

1	<b>Contents</b>	
2	<b>A Additional Material: Media Gallery</b>	<b>2</b>
3	<b>B The Join-Attention in Multi-Modal Diffusion Transformers</b>	<b>2</b>
4	<b>C Analyses Details</b>	<b>3</b>
5	C.1 Setup Details .....	3
6	C.2 Further Discussion for Local Isotropy .....	4
7	C.3 Further Discussion for Global Anisotropy .....	5
8	C.4 Pitfalls of Variance Scale-Up via Semantic Vector Analysis .....	6
9	<b>D Methods Details</b>	<b>7</b>
10	D.1 Implementation Details and GPU Time/Memory Analysis .....	8
11	D.2 Effects of Token Spacing for Local Isotropy .....	8
12	D.3 Effects of Residual Alignment for Mitigating the Side Effects from Token Spacing.	9
13	<b>E Evaluation Details</b>	<b>9</b>
14	E.1 Human Study .....	9
15	E.2 GPT-4o Evaluation .....	9
16	E.3 Evaluation Prompts for Each Benchmarks .....	11
17	<b>F Additional Experiement Results</b>	<b>11</b>
18	F.1 Additional Ablation Studies .....	11
19	F.2 More qualitative results .....	14
20	F.2.1 T2ICompbench .....	14
21	F.2.2 Text-to-Image Generation .....	14
22	F.2.3 Text-to-Video Generation .....	14
23	F.2.4 Text-driven Image Editing .....	14
24	<b>G Limitations and Further Discussions</b>	<b>14</b>
25	G.1 Limitations and Future Works .....	14
26	G.2 Further Discussions .....	14
27	<b>H Qualitative Comparison with GPT-4o</b>	<b>15</b>

## 28 Correction for Clarity

29 To ensure the utmost clarity and prevent any technical misunderstandings of our work, we want  
30 to address a few misleading or misplaced expressions in our main manuscript. We believe these  
31 corrections will help readers interpret our findings precisely. These issues will be comprehensively  
32 addressed and corrected during the revision process.

33 **Main results table notations.** We report quantitative performance for each RareBench category  
34 in our main manuscript; Table 1 lists the eight categories (*Property, Shape, Texture, Action, Single*  
35 *Complex, Concat, Relation, and Multi Complex*) arranged from left to right across the columns.

36 **Image Editing results table notations.** In this Technical Appendices, we present additional abla-  
37 tion studies on image editing in Section F. Before proceeding, we would like to clarify a correction  
38 related to Table 2 in the main manuscript. Upon further review, we identified that certain values  
39 were inadvertently misplaced as shown below: specifically, the blue values were swapped among  
themselves, as were the red values.

Model	CLIP <sub>img</sub>	CLIP <sub>text</sub>	CLIP <sub>dir</sub>	Human	GPT4o
Stable Flow	0.87	0.23	0.08	50.77	62.1
+ Ours	0.80	0.28	0.20	87.69	82.8

40

41 The version of the table presented here resolves this issue. For further details and extended ablation  
42 results, please refer to Section F.

## 43 A Additional Material: Media Gallery

44 As displaying video content frame by frame within the paper offers only limited insight into temporal  
45 coherence and visual quality, we provide a Media Gallery Page featuring the full video outputs from  
46 both the main and additional experiments. This page allows for a more faithful assessment of motion  
47 consistency and prompt alignment. The results can be viewed at: [https://neurips2025-1573.](https://neurips2025-1573.github.io/)  
48 [github.io/](https://neurips2025-1573.github.io/)

## 49 B The Join-Attention in Multi-Modal Diffusion Transformers

50 Multi-modal Diffusion Transformer (MM-DiT) extends Diffusion Transformer (DiT) to the multi-  
51 modal text-to-vision setting, jointly processing textual and visual representations. Conditioning text  
52 embeddings come from CLIP (L/14, G/14) [1] and T5-XXL [2] in Stable Diffusion 3 (SD3) [3],  
53 and from T5-XXL alone in FLUX.1 [4]. The initial text embedding  $e^{\text{init}} \in \mathbb{R}^{V \times d}$ , where  $V$  denotes  
54 the number of text tokens, and the latent-noise embedding  $x^{\text{init}} \in \mathbb{R}^{N \times d}$ , where  $N$  represents the  
55 number of image tokens, are processed through  $B$  joint-attention blocks.

56 For every block  $b \in \{1, 2, \dots, B\}$ , the text condition  $e^{b-1}$  and latent noise  $x^{b-1}$  are updated. These  
57 dual-stream joint-attention blocks (also referred to as multi-modal attention layers or MMATTNs in  
58 some literature [5]) are designed to keep the image and text modalities in separate residual streams  
59 while allowing for interaction. The process within each block can be detailed as follows:

- 60 **1. Adaptive Layer Normalization (AdaLN):** The input image patch representations  $x^{b-1} \in$   
61  $\mathbb{R}^{N \times d}$  and prompt token embeddings  $e^{b-1} \in \mathbb{R}^{V \times d}$  are first processed by adaptive layer  
62 norm (AdaLN) operations. As described by Peebles and Xie [6] for DiTs, these AdaLN  
63 layers are conditioned on the diffusion time-step  $t$  and a global CLIP vector embed-  
64 ding. The AdaLN operation effectively applies a LayerNorm and then modulates the nor-  
65 malized embeddings using learned affine transformation parameters (scales  $\gamma$  and shifts  
66  $\beta$ , though sometimes only scales are used) derived from  $t$  and the CLIP vector. Let  
67  $h_{\text{txt}}^{b-1} = \text{AdaLN}(e^{b-1})$  and  $h_{\text{img}}^{b-1} = \text{AdaLN}(x^{b-1})$  be the modulated embeddings. These  
68  $\gamma, \beta$  parameters (or just  $\gamma$ ) are also used to scale the residual connections later.
- 69 **2. Query, Key, and Value Projections:** Separate learned projection matrices are used for text  
70 and image modalities to generate queries ( $Q$ ), keys ( $K$ ), and values ( $V$ ). These are applied



to the modulated embeddings from the AdaLN step:

$$\begin{aligned} Q_\mu^{b-1} &= h_\mu^{b-1} W_{q,\mu}^b \\ K_\mu^{b-1} &= h_\mu^{b-1} W_{k,\mu}^b \\ V_\mu^{b-1} &= h_\mu^{b-1} W_{v,\mu}^b \end{aligned}$$

for each modality  $\mu \in \{\text{txt}, \text{img}\}$ . Here,  $W_{q,\mu}^b, W_{k,\mu}^b, W_{v,\mu}^b \in \mathbb{R}^{d \times d}$  are the modality-specific projection weights at block  $b$ .

3. **Joint Attention Mechanism:** The core of the block is the attention operation. The attention outputs for text ( $o_e$ ) and image ( $o_x$ ) are computed as:

$$o_e = \left[ \text{softmax} \left( \frac{Q_{\text{txt}}^{b-1} (K_{\text{img}}^{b-1})^\top}{\sqrt{d}} \right); \text{softmax} \left( \frac{Q_{\text{txt}}^{b-1} (K_{\text{txt}}^{b-1})^\top}{\sqrt{d}} \right) \right] V_{\text{txt}}^{b-1} \quad (4)$$

$$o_x = \left[ \text{softmax} \left( \frac{Q_{\text{img}}^{b-1} (K_{\text{txt}}^{b-1})^\top}{\sqrt{d}} \right); \text{softmax} \left( \frac{Q_{\text{img}}^{b-1} (K_{\text{img}}^{b-1})^\top}{\sqrt{d}} \right) \right] V_{\text{img}}^{b-1} \quad (5)$$

where the  $[\cdot; \cdot]$  operation indicates that the attention scores (maps) from cross-modal attention (e.g., text queries attending to image keys) and self-modal attention (e.g., text queries attending to text keys) are combined (e.g., concatenated or summed) before being applied to the values of the query’s own modality.

4. **Output Linear Layer and Residual Connection:** The attention outputs  $o_e$  and  $o_x$  are then passed through another linear projection layer. These projected outputs are then added back to the original input embeddings of the block (before AdaLN, i.e.,  $e^{b-1}$  and  $x^{b-1}$ ), scaled by a factor (e.g.,  $\gamma'_{\text{txt}}, \gamma'_{\text{img}}$ ) derived from the AdaLN modulation parameters, thus completing the residual path for this attention stage:

$$\begin{aligned} e^b &= e^{b-1} + \gamma'_{\text{txt}} \cdot \text{Linear}(o_e) \\ x^b &= x^{b-1} + \gamma'_{\text{img}} \cdot \text{Linear}(o_x) \end{aligned}$$

It’s important to note that a full joint-attention block, similar to standard Transformer blocks, would typically also include a position-wise Feed-Forward Network (FFN) or MLP, itself conditioned via AdaLN and with its own residual connection, for both the text and image streams after the attention mechanism described above. The description provided focuses on the multi-modal attention aspects.

This dynamic embedding evolution distinctly differentiates MM-DiT from traditional UNet-based diffusion models. These joint-attention blocks operate sequentially across  $B$  layers at each timestep for latent noise prediction.

## C Analyses Details

### C.1 Setup Details

As stated in the main paper, we retain the default settings provided by the baseline models [3, 4, 7, 8] and only adjust the  $\sigma$  values derived from our proposed method through ablation studies. In our analysis, we utilize both prompts from the original RareBench [9] and additional prompts generated using GPT-4o [10].

In total, our analysis considers 100 text prompts. 40 prompts come directly from RareBench’s eight benchmark categories (five per category) [9], and the remaining 60 prompts are generated with GPT-4o [10], comprising 30 *rare* prompts, created strictly under RareBench [9]’s rarity guidelines, and 30 *common* prompts added as a bias check to ensure that our analysis is not limited to rare cases. A concise overview of the entire prompt set appears in Fig. 11. While the main manuscript reports results with random seed 42 for reproducibility, following the evaluation protocol of R2F [9], the present analysis probes robustness by sampling additional seeds other than 42, thereby capturing stochastic variability that could influence generation quality.

Due to the intrinsic characteristics of the MM-DiT architecture, the initial text embeddings are consistently fed into the first block of each timestep. Additionally, we observed that embedding patterns across the 24 blocks within each timestep exhibit strong similarities from one timestep to the next. Based on this insight, our analyses primarily focus on the behavior observed at the block level.

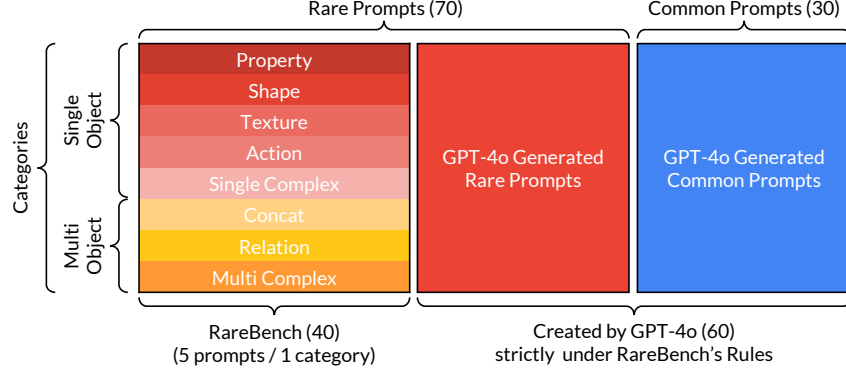


Figure 11: Distribution diagram of prompt samples used in the analysis.

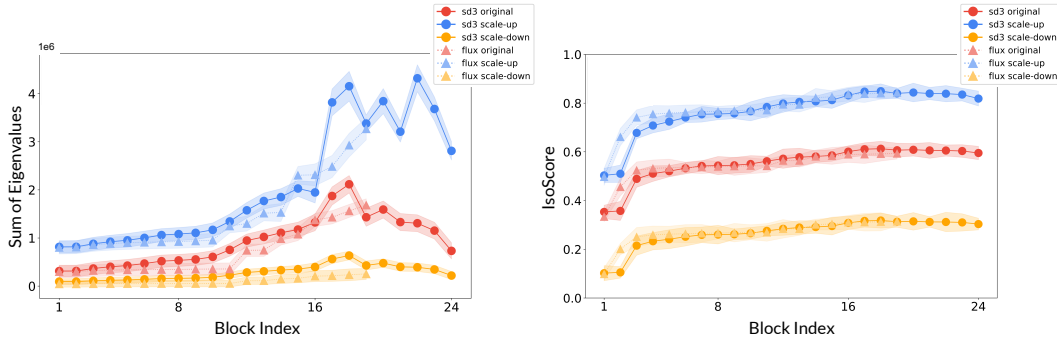


Figure 12: Results of measuring isotropy based on the IsoScore [11] metric with variance scaling. *Left*: A comparative plot of the sum of eigenvalues for the original and variance-scaled data. *Right*: A comparison of IsoScore values before and after variance scaling.

## 111 C.2 Further Discussion for Local Isotropy

112 To rigorously verify the isotropic properties of the text semantic space discussed in Section 2.3, we  
 113 experimentally analyze how variance scale-up impacts isotropy using the IsoScore metric proposed  
 114 by Rudman and Eickhoff [11], noting that *this metric does not locally measure isotropy*. Addition-  
 115 ally, beyond the examples presented in Fig. 3, we provide supplementary examples consisting of  
 116 generated images and their corresponding text self-attention maps to further assess the generality of  
 117 our approach across diverse prompts.

