\,\epsilon}\Big[\big\|\epsilon - \epsilon_\theta(\sqrt{\bar\alpha_t}\,z_0 + \sqrt{1-\bar\alpha_t}\,\epsilon,\, t)\big\|\Big],$$

$$L_{adv}(\theta) = \mathbb{E}_{z_0,\,t,\,\epsilon}\Big[\big[m - \big\|\epsilon_\theta\big(\sqrt{\bar\alpha_t}\,z_0 + \sqrt{1-\bar\alpha_t}\,\epsilon,\, t\big) - \epsilon\big\|\big]_+\Big],$$

where $[\cdot]_+ = \max(0, \cdot)$ and $m > 0$ is a positive margin.

Interpretation.

- $L_{ldm}(\theta)$ is the standard latent diffusion objective, which encourages ϵ_θ to predict the noise ϵ_θ accurately.
- $L_{adv}(\theta)$ imposes a penalty whenever the predicted noise and the true noise are 'too close' (*i.e.*, within distance m). Hence, it effectively enforces the separation between 'real' noisy latents and 'fake' (denoised) latents.

We consider the overall objective

$$L_{gad}(\theta) = L_{ldm}(\theta) + L_{adv}(\theta).$$

Minimizing $L_{gad}(\theta)$ balances the primary denoising objective with a margin-based adversarial constraint.

11.2. Theoretical Analysis

Theorem 1 (Margin-based Separation)**.** *Suppose θ^* is a global minimizer of $L_{gad}(\theta) = L_{ldm}(\theta) + L_{adv}(\theta)$. Let*

$$D_\theta(z) = \big\|\epsilon_\theta(z) - \epsilon\big\|$$

denote the noise-prediction discrepancy at some perturbed latent z (where, for brevity, $z = \sqrt{\bar\alpha_t}\,z_0 + \sqrt{1-\bar\alpha_t}\,\epsilon$). If m

is chosen such that the training data covers noise scales up to m, then under mild continuity assumptions on ϵ_θ,

$$D_{\theta^*}(z) \geq m \quad \textit{for almost all fake latents } z.$$

In other words, the distance between the fake latents' predicted noise and the true noise is encouraged to exceed m whenever it is beneficial to minimize L_{gad}.

Proof. Let θ^* be a global minimizer of $L_{gad}(\theta)$. By definition,

$$L_{gad}(\theta^*) = \min_\theta \Big\{ L_{ldm}(\theta) + L_{adv}(\theta) \Big\}.$$

Suppose, for the sake of contradiction, there exists a set of latent variables $\mathcal{Z}$ (of non-negligible measure) such that for any $z \in \mathcal{Z}$, we have $D_{\theta^*}(z) = \|\epsilon_{\theta^*}(z) - \epsilon\| < m$.

Since $\big[m - D_{\theta^*}(z)\big]_+ > 0$ whenever $D_{\theta^*}(z) < m$, these latents would contribute a strictly positive penalty to $L_{adv}(\theta^*)$.

However, because θ^* is a *global* minimizer, if there is a simple parameter perturbation $\delta\theta$ such that *increases* $D_{\theta^*}(z)$ (so that $D_{\theta^*+\delta\theta}(z) \geq m$) without significantly increasing L_{ldm}, then L_{adv} would decrease (since $[m - D_{\theta^*+\delta\theta}(z)]_+ = 0$ when $D_{\theta^*+\delta\theta}(z) \geq m$), leading to a lower overall cost.

Under mild smoothness assumptions on ϵ_θ and the fact that the margin penalty $[m-x]_+$ is subdifferentiable almost everywhere, it is generally possible for the model to adjust the parameters so as to increase $D_\theta(z)$ to at least m (especially for 'fake' latents that do not strongly affect real data reconstruction). Hence, retaining latents with $D_{\theta^*}(z) < m$ would *not* minimize the objective unless it concurrently and significantly lowered L_{ldm}.

Therefore, in equilibrium, we must have $D_{\theta^*}(z) \geq m$ for almost all relevant fake latents z, otherwise the penalty in L_{adv} could be further reduced. This shows that the margin-based term enforces a separation condition in the global minimizer θ^*. □

Remarks.