118 **Another metric to measure isotropy: IsoScore [11].** While our analysis in Section 2.3 verified the  
 119 isotropic properties of textual semantics by calculating local isotropy through embedding clustering,  
 120 Rudman and Eickhoff [11] provides a complementary global perspective by introducing *IsoScore*, an  
 121 isotropy metric computed without relying on clustering or cosine similarity-based methods. Specif-  
 122 ically, given a  $V$  text token embeddings represented by the matrix  $e \in \mathbb{R}^{V \times d}$  with sample mean  $\bar{e}$ ,  
 123 IsoScore is calculated as follows:

124 First, embeddings are projected onto their first  $k$  principal components to obtain  $e^{\text{PCA}} \in \mathbb{R}^{V \times k}$ , and  
 125 the covariance diagonal  $\Sigma_D$  is computed:

$$\Sigma_D = \sqrt{k} \cdot \frac{\text{diag}(\text{Cov}(e^{\text{PCA}}))}{\|\text{diag}(\text{Cov}(e^{\text{PCA}}))\|}, \quad (6)$$

126 where  $\|\cdot\|$  is the Euclidean norm. We determine the number of principal components  $k$  for each  
 127 block using the Maximum Distance to Chord (MDC) method [12], as described in Section 3.1. Then,  
 128 isotropy defect  $\delta(e)$  and dimension occupancy  $\varphi(e)$  are defined as:

$$\delta(\mathbf{e}) = \frac{\|\Sigma_D - \mathbf{1}\|}{\sqrt{2(k - \sqrt{k})}}, \quad \text{where } \mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^k \quad (7)$$

$$\varphi(\mathbf{e}) = \frac{(k - \delta(\mathbf{e})^2(k - \sqrt{k}))^2}{k^2}. \quad (8)$$

Finally, the metric is rescaled to range between 0 and 1, defining IsoScore\* as:

$$\xi_{\text{IsoScore}}(\mathbf{e}) = \frac{k \cdot \varphi(\mathbf{e}) - 1}{k - 1}. \quad (9)$$

Here,  $\xi_{\text{IsoScore}} \in [0, 1]$ , with  $\xi_{\text{IsoScore}} = 1$  indicating perfect isotropy (uniform distribution across all principal directions) and values near 0 indicating high anisotropy.

Fig. 12 shows the block-wise IsoScore results computed from MM-DiT using the procedure described above. The results with variance **scale-up** exhibit higher isotropy across blocks compared to the **original** and **scale-down**. This aligns with our findings using other isotropy metrics, confirming that variance scale-up improves the isotropic properties of the text semantic space.

**More analysis results for local isotropy.** In Fig. 3 of Section 2.2, we observed that variance scale-up in MM-DiT facilitates rare semantic emergence for rare prompts—a phenomenon not occurring under the original configuration. Additionally, variance scale-up intensifies activation among text tokens, as reflected in the text self-attention maps, thereby enhancing inter-token relationships. Fig. 18 provides supplementary results supporting this observation, presenting additional examples and corresponding text self-attention maps from SD 3.0 [3] and FLUX [4], beyond those included in the main manuscript.

### C.3 Further Discussion for Global Anisotropy

In the main manuscript we demonstrated, from a global standpoint, that the text-semantic space of MM-DiT exhibits pronounced block-wise anisotropy. Empirically, this anisotropy does not impair the fidelity of the generated images; rather, an overly aggressive attempt to eliminate it produces almost entirely noisy outputs (see Section 2.4 in the main paper). In this section, we present a mathematical derivation of the global anisotropy-reduction procedure adopted in our study, namely the post-processing technique proposed by Mu et al. [13]. We further examine related studies [11, 14–16] to explain, from a global perspective, why text semantic space’s anisotropic representations arise naturally in transformer architectures and how moderate regularization, instead of wholesale suppression, preserves the useful structure of the embedding space.

**Reducing global anisotropy via principal component removal.** Mu et al. [13] identify pronounced *global anisotropy* in standard word-embedding spaces, showing that a shared mean vector and a few high-variance principal directions dominate the geometry. They propose a lightweight post-processing step, mean subtraction followed by removal of the top principal components, that markedly restores isotropy and enhances downstream performance. As outlined earlier in our main manuscript, we apply the same strategy to MM-DiT; the detailed mathematical derivation used in our implementation is presented below.

Let the original embedding matrix be  $\mathbf{e} = [\mathbf{e}_1^\top \dots \mathbf{e}_{|V|}^\top] \in \mathbb{R}^{|V| \times d}$ .

$$\bar{\mathbf{e}} = \frac{1}{|V|} \sum_{i=1}^{|V|} \mathbf{v}_i, \quad \dot{\mathbf{e}}_i = \mathbf{e}_i - \bar{\mathbf{e}}. \quad (10)$$

$$\mathbf{C} = \frac{1}{|V|} \sum_{i=1}^{|V|} \dot{\mathbf{e}}_i \dot{\mathbf{e}}_i^\top = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top, \quad (11)$$

where  $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_d]$  holds eigenvectors in descending order of eigenvalue. Choosing  $D = \lfloor d/100 \rfloor$ , we form  $\mathbf{U}_D = [\mathbf{u}_1, \dots, \mathbf{u}_D]$  and project out the dominant subspace:

$$\mathbf{e}'_i = \dot{\mathbf{e}}_i - \mathbf{U}_D \mathbf{U}_D^\top \dot{\mathbf{e}}_i. \quad (12)$$

Following Mu et al. [13], we use  $D=15$  when  $d=1536$  in Stable Diffusion 3.0 [3]. The resulting block-wise anisotropy scores and corresponding generations are shown in Fig. 4 of our main manuscript, highlighted in yellow.

**Global anisotropy revisited: Extended discussion and related work.** In this section, we further explore the nature of anisotropy within the text semantic space of MM-DiT, building upon long-standing discussions in the Natural Language Processing (NLP) community.

Intuition might suggest that anisotropy, where text embeddings are predominantly distributed across a few key dimensions, could negatively impact downstream tasks. However, our experiments, detailed in Section 2.4 of the main manuscript, reveal its harmlessness concerning the *semantic emergence* effect we proposed. We aim to provide a more comprehensive explanation, drawing upon related works [11, 14–16], of why this characteristic, particularly in transformer models, can be a *natural global phenomenon* that doesn’t necessarily correlate with downstream model performance.

**The Implication of Global Anisotropy in Neural Networks:** Contrary to certain traditional NLP literature [13, 17], global anisotropy in text semantic spaces is not inherently detrimental to downstream task performance. Instead, it is increasingly recognized as an intrinsic property of transformer architectures and a direct consequence of stochastic gradient descent (SGD) optimization [16]. This perspective posits that anisotropy, arising from SGD, facilitates the model’s convergence to flat minima in the loss landscape, which is known to promote superior generalization compared to sharp minima [11, 16]. Furthermore, extensive research indicates that neural networks inherently compress data into lower-dimensional manifolds [11, 14]. These compressed representations often lead to enhanced performance on downstream tasks. Notably, a lower intrinsic dimensionality (ID) in the final layers has been identified as a robust predictor of classification accuracy on test data [15]. This perspective directly contrasts with earlier claims suggesting that *increasing global isotropy (decreasing global anisotropy)* improves model representations, potentially at the cost of performance degradation [11]. Conversely, a growing body of evidence suggests that heightened anisotropy, through this very compression into lower-dimensional manifolds, can lead to performance gains [16].

**Local Isotropy and Generalization:** Several studies [11, 16] corroborate these observations: SGD intrinsically introduces anisotropic noise, which aids in escaping sharp minima and converging to flatter, more generalizable minima. Deep neural networks progressively compress data representations into manifolds of progressively lower intrinsic dimensionality in later layers. This dimensional compression in later layers strongly correlates with improved generalization performance [16]. The low-dimensional representations learned by neural networks are thus believed to prevent overfitting and facilitate generalization by effectively discarding task-irrelevant dimensions. These discrepancies collectively motivated our investigation into *local isotropy* to accurately characterize the text semantic space.

**Our Conclusion:** In conclusion, the evidence strongly suggests two key points. First, global anisotropy naturally emerges from neural network training. Second, the compression of representations into lower intrinsic dimensions is intrinsically linked to enhanced generalization performance. These findings hold true across general neural network learning and classification tasks, as supported by existing literature [11, 14–16]. Lastly, we find that these insights extend naturally to flow- and diffusion-based generative models for text-to-vision tasks that utilize natural language input.

## C.4 Pitfalls of Variance Scale-Up via Semantic Vector Analysis

In Section 2.5 of our main manuscript, we argued that, although variance scale-up demonstrably benefits semantic emergence, it does not invariably produce positive results for every text prompt or random seed. Here, we provide a mathematical derivation that clarifies this limitation and present additional experimental examples beyond Fig. 5 in the main paper.

**Mathematical derivations.** We restate the key definitions from the main paper before delving into new analyses. We first introduce the semantic vector:  $s = e_{\text{cond}} - e_{\emptyset}$ , the difference between the conditional text embedding  $e_{\text{cond}}$  and its unconditional (null) counterpart  $e_{\emptyset}$  used in classifier-free guidance (CFG). Let  $e$  be the original text embedding and  $\hat{e}$  its variance scale-up version. The cosine difference quantifies the effect of variance scale-up on semantic alignment:  $\Delta\gamma = \cos(s, \hat{e}) - \cos(s, e)$ . A positive value of  $\Delta\gamma$  indicates that variance scale-up enhances alignment with the semantic direc-

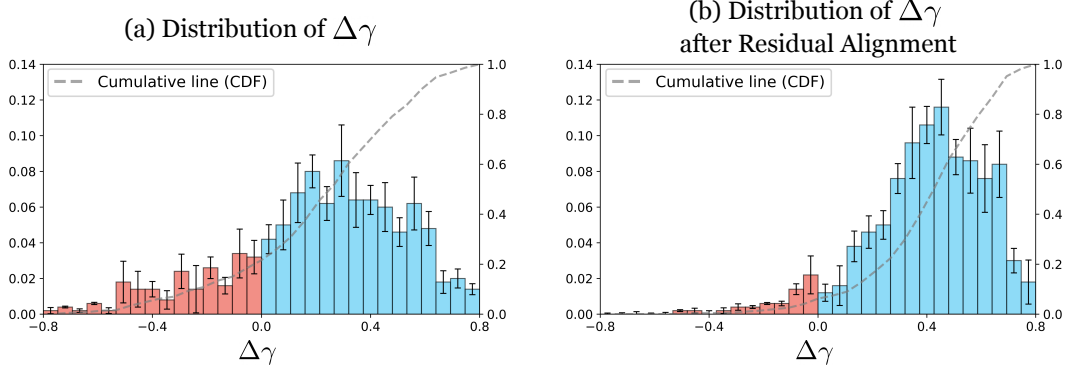


Figure 13: Pitfalls of variance scale-up as seen through  $\Delta\gamma$  in additional examples (random sampled 50 prompts from Fig. 11). (a) Analysis of extended samples: It presents results from 50 randomly selected samples used in our analysis. The distribution illustrates the average frequency value for each sample, with error bars indicating the standard deviation. (b) Impact of *residual alignment* on side effects: We examine the change in  $\Delta\gamma$  values across samples after applying Residual Alignment, demonstrating its effect on mitigating side effects.

tion  $s$ , whereas a negative value signals a degradation.  $\Delta\gamma$  can be rewritten as

$$\Delta\gamma = \cos(s, \hat{e}) - \cos(s, e) = \frac{s \cdot \hat{e}}{\|s\| \|\hat{e}\|} - \frac{s \cdot e}{\|s\| \|e\|} = \frac{s \cdot \hat{e} \cdot \|e\| - s \cdot e \cdot \|\hat{e}\|}{\|s\| \|e\| \|\hat{e}\|}. \quad (13)$$

Because the denominator is strictly positive, the sign of  $\Delta\gamma$  is governed solely by the numerator  $s \cdot \hat{e} \cdot \|e\| - s \cdot e \cdot \|\hat{e}\|$ .

Now let the two comparison vectors share a common component  $\bar{e}$ , the mean of  $e$ , and differ in a residual direction  $u = \frac{e - \bar{e}}{\sqrt{\text{Var}(e)}}$ , writing  $\hat{e} = \bar{e} + \sigma u$  and  $e = \bar{e} + u$  with scaling factor  $\sigma > 0$ .

Substituting these forms gives the inner products  $s \cdot \hat{e} = s \cdot \bar{e} + \sigma(s \cdot u)$  and  $s \cdot e = s \cdot \bar{e} + s \cdot u$ . Keeping the norms symbolic,  $\|\hat{e}\|$  and  $\|e\|$  depend on  $\bar{e}$ ,  $u$ , and  $\sigma$  but remain positive, we obtain

$$\begin{aligned} s \cdot \hat{e} \cdot \|e\| - s \cdot e \cdot \|\hat{e}\| &= (s \cdot \bar{e} + \sigma s \cdot u) \|e\| - (s \cdot \bar{e} + s \cdot u) \|\hat{e}\| \\ &= (s \cdot \bar{e})(\|e\| - \|\hat{e}\|) + (s \cdot u)(\sigma \|e\| - \|\hat{e}\|). \end{aligned} \quad (14)$$

Hence,

$$\text{sign}(\Delta\gamma) = \text{sign}((s \cdot \bar{e})(\|e\| - \|\hat{e}\|) + (s \cdot u)(\sigma \|e\| - \|\hat{e}\|)) \quad (15)$$

In words, whether  $\cos(s, \hat{e})$  exceeds  $\cos(s, e)$  is determined by two weighted gaps: the shared-component inner product  $s \cdot \bar{e}$  multiplied by the difference of norms, and the residual inner product  $s \cdot u$  multiplied by the scaled norm gap. Plugging explicit formulas for the norms, if desired, turns this boxed condition into concrete inequalities for  $\Delta\gamma > 0$  or  $\Delta\gamma < 0$ . This sign rule shows that scaling the residual ( $\sigma$ ) improves alignment ( $\Delta\gamma > 0$ ) when  $s$  is more aligned with the residual direction  $u$ , and worsens it ( $\Delta\gamma < 0$ ) when  $s$  is closer to the mean component  $\bar{e}$ . Consequently, as we mentioned in the main manuscript, even though variance scale-up often helps in a broad sense, it does not automatically improve alignment along the semantic direction—mathematically,  $\Delta\gamma$  can still turn negative, so a semantic gain is not guaranteed.

**More analysis results.** Fig. 13(a) extends the analysis of the  $\Delta\gamma$  distribution for text embeddings, originally presented in Fig. 5 of the main manuscript, to a larger sample set (50 prompts and 100 random seed per prompt). When  $\Delta\gamma < 0$ , the average CLIP score registered at  $\text{CLIP}_{\text{text}} = 0.16 \pm 0.1$ , whereas for  $\Delta\gamma > 0$ , it consistently reached  $\text{CLIP}_{\text{text}} = 0.45 \pm 0.1$ .

## D Methods Details

In this section, we provide an analysis of our proposed method, *Token Spacing* and *Residual Alignment*. Specifically, we investigate how isotropy, previously analyzed in the main paper in Section 2,

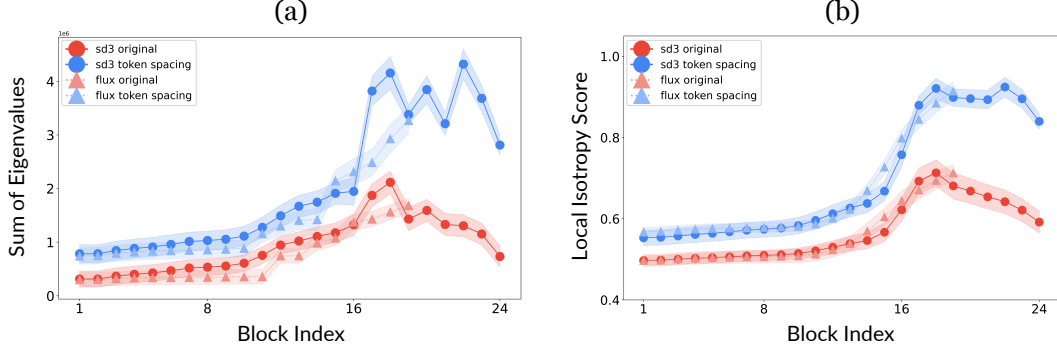


Figure 14: We analyzed the geometric properties of the text embedding space after applying Token Spacing. (a) The left plot shows the sum of eigenvalues for the **original** text embeddings compared to those with **Token Spacing** applied. (b) The right plot illustrates the local isotropy score.

is affected by our proposed approach. We further analyze how Residual Alignment mitigates potential side effects arising from Token Spacing, quantifying its effectiveness in enhancing semantic alignment within the embedding space.

## D.1 Implementation Details and GPU Time/Memory Analysis

Our experiments and results derivation were based on Stable Diffusion 3.0 [3] and FLUX.1 [4], both publicly available on HuggingFace. We reproduced the results of R2F [9] using their publicly accessible code from their official GitHub repository. It is worth noting that R2F exclusively supports SD 3.0 and FLUX-schnell within the MM-DiT model family, and does not provide support for FLUX-dev. Our computational resources included a single NVIDIA 48GB A6000 GPU for general experiments and results, with a single NVIDIA 80GB A100 GPU dedicated to Text-to-Video generation.

Table 4: Evaluation of GPU performance across diverse methods for text-to-image generation with Stable Diffusion 3.0.

Models	SD 3.0	R2F [9]	Ours
Peak Memory (GB)	22.02	38.49	22.17
GPU Time (sec)	$21.30 \pm 1.1$	$78.13 \pm 2.36$	$44.74 \pm 1.9$

Beyond this, for a practical demonstration of our algorithms’ performance, we analyzed their GPU time and memory usage on an NVIDIA 48GB A6000 GPU, conducting all experiments ourselves. We measured the resources needed to generate the complex prompt “a horned lion and a wigged elephant”, which features two rare concepts, consistent with the experimental setup in Park et al. [9]. For each measurement, we performed 100 trials using different random seeds for the same prompt and report the average values and their standard deviations in Table. 4. These results are limited to the diffusion sampling steps.

## D.2 Effects of Token Spacing for Local Isotropy

We analyze our method, Token Spacing, focusing particularly on its influence on local isotropy and inter-token relationships. We demonstrate that increasing the distances between embeddings enhances the local isotropy, effectively improving semantic emergence. Here, we investigate whether Token Spacing yields comparable benefits. To briefly summarize Token Spacing, we first perform Principal Component Analysis (PCA) on the embedding space, decomposing it into principal and residual spaces. Token Spacing then expands the distances between token embeddings within the principal space, effectively amplifying their variance along principal directions. While this approach conceptually parallels a simple variance scaling, our goal is to specifically examine whether Token Spacing also improves local isotropy, an important factor associated with clearer semantic differentiation among tokens.



As shown in Fig. 14, we observe results that closely align with our intended outcomes from the variance scale-up analysis in Fig. 3 of the main manuscript. Token Spacing elevates local isotropy to a level comparable with direct variance scale-up, validating that Token Spacing effectively boosts the isotropic structure within local cluster neighborhoods.

### D.3 Effects of Residual Alignment for Mitigating the Side Effects from Token Spacing

To mitigate potential side effects (see Section 2.5 and C.4) introduced by variance scaling, we propose an additional technique named Residual Alignment. Residual Alignment involves rotating the residual space, previously obtained through PCA, toward a meaningful semantic direction, enhancing semantic coherence while preserving variance-increased embeddings.

Specifically, as depicted in Fig. 13(a) and (b), we compare the values of  $\Delta\gamma$  before and after applying Residual Alignment. Here,  $\Delta\gamma$  is defined as  $\cos(\mathbf{s}, \tilde{\mathbf{e}}) - \cos(\mathbf{s}, \mathbf{e})$ , where  $\tilde{\mathbf{e}}$  denotes the adjusted embedding. More explicitly, before Residual Alignment,  $\tilde{\mathbf{e}}$  represents embeddings modified solely by Token Spacing. After applying Residual Alignment, however,  $\tilde{\mathbf{e}}$  reflects embeddings adjusted by both Token Spacing and Residual Alignment. This distinction allows us to directly evaluate the incremental impact of Residual Alignment on semantic coherence. The results clearly illustrate that Residual Alignment effectively reduces discrepancies, refining semantic alignment and demonstrating its complementary role alongside Token Spacing.

## E Evaluation Details

We follow the same evaluation protocol as RareBench [9] for both Human Study and GPT-4o evaluation [10]. In the following sections, we describe the scoring criteria for Human Study and GPT-4o, as well as provide a brief explanation of the prompts used for each task. Basically, our evaluation methodology followed Park et al. [9]: all evaluations were initially scored on a [1, 2, 3, 4, 5] point scale by GPT-4o and Human, which were then normalized to [0, 25, 50, 75, 100] for reporting the final benchmark performance.

### E.1 Human Study

For the human study, we recruited 23 distinct participants. Participants evaluated the alignment between the given prompt and the generated images using scores ranging from 1 to 5, where a score of 5 indicates a perfect alignment between the image and text, while a score of 1 means that the image fails to capture any aspect of the prompt. We utilized the same prompt categories defined by RareBench, and participants assessed outputs generated by our model and baseline models under identical conditions within each prompt category. To ensure unbiased evaluation, model identities were anonymized, and the presentation order of generated outputs was randomly shuffled within each category and prompt. The detailed scoring guidelines are identical to those used in the GPT-4o evaluation; please refer to the GPT-4o evaluation example provided below.

### E.2 GPT-4o Evaluation

We conducted an identical evaluation procedure using GPT-4o, employing the same scoring guidelines and assessment protocol as in the human evaluation described above. To ensure consistency and reproducibility, we set the random seed to 42, following the exact methodology outlined in RareBench [9].

#### GPT-4o Instruction: RareBench for Text-to-Image

You are my assistant evaluating the correspondence of an image to a given text prompt.  
Focus specifically on:

- Objects in the image and their attributes (e.g., color, shape, texture)
- Spatial layout and positioning
- Action relationships among objects

Evaluate how well the provided image aligns with the following prompt:

#### [PROMPT]

Assign a score from 1 to 5 based on the criteria below:

- 5** : Image perfectly matches the content of the text prompt with no discrepancies.
- 4** : Image portrays most of the content with minor discrepancies.
- 3** : Image depicts some elements, but omits key parts or details.
- 2** : Image depicts few elements, omitting many key parts or details.
- 1** : Image fails to convey the main scope of the text prompt.

Provide your evaluation clearly within 20 words using the format below:

### SCORE: [your score]

### EXPLANATION: [brief justification]

307

#### GPT-4o Instruction: RareBench for Text-to-Video

You are my assistant evaluating the correspondence of a time-lapse video to a given text prompt.

You will receive eight key frames extracted from the video, each filename indicating its position in a sequence.

Focus specifically on:

- Objects in each frame and their attributes (e.g., color, shape, texture)
- Spatial layout and positioning
- Action relationships among objects
- Consistency and appearance/disappearance of elements across frames

Evaluate how well the provided video aligns with the following prompt:

#### [PROMPT]

Assign a score from 1 to 5 based on the criteria below:

- 5** : All frames perfectly match the text prompt with no discrepancies.
- 4** : Most content matches, but minor discrepancies exist in a few frames.
- 3** : Some key elements match, but several important details are missing or incorrect.
- 2** : Only a few prompt elements appear; many key details are absent or wrong.
- 1** : The video largely fails to convey the prompt's content.

Provide your evaluation clearly within 20 words using the format below:

### SCORE: [your score]

### EXPLANATION: [brief justification]

308



### GPT-4o Instruction: RareBench for Text-Driven Image Editing

You are my assistant evaluating the effectiveness of text-driven editing from a reference image (first) to an edited image (second), guided by the following text prompt:

#### [PROMPT]

Focus specifically on:

- Changes in objects and their attributes (e.g., color, shape, texture)
- Adjustments in spatial layout and positioning
- Modifications in action relationships among objects

Evaluate how effectively the edited image reflects the intended transformation described by the prompt compared to the reference image.

Assign a score from 1 to 5 based on the criteria below:

- 5** : Edited image perfectly matches the intended transformation described by the prompt.
- 4** : Edited image effectively conveys the transformation with minor discrepancies.
- 3** : Edited image captures some intended transformations but misses key details.
- 2** : Edited image reflects few intended changes, omitting many key transformations.
- 1** : Edited image fails to convey the intended transformation from the reference image.

Provide your evaluation clearly within 20 words using the format below:

### SCORE: [your score]

### EXPLANATION: [brief justification]

309

### 310 E.3 Evaluation Prompts for Each Benchmarks

311 **RareBench.** We primarily utilize RareBench [9] to evaluate rare prompts. RareBench comprises  
312 prompts featuring single objects across five categories: property, shape, texture, action, and complex.  
313 Multi-object prompts are categorized into concat, relation, and complex. Each category contains 40  
314 diverse prompts, totaling a comprehensive set for evaluation.

315 **T2I-Compbench.** To assess performance on more common prompts, we also evaluate using T2I-  
316 CompBench [18]. T2I-CompBench offers a holistic evaluation framework, encompassing six cate-  
317 gories: color, shape, texture, spatial relationships, non-spatial relationships, and complex composi-  
318 tions. Each category provides tailored prompts, such as “a blue bench and a green cake” for the color  
319 category. For evaluating attributes like color, shape, and texture, we employ BLIP. Spatial relation-  
320 ships are assessed using UniDet for object detection, while non-spatial relationships are evaluated  
321 with CLIP. Complex compositions are analyzed using the 3-in-1 evaluation method proposed by  
322 T2I-CompBench.

323 **GenEval.** We also evaluated with GenEval [19], which is an object-focused framework designed to  
324 evaluate compositional image properties, including object co-occurrence, position, count, and color.  
325 It leverages object detection models to verify the presence and attributes of objects in generated im-  
326 ages, facilitating fine-grained, instance-level analysis. GenEval’s evaluation pipeline includes tasks  
327 such as single object recognition, two-object co-occurrence, counting, color classification, position  
328 assessment, and attribute binding. This comprehensive approach allows for detailed evaluation of  
329 text-to-image models’ capabilities in generating semantically accurate and compositionally coher-  
330 ent images.