- This result indicates that once the model finds parameters θ^* that minimize L_{gad}, any fake latent noise whose predicted noise is too close to the true noise (*i.e.*, within the margin m) will incur a penalty. Consequently, the model is incentivized to maintain or even increase the gap between predicted and true noise, provided that doing so does not significantly deteriorate L_{ldm}.
- In practice, if λ_{adv} (from the formulation $L_{gad} = L_{ldm} + \lambda_{adv}\, L_{adv}$) is appropriately tuned, the model can balance accurate noise prediction with the enforcement of a margin. This balance ensures that the network both reconstructs noise effectively (yielding good diffusion performance) and preserves sufficient separation between fake and real latents.
- Although Theorem 1 guarantees a global separation property, real-world training typically converges to local minima or saddle points. Nevertheless, the margin penalty acts as a stabilizer by providing a practical mechanism to push fake latents away from the real ones, thus improving the convergence of the training and overall stability.

11.3. Mode Collapse and Convergence

Theorem 1 provides a simplified insight: By adding L_{adv} to the overall objective, the model is encouraged to maintain a margin between real and fake latents. This can mitigate mode collapse, as the generator (*i.e.*, U-Net) cannot trivially assign every latent value to a small cluster around the real noise values, *i.e.* doing so would produce a large penalty from L_{adv}.

Moreover, the diffusion objective L_{ldm} is still preserved, ensuring that ϵ_θ learns to denoise effectively. In combination, these two goals can help training converge more stably in practice, even though a complete global-optimality proof (*i.e.*, covering all aspects of GAN and diffusion) is beyond the scope of this simplified analysis.

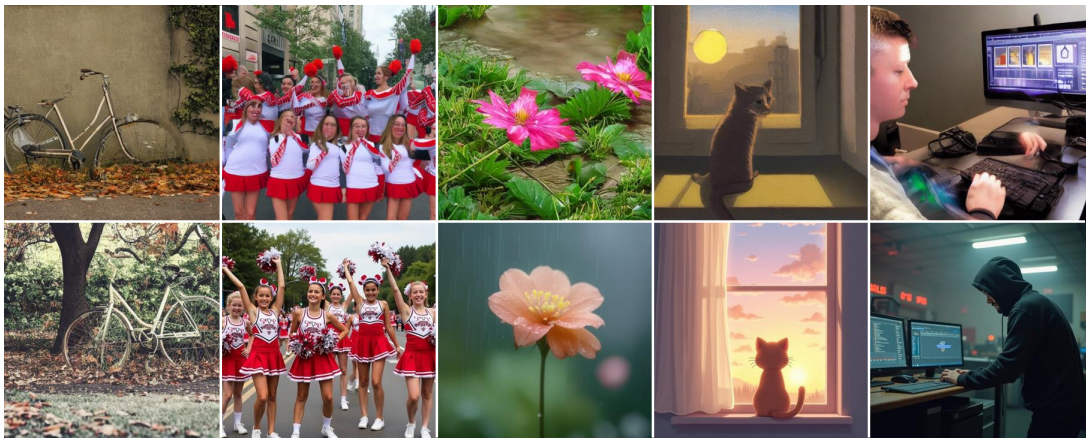
(a) A man and woman sitting on a lake deck.

(b) Woman sitting in a red chair with a dressy purple dress on and a book sitting on the arm rest.

(c) A soft pastel watercolor of a tiny fairy with glowing wings, sitting on a mushroom in a forest, surrounded by soft rays of sunlight filtering through the trees.

(d) Three men are riding on an elephant with a shack in the background and they are wearing hats.

(e) A woman smiles while holding several green peppers at a grocery stand..



(f) A vintage bicycle leaning against a tree in a quiet park, with leaves scattered on the ground.

(g) A group of local cheerleaders in red and white kept our spirits up during the parade.

(h) A single flower blooming under a soft rain in a peaceful garden.

(i) A animation of a cute cat sitting on a windowsill, gazing out at the sunset.