## 331 F Additional Experiment Results

### 332 F.1 Additional Ablation Studies

333 **Experiments for extended  $\sigma$  range on Text-to-Vision.** Fig. 15 presents an extended ablation study,  
334 building upon Fig. 10 of the main manuscript, to thoroughly investigate the impact of the hyperpa-  
335 rameter  $\sigma$  on both Text-to-Image and Text-to-Video tasks. In our method, TORa,  $\sigma$  is critically

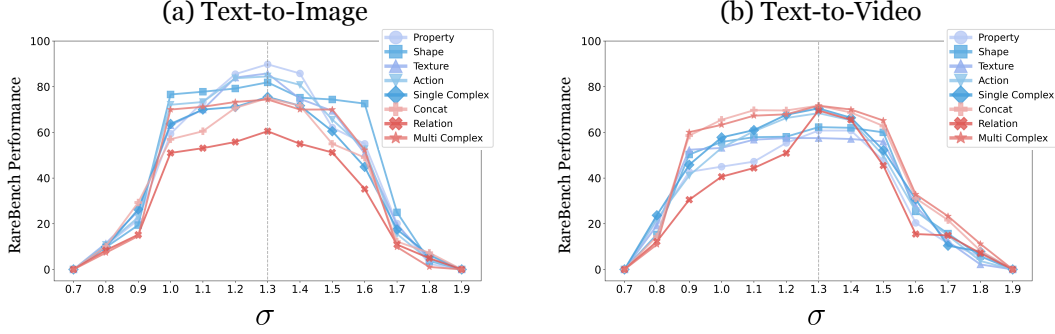


Figure 15: Ablation study investigating the impact of varying the hyperparameter  $\sigma$  on our method’s performance across the RareBench [9] benchmark. (a) Results plot depicting Text-to-Image performance for each category in RareBench, generated using the SD 3.0 [3]. (b) Corresponding results plot for Text-to-Video performance across RareBench categories, utilizing the CogVideoX-5B [7]. In both tasks,  $\sigma = 1.3$  was consistently identified as the optimal hyperparameter setting.

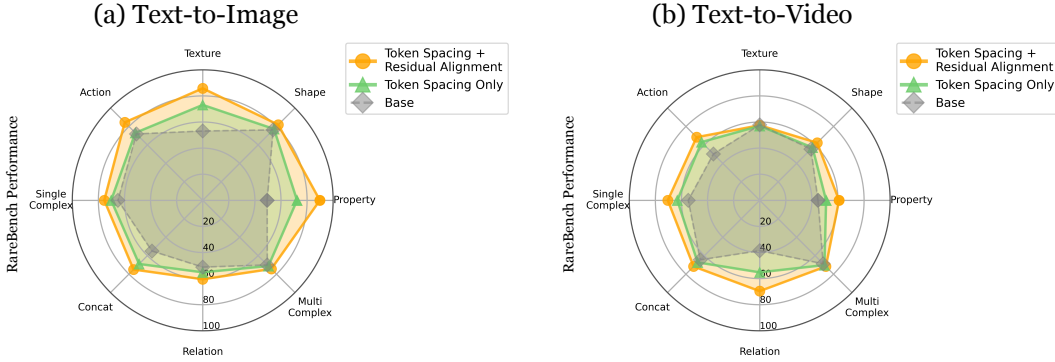


Figure 16: Ablation studies investigating the impact of Residual Alignment within our proposed method, TORA, evaluated on RareBench [9]. (a) For the Text-to-Image task, incorporating Residual Alignment consistently delivered superior performance across all categories, as measured on the SD 3.0 [3]. (b) Similarly, the Text-to-Video task also demonstrated the best quantitative results when Residual Alignment was utilized, with measurements taken on the CogVideoX-5B [7].

utilized for Token Spacing. Our experiments consistently demonstrate that optimal performance for both tasks is achieved at  $\sigma = 1.3$ . A notable observation is the significant performance degradation when  $\sigma < 1.0$ . Specifically, for  $\sigma \leq 0.7$  or  $\sigma \geq 1.9$ , the generated outputs consistently exhibit characteristics close to pure noise or bear no relevance to the input prompt whatsoever, resulting in the lowest possible score of 0.0. While our method exhibits some sensitivity to the value of  $\sigma$ , it is crucial to highlight that within the range of  $1.0 < \sigma \leq 1.5$ , our model consistently outperforms the baseline across all evaluated categories.

**Effects of Residual Alignment on Text-to-Vision.** Fig. 16 presents additional quantitative results demonstrating the impact of residual alignment on our method, building upon the insights from Fig. 10. As evidenced in Fig. 16(b), the integration of residual alignment significantly enhances performance in Text-to-Video generation when combined with token spacing within our framework. This outcome demonstrates that employing only token spacing in our method, TORA, does not achieve optimal performance due to the side effects we previously identified in Section 2.5 of the main paper and Appendix C.4. Consequently, this quantitative ablation study further validates that our proposed residual alignment effectively mitigates these undesirable symptoms.

**Ablations on Text-Driven Image Editing.** Table 5 presents an extended ablation study, building upon Table 2 from the main manuscript, to comprehensively examine the influence of the hyperparameter  $\sigma$  on text-driven image editing performance. Consistent with the findings reported in the main manuscript, the  $\text{CLIP}_{\text{img}}$  scores generally show lower values compared to the baseline across various settings. However, our method achieves the highest directional alignment, measured

Table 5: CLIP similarity scores for image, text, and directional alignment across different  $\sigma$  values of Stable Flow [8] + Our approach on Text-Driven Image Editing. The analysis also demonstrates the impact of residual alignment in our methods. The values highlighted in yellow are taken from our main paper, and those marked in blue indicate the highest-performing results per metric.

Experiments	CLIP <sub>img</sub> ↑	CLIP <sub>text</sub> ↑	CLIP <sub>dir</sub> ↑	GPT-4o↑
<i>Ablation Studies for Various <math>\sigma</math> (w/ residual alignment)</i>				
0.7	0.77	0.23	0.13	55.7
0.8	0.82	0.25	0.14	68.4
0.9	0.82	0.25	0.13	72.9
1.0	0.83	0.25	0.14	73.4
1.1	0.82	0.28	0.17	77.2
1.2	0.81	0.28	0.18	79.3
1.3	0.80	0.28	0.20	82.8
1.4	0.81	0.28	0.20	84.5
1.5	0.82	0.29	0.18	87.2
1.6	0.81	0.29	0.18	86.1
1.7	0.82	0.29	0.16	81.4
1.8	0.81	0.29	0.17	70.6
1.9	0.80	0.29	0.17	68.4
<i>Ablation Studies for Residual Alignment (<math>\sigma = 1.3</math>)</i>				
w/o residual alignment	0.83	0.25	0.14	68.4

by CLIP<sub>dir</sub>, at  $\sigma = 1.3$  and  $\sigma = 1.4$ . Additionally, we observe a clear trend where increasing  $\sigma$  consistently improves CLIP<sub>text</sub> and CLIP<sub>img</sub> performance, whereas significantly lower values, such as  $\sigma = 0.7$ , result in substantial performance degradation. For GPT-4o score, performance rises when the scale factor  $\sigma$  lies between 1.3 and 1.6; pushing  $\sigma$  below 1.3 or above 1.6 consistently degrades performance, a trend that matches what we observe on other diverse tasks.

Looking at these results, what we find particularly noteworthy is that, given that Stable Flow [8] only controls a subset of layers rather than all layers simultaneously, the variations in performance relative to baseline methods appear limited. Nevertheless, applying either Token Spacing or Residual Alignment methods improves the baseline performance. As previously discussed, choosing an appropriate  $\sigma$  value remains crucial for optimal model performance, and it’s important to note that while  $\sigma = 1.3$  demonstrates strong results overall, it does not universally produce superior performance across all metrics.

**Evaluating robustness across random seeds.** To check if our method consistently performs well no matter the random seeds, we tested its strength on 20 prompts from RareBench [9]. We used 50 different random seeds for each prompt and then took the average score. As shown in Fig. 17, our approach consistently bolsters the baseline’s reliability. Moreover, when integrated with other methods [9], we observe a general uplift in benchmark performance.

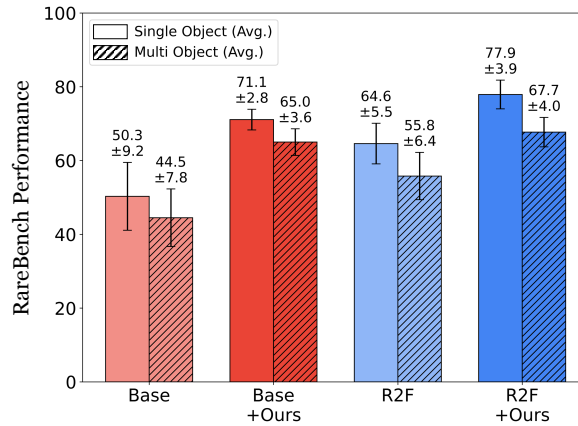


Figure 17: Quantitative results evaluating robustness across random seeds on RareBench.

## F.2 More qualitative results

In this section, we present further qualitative results. These include insights into compositional alignment, alongside Text-to-Image generation, Text-to-Video generation, and Text-Driven Image Editing with rare prompts.

*Note.* Full-page figures are placed at the bottom of the document.

### F.2.1 T2ICompbench

Qualitative results in Fig. 19-20 demonstrate how our method, TORA, influences the model’s compositional alignments in Text-to-Image generation.

### F.2.2 Text-to-Image Generation

Supplementary Text-to-Image generation results for rare prompts are shown in Fig 21-29.

### F.2.3 Text-to-Video Generation

We present further Text-to-Video generation outcomes for rare prompts in Fig. 30-34.

### F.2.4 Text-driven Image Editing

Fig. 35 and 36 illustrate additional Text-Driven Image Editing results for rare prompts.

## G Limitations and Further Discussions

### G.1 Limitations and Future Works

*Architectural Scope:* Our technique is presently tailored to generative models with joint text–image self-attention. Extending it to U-Net backbones [20, 21] and other emerging diffusion variants is a natural next step as the architectural landscape evolves. Moreover, as Diffusion Transformers are an actively researched area in generative modeling, similar to REPA [22], it is essential to further explore our method’s applicability to these emerging architectures.

*Prompt Length:* We have not yet stress-tested extremely long prompts [23]. A systematic evaluation of this regime will clarify the method’s conditioning limits. Investigating such cases would provide further insights into prompt robustness and conditioning boundaries.

*Broader Applicability for Training Phase with TORA:* Our method was developed as a training-free approach. It would be interesting to investigate how its underlying principles might transfer to the training process itself. Extending our approach to diverse learning settings remains an important direction for future work.

*Deeper Theoretical Insights for TORA:* Finally, while our empirical findings reveal the role of isotropy and anisotropy in semantic representations, exploring deeper theoretical insights into why variance scaling is effective could more concretely explain these observations. Such insights might also bring clarity to the phenomenon of *semantic emergence*, referring to how meaningful semantic properties arise through the interplay of these representational characteristics, offering an exciting avenue for future exploration.

### G.2 Further Discussions

Our investigation centered on the *semantic emergence* within text embeddings in vision generative models, particularly MM-DiT. We found that this phenomenon, where intrinsic meaning naturally surfaces within the model, can be effectively induced through a relatively simple yet potent technique: *variance scale-up*. This effect, we suggest, can be explained by the established properties of isotropy and anisotropy as discussed in natural language processing research.

The significance of this *semantic emergence* finding lies in its potential to facilitate successful textual semantic and often elusive compositional alignment internally within the model’s embedding space, achieved through a simple yet effective intervention. This is accomplished without the need for

external modules, such as large language models (LLMs), to forcibly inject semantic alignment. This inherent capability appears to offer broad generalizability, allowing for seamless integration with other methods. Moreover, its applicability is not confined to a single output data type or task, extending to a wide array of text-to-vision tasks that leverage natural language input. We believe this work thus presents an exciting avenue for elevating the intrinsic capabilities and potentially pushing the upper bounds of performance for MM-DiT architectures, a vibrant area of research in modern generative modeling.

## H Qualitative Comparison with GPT-4o

As GPT-4o [10] currently represents the state-of-the-art in generative models, we evaluate our method in direct comparison to it. Although GPT-4o has recently demonstrated remarkable performance across a variety of generative tasks, its closed-source nature limits direct reproducibility and integration. Our method is built entirely on open-source components, yet achieves results that are qualitatively comparable to GPT-4o across a wide range of prompts, as shown in Fig. 37-40. This highlights the potential of open models to narrow the performance gap while maintaining accessibility and transparency. We also find that in certain instances where prompts require precise numerical understanding, such as “A four armed *ninja*” in Fig. 40, GPT-4o tends to misrepresent the intended structure. In contrast, our method accurately depicts the number of arms, highlighting improved semantic fidelity in these challenging cases.

**Note.** Despite being developed entirely with open-source components, our method achieves performance that is *comparable to GPT-4o*, the current state-of-the-art in text-to-image generation. Unlike GPT-4o, which often produces visually similar outputs for a given prompt, our method exhibits greater generative diversity across different random seeds, showing its robustness and flexibility. Furthermore, its plug-and-play design allows seamless integration into various generative pipelines, enabling easy experimentation and broader applicability. We believe this accessibility positions our method as a practical and versatile contribution toward the advancement of future generative modeling research.

434



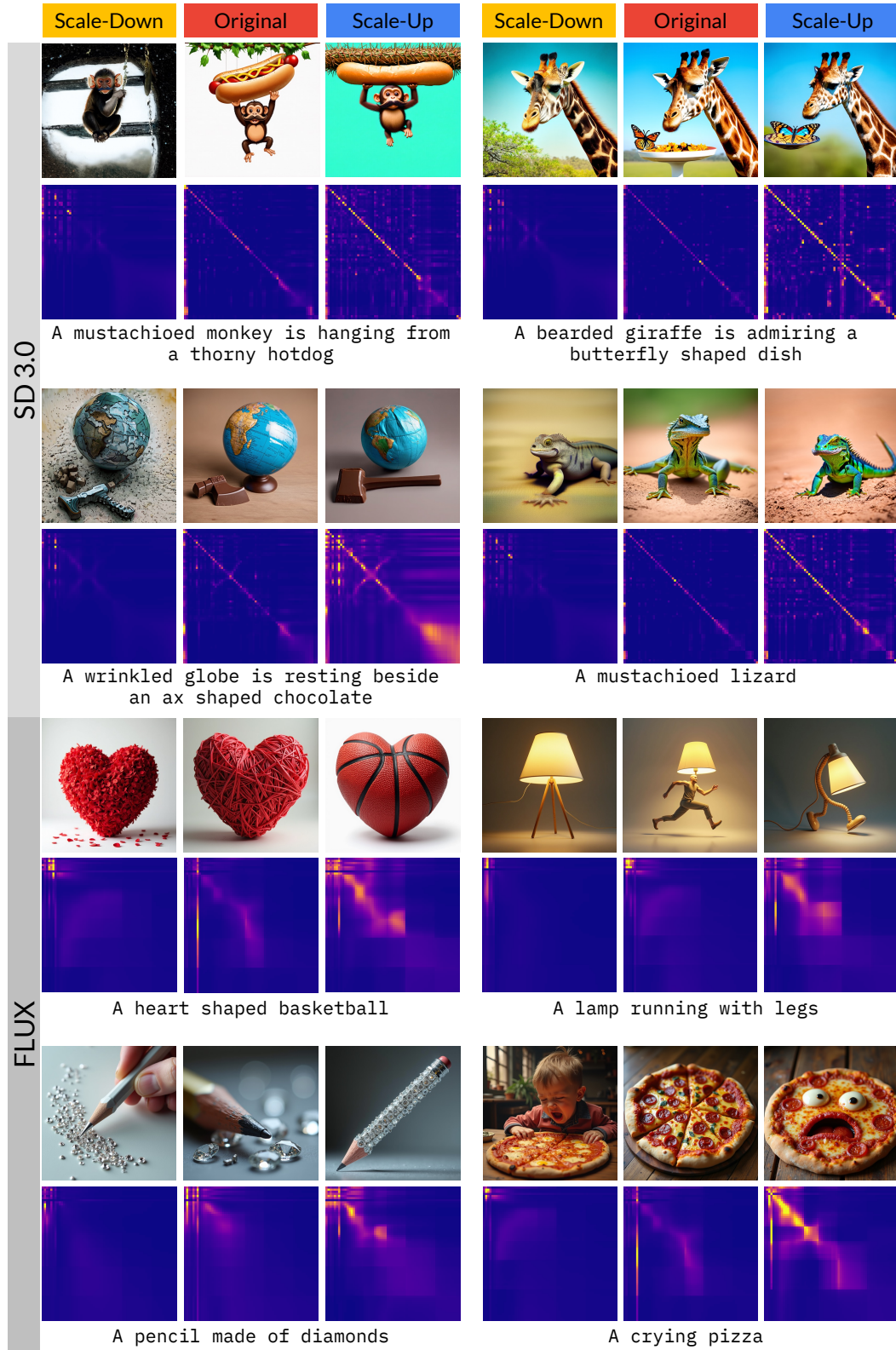


Figure 18: Further results illustrating the influence of variance scaling (Scale-Down and Scale-Up) on text embeddings within the original configuration, examining its effects on both generated images and text-to-text self-attention maps.

Left: Original (SD 3.0)  
Right: Original + Ours



A woman behind a man



A computer in front of a cow



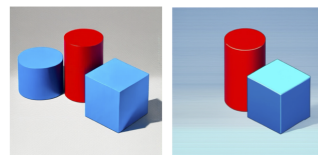
A giraffe behind a sofa



A sheep in front of a cat



A train behind a chair



A red cylinder and a blue cube



A big balloon and a small marble



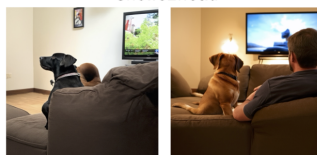
A big bathtub and a square showerhead



A brown snake and a green log



A black gold and white vase sitting on a counter



A dog is sitting on a couch and watching TV with its owner



Four trucks



Six bowls



Two plates



Eight guitars



The fluffy marshmallow melted over the warm cocoa and the crunchy graham cracker



The fuzzy sphere was nestled between the spiky cube and the smooth cylinder



A long rectangular table and a small circular vase were placed in the center of the room

Figure 19: Qualitative comparisons of text-to-image compositional alignments: baseline (SD 3.0) vs. baseline + our method.

Left: Original (FLUX-schnell)  
Right: Original + Ours

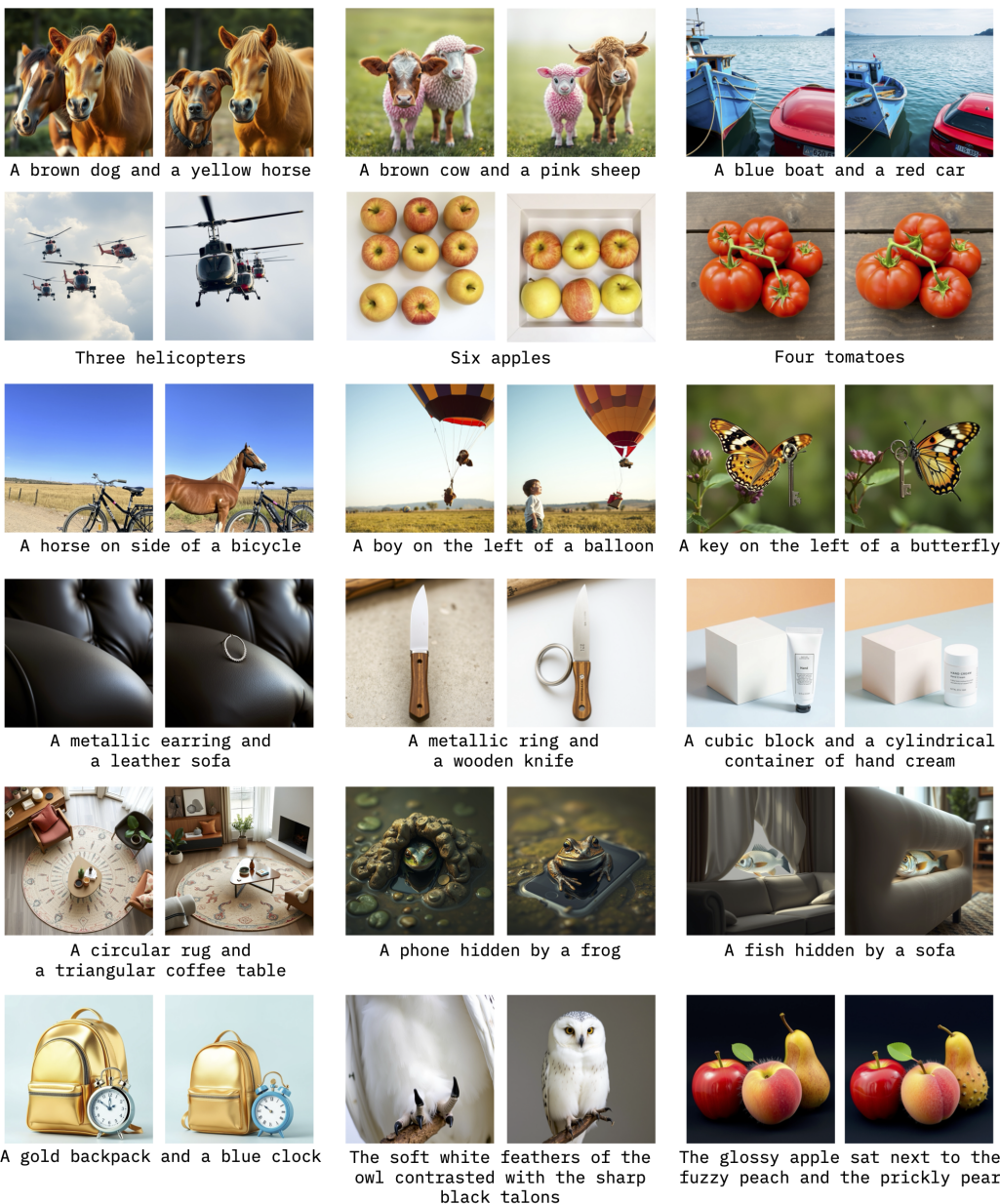


Figure 20: Qualitative comparisons of text-to-image compositional alignments: baseline (FLUX-schnell) vs. baseline + our method.





Figure 21: Further Text-to-Image generation comparisons for rare prompts.

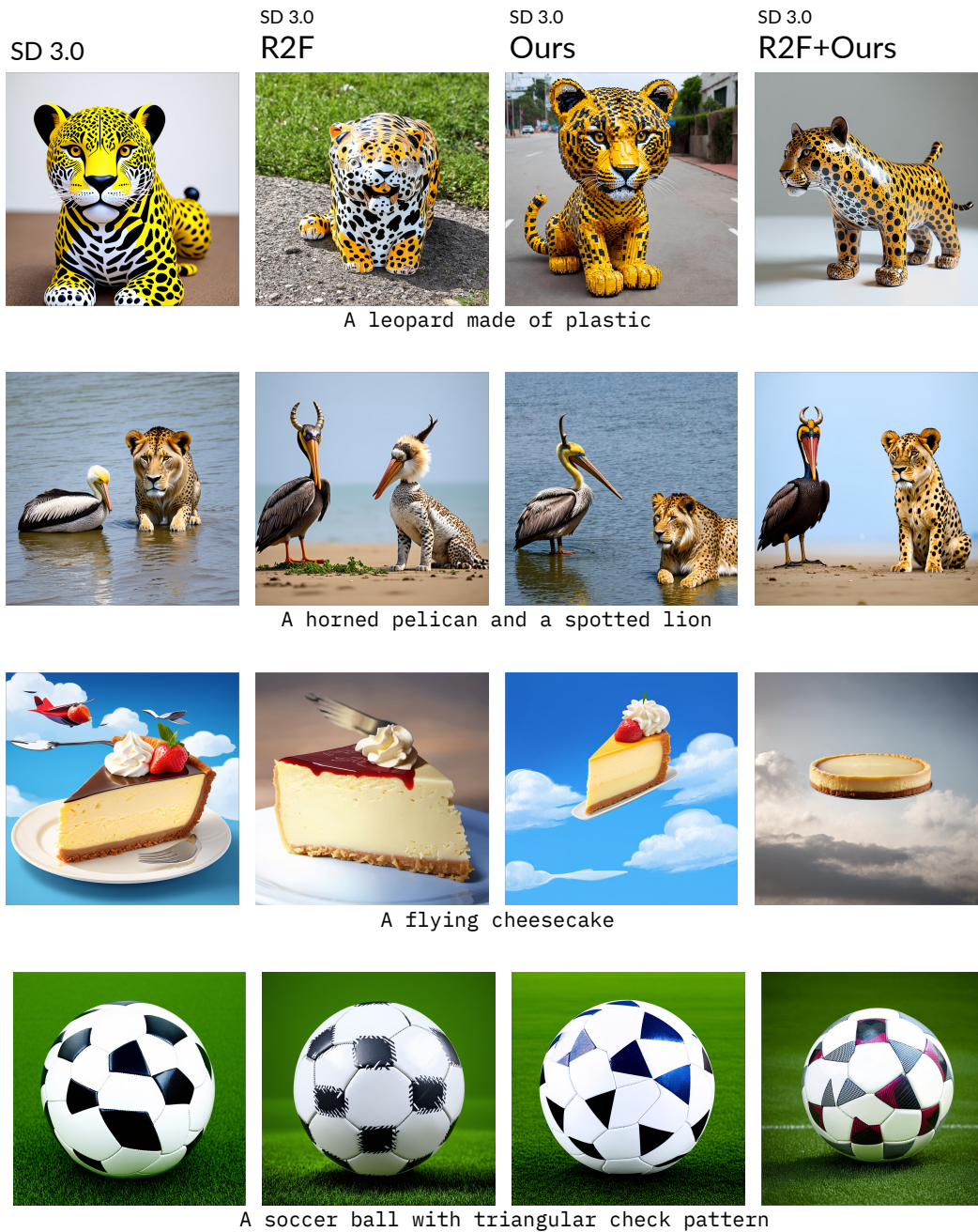


Figure 22: Further Text-to-Image generation comparisons for rare prompts.





Figure 23: Further Text-to-Image generation comparisons for rare prompts.