(j) The man in the hooded sweatshirt is working on a computer

(k) A digital painting of whimsical treehouse high in the sky, surrounded by clouds and glowing lights.

(l) Three children, baby in the center, sit in an outdoor swing chair.

(m) A woman wearing a pink sweater and eyeglasses is wiping a wooden table with a pink sponge.

(n) A digital painting of a cute fox sitting in the snow, with snowflakes falling gently around

(o) A dog is wearing a red number 6 and running through a field of green grass while being chased by a bird.

Figure 7. Additional qualitative comparisons with Stable diffusion.

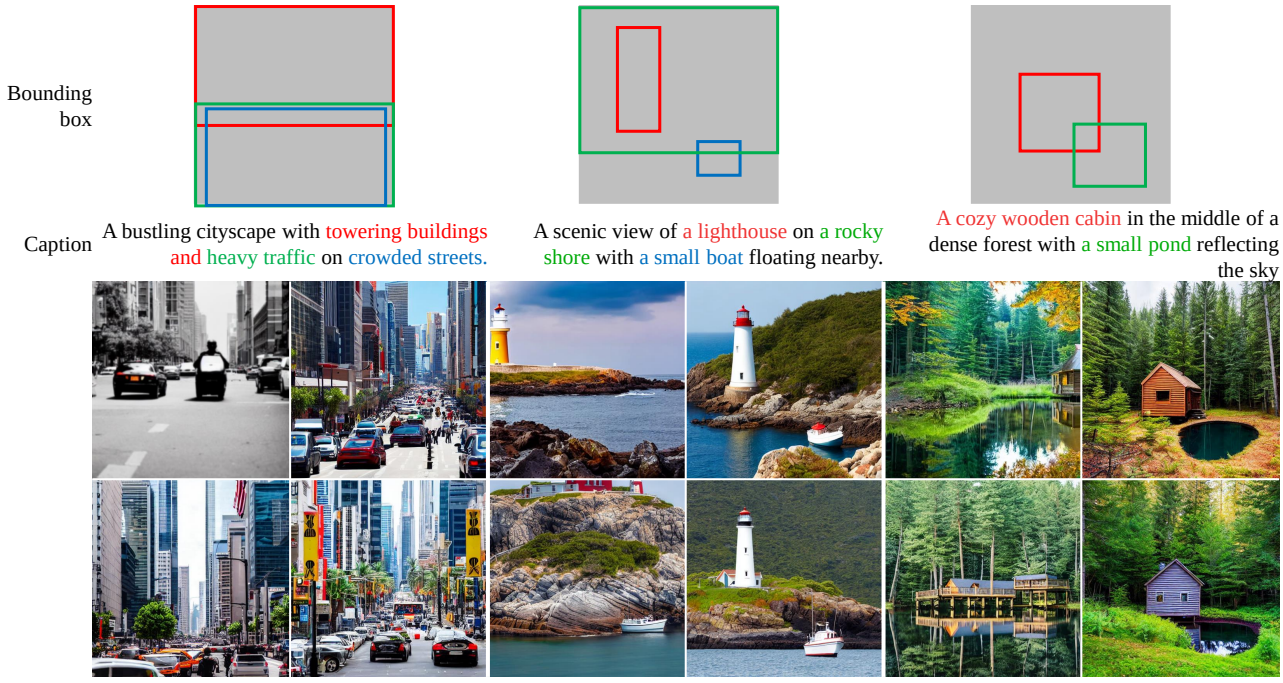
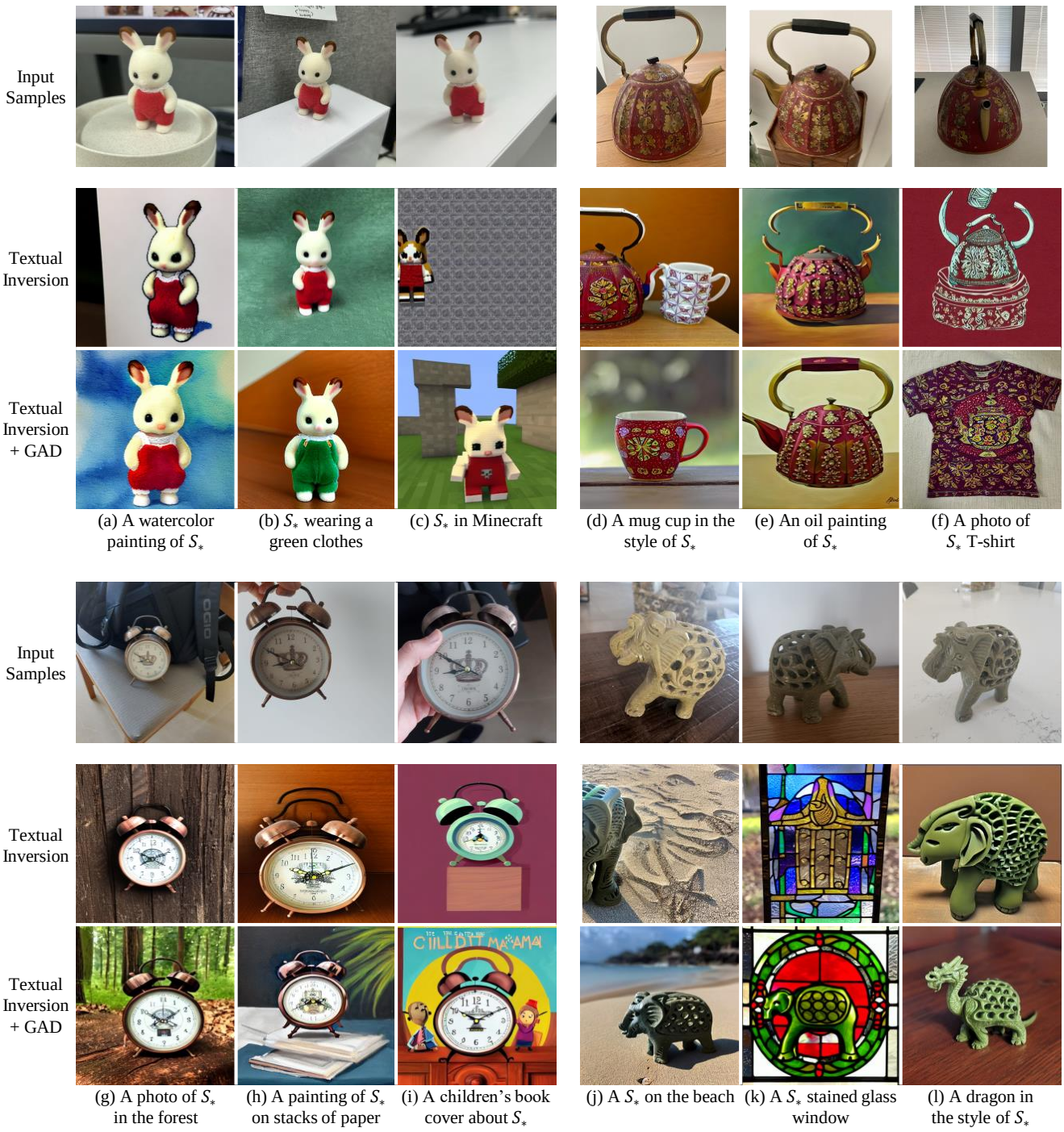


Figure 8. Additional qualitative comparisons with GLIGEN.



(a) A watercolor painting of S_* (b) S_* wearing a green clothes (c) S_* in Minecraft (d) A mug cup in the style of S_* (e) An oil painting of S_* (f) A photo of S_* T-shirt

(g) A photo of S_* in the forest (h) A painting of S_* on stacks of paper (i) A children's book cover about S_* (j) A S_* on the beach (k) A S_* stained glass window (l) A dragon in the style of S_*

Figure 9. Additional qualitative comparisons with Textual Inversion.

Figure 10. Additional qualitative comparisons with SyncDreamer.

Figure 11. Additional qualitative comparisons with Era3D.