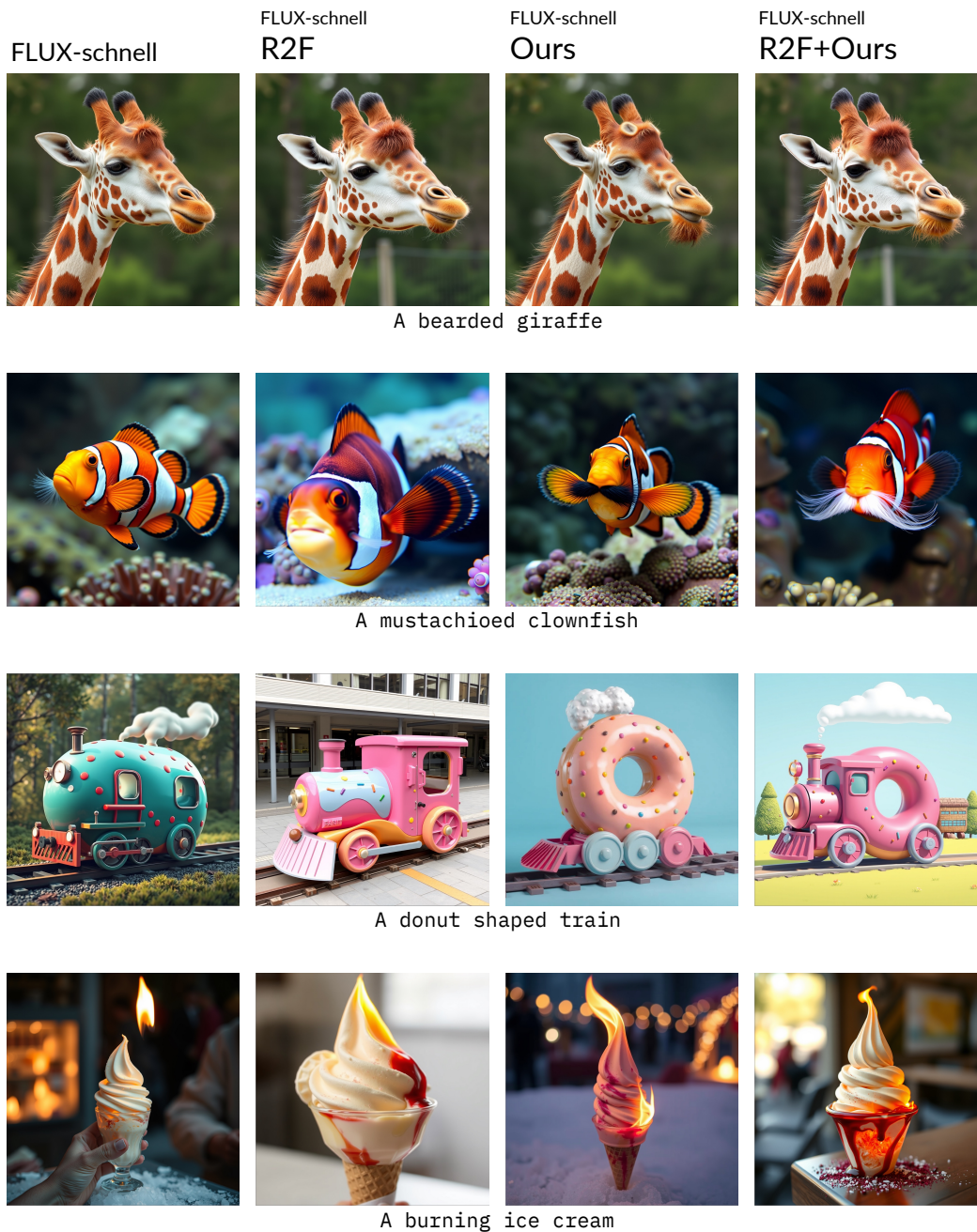


Figure 24: Further Text-to-Image generation comparisons for rare prompts.



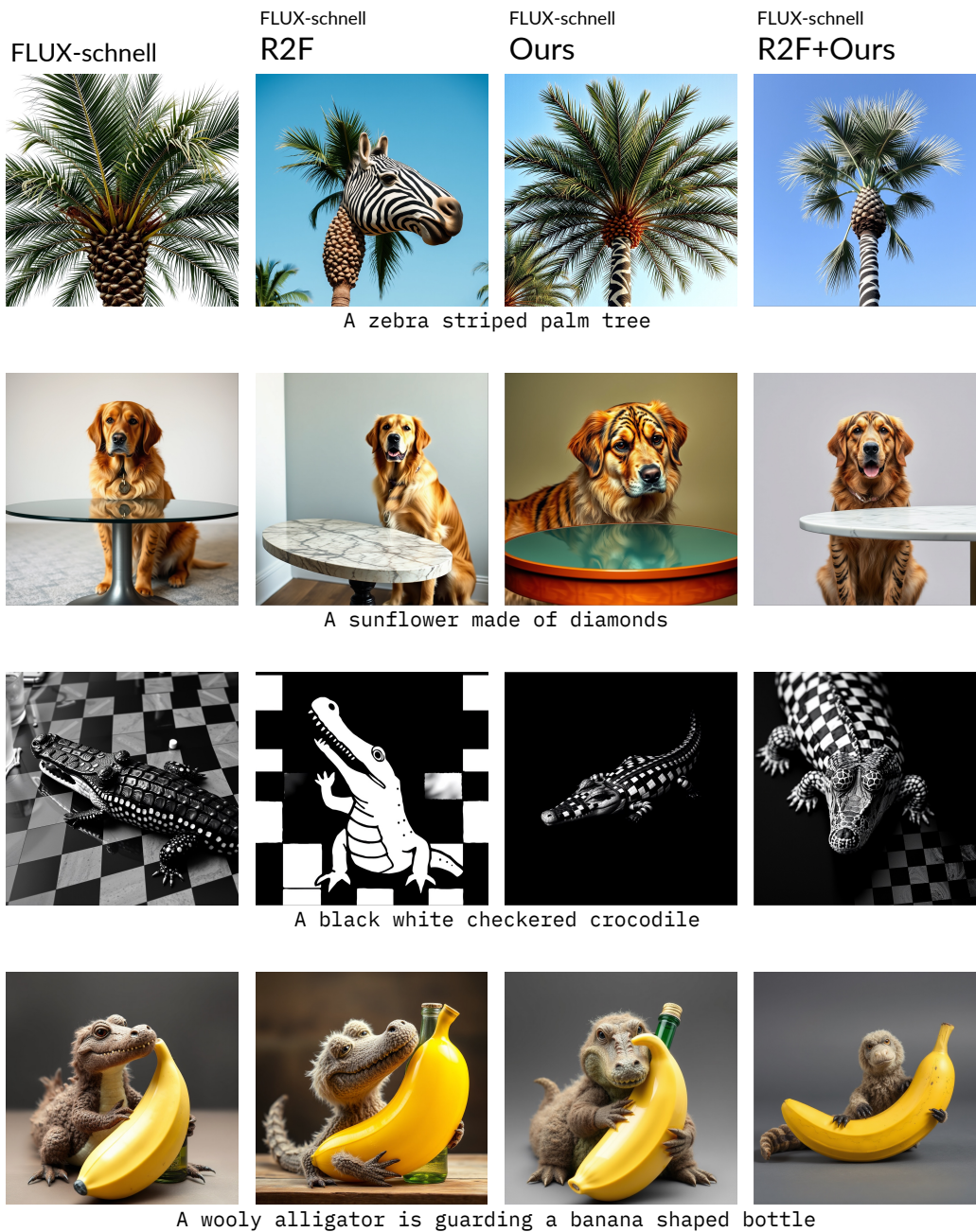


Figure 25: Further Text-to-Image generation comparisons for rare prompts.



A thorny dolphin is leaping over a star shaped bed



A two decorated indian elephants playing chess on a floating island surrounded by pink dolphins painting rainbows in the sky



A blue ant hides in the forest in military uniform aiming a gun at a red enemy ant on horseback



A pirate sailing on a pyramid-shaped boat

Figure 26: Further Text-to-Image generation comparisons for rare prompts.



Figure 27: Further Text-to-Image generation comparisons for rare prompts.



FLUX-dev



FLUX-dev  
Ours



A cactus made of steel and a flower patterned mirror



An orange made of marble



A skyblue unicorn doing kung fu



A red bird with blue fish tail

Figure 28: Further Text-to-Image generation comparisons for rare prompts.





Figure 29: Further Text-to-Image generation comparisons for rare prompts



Figure 30: Further Text-to-Video generation comparisons for rare prompts.



Figure 31: Further Text-to-Video generation comparisons for rare prompts.



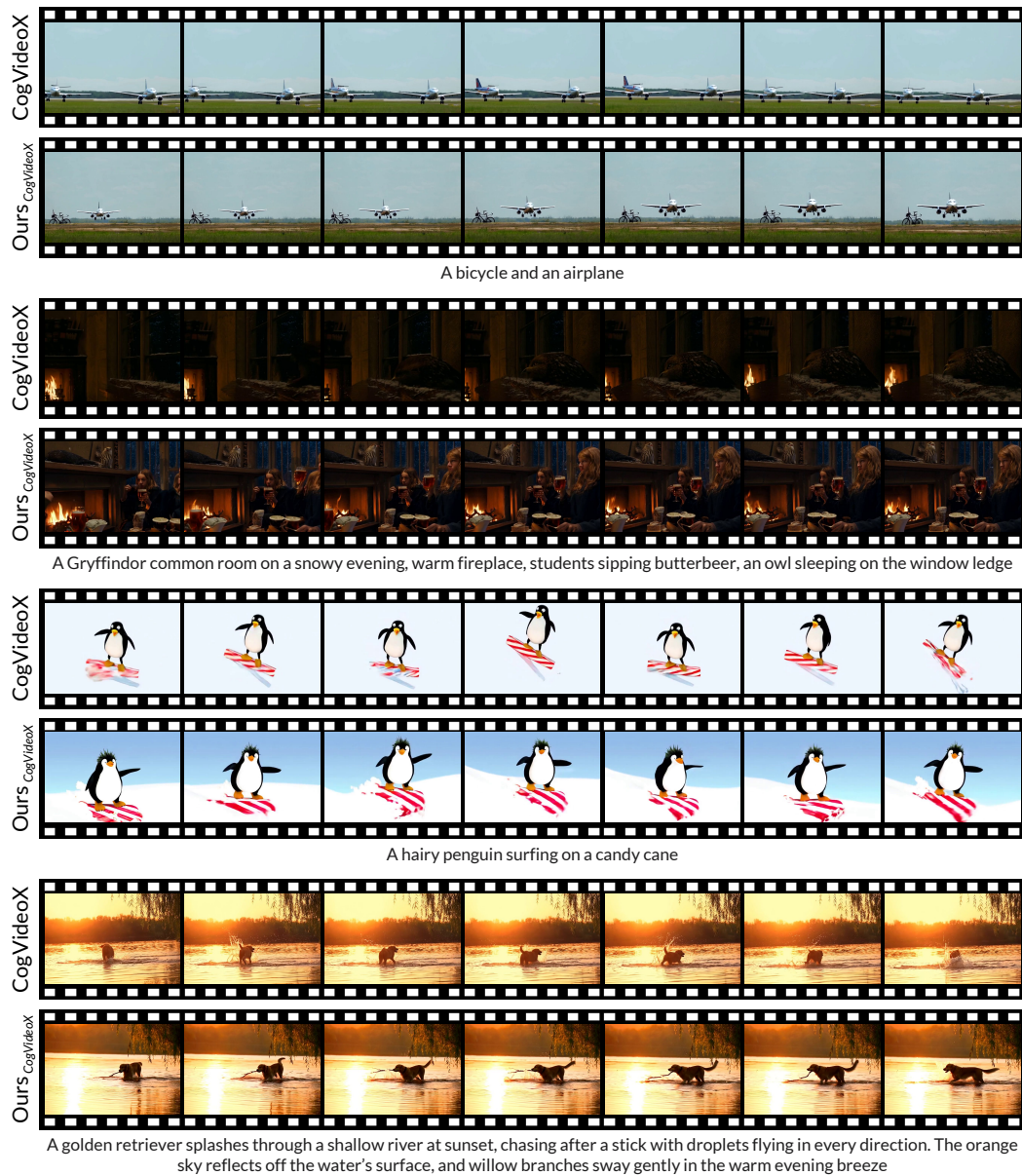


Figure 32: Further Text-to-Video generation comparisons for rare prompts.

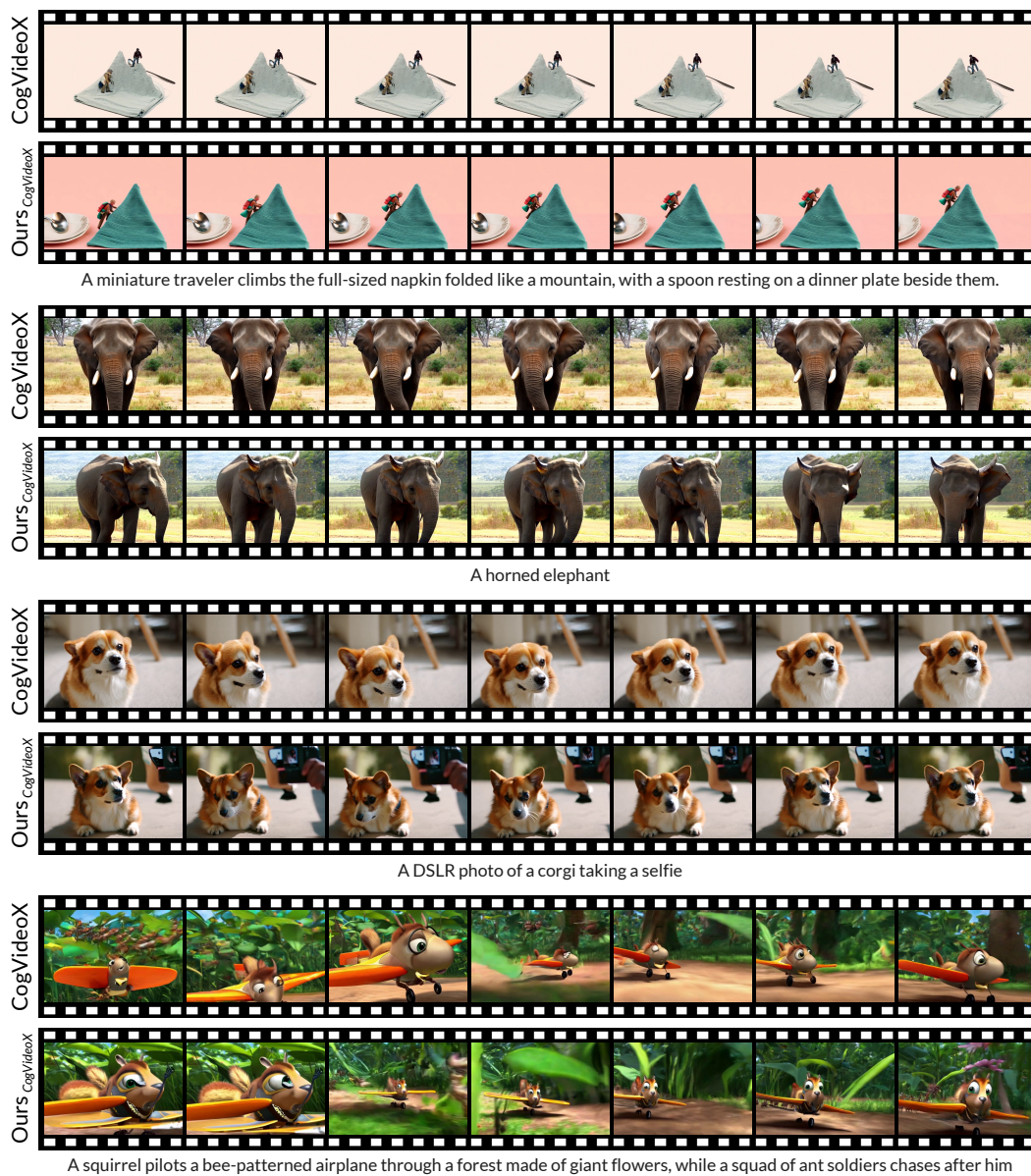


Figure 33: Further Text-to-Video generation comparisons for rare prompts.

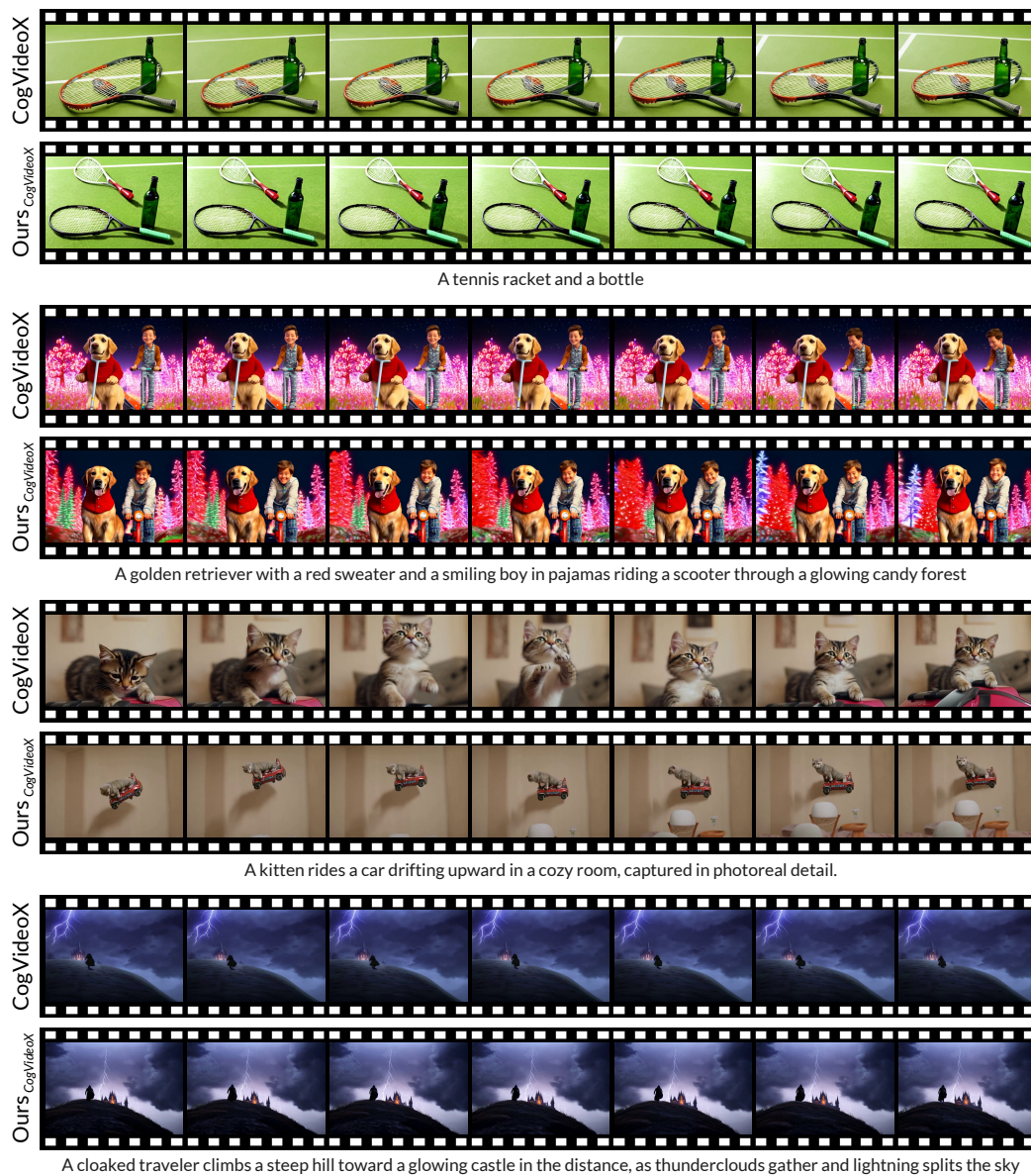


Figure 34: Further Text-to-Video generation comparisons for rare prompts.



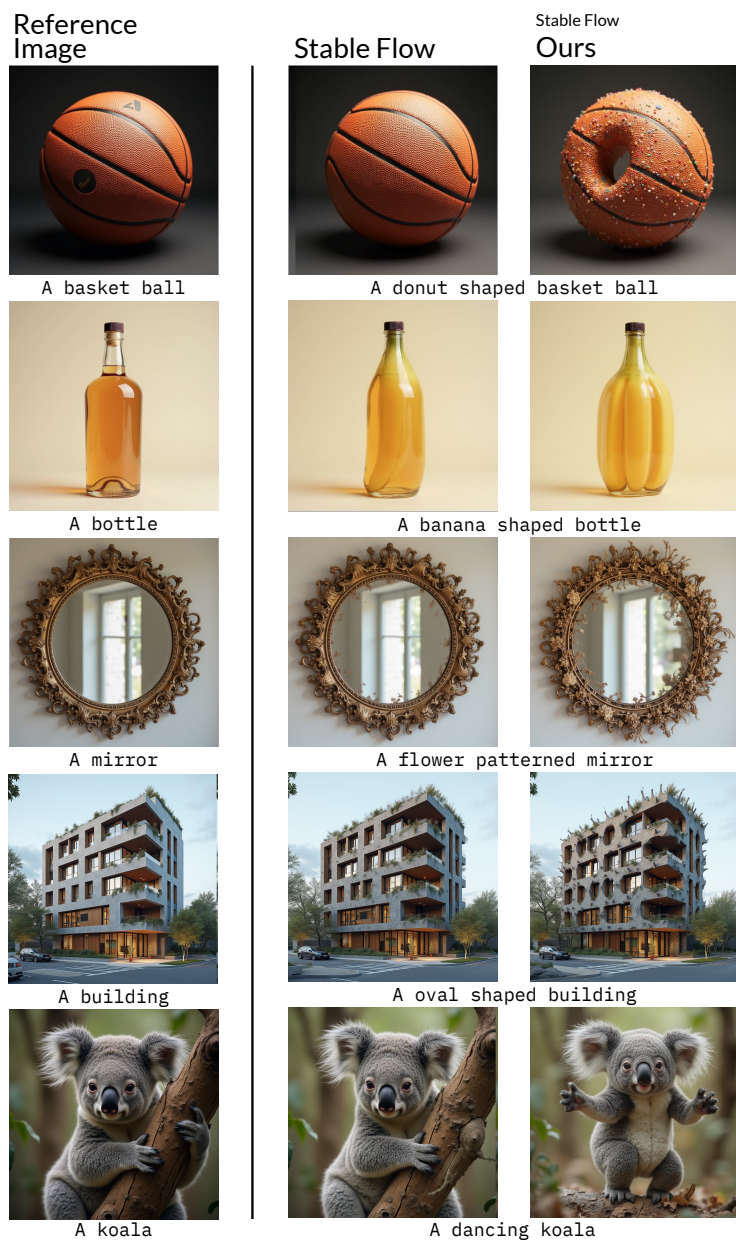


Figure 35: Further Text-Driven Image Editing comparisons for rare prompts.

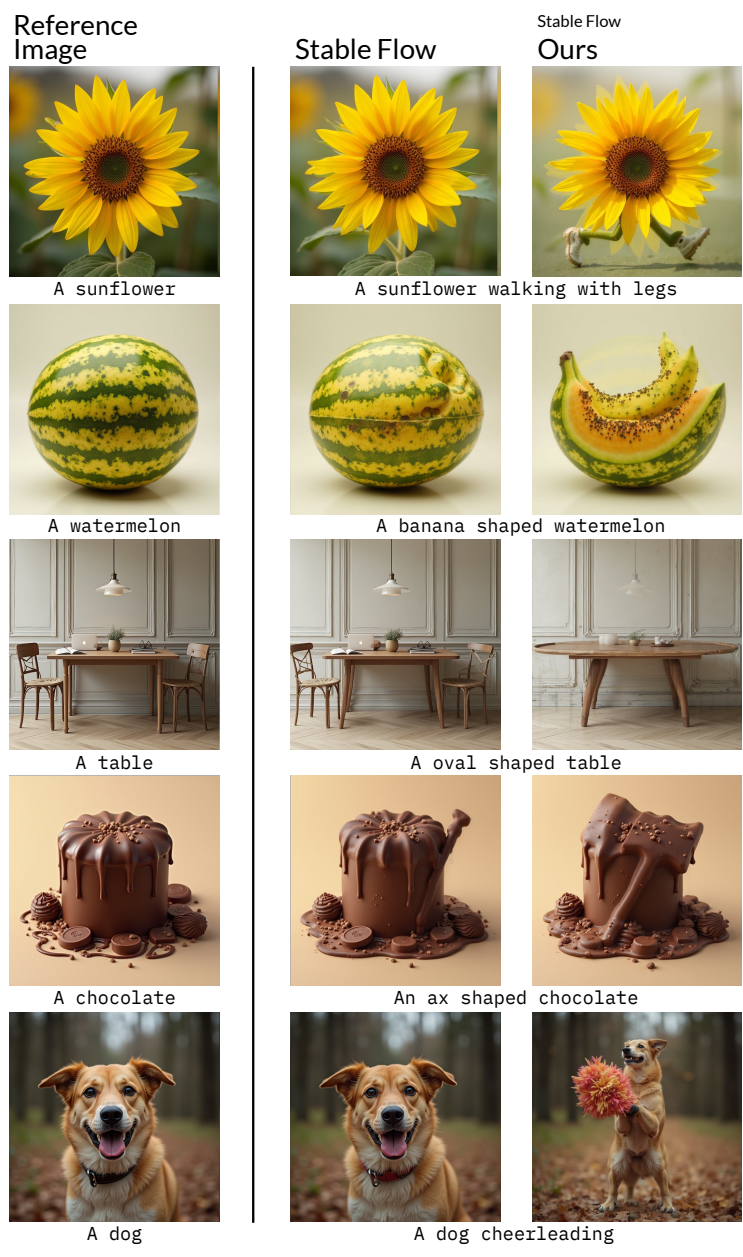


Figure 36: Further Text-Driven Image Editing comparisons for rare prompts.














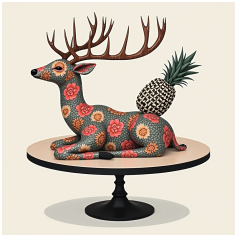
Prompt	GPT-4o	FLUX-dev	FLUX-dev Ours
A thorny shark and a mustachioed dolphin			
A thorny snake is coiling around a star shaped drum			
A hairy octopus dancing with a zebra striped duck is sitting on the top of a star shaped cheesecake			
A flower patterned deer flying with a black white checkered pineapple is resting on an oval table			

Figure 37: Qualitative comparisons with GPT-4o-generated images.

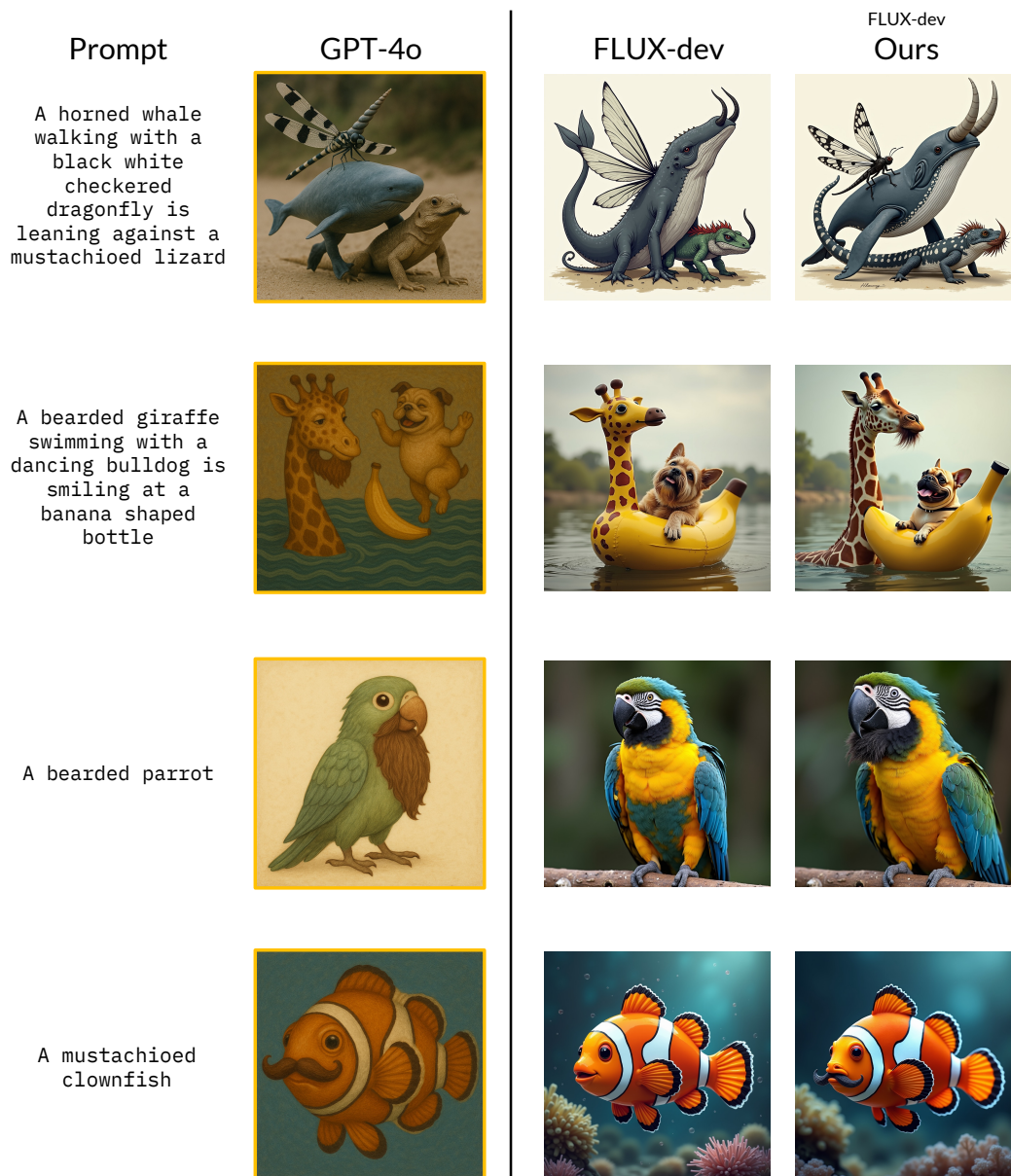


Figure 38: Qualitative comparisons with GPT-4o-generated images.

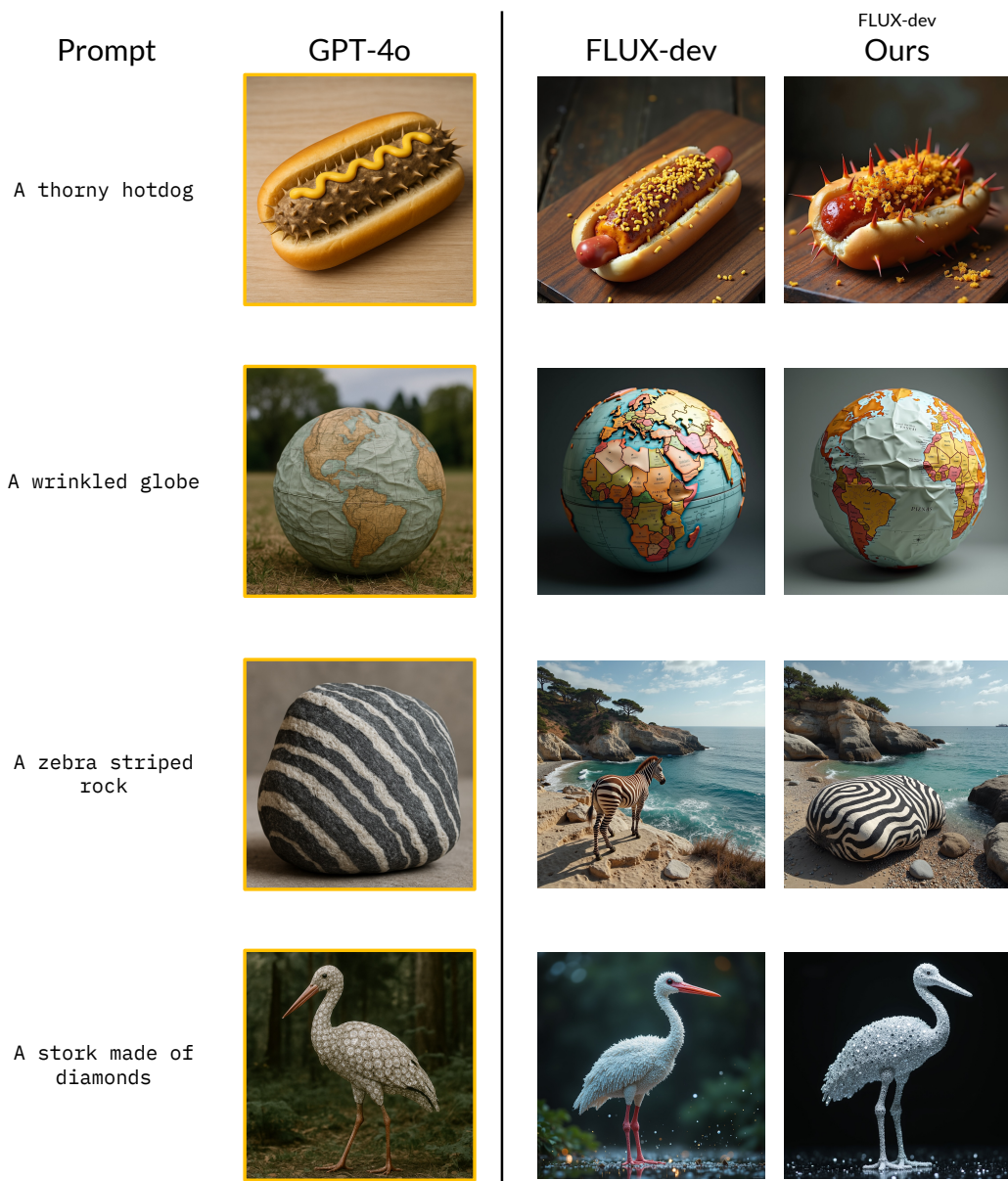


Figure 39: Qualitative comparisons with GPT-4o-generated images.



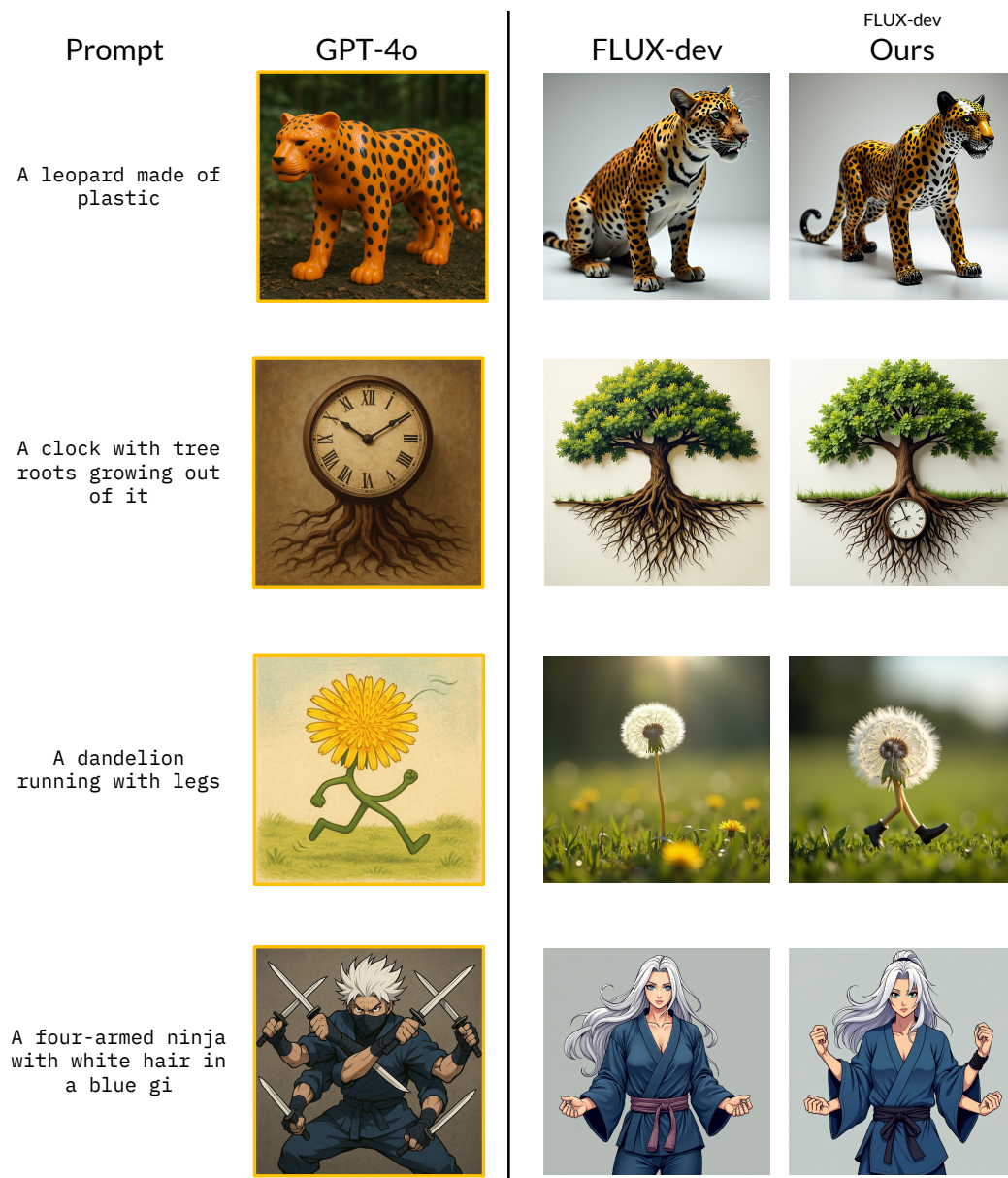


Figure 40: Qualitative comparisons with GPT-4o-generated images.

## References

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR, 2021.
- [2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.
- [3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [4] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [5] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. *arXiv preprint arXiv:2502.04320*, 2025.
- [6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [7] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [8] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. *arXiv preprint arXiv:2411.14430*, 2024.
- [9] Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. *arXiv preprint arXiv:2410.22376*, 2024.
- [10] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [11] William Rudman and Carsten Eickhoff. Stable anisotropic regularization. In *The Twelfth International Conference on Learning Representations*.
- [12] David H. Douglas and Thomas K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10:112–122, 1973.
- [13] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.
- [14] Stefano Recanatesi, Matthew Farrell, Madhu Advani, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. Dimensionality compression and expansion in deep neural networks. *arXiv preprint arXiv:1906.00443*, 2019.
- [15] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- [16] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, pages 7654–7663. PMLR, 2019.
- [17] Kavin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, 2019.
- [18] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

- 488 [19] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused frame-  
489 work for evaluating text-to-image alignment. *Advances in Neural Information Processing Sys-*  
490 *tems*, 36:52132–52152, 2023.
- 491 [20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-  
492 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF*  
493 *conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- 494 [21] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mc-  
495 Grew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and  
496 editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- 497 [22] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin,  
498 and Saining Xie. Representation alignment for generation: Training diffusion transformers is  
499 easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.
- 500 [23] Luping Liu, Chao Du, Tianyu Pang, Zehan Wang, Chongxuan Li, and Dong Xu. Improv-  
501 ing long-text alignment for text-to-image diffusion models. *arXiv preprint arXiv:2410.11817*,  
502 2024.