Rethinking Cross-Modal Interaction in Multimodal Diffusion Transformers

Supplementary Material

Overview. In the supplementary material, we provide further details to support our work. Section A elaborates on the implementation of TACA, including code snippets and a speed comparison of different approaches. Section B presents additional ablation studies focusing on text alignment, examining the effect of CFG guidance scale and the content/length of prompts. Section C explains why we choose LoRA rather than full-parameter finetune. Finally, Section D showcases more qualitative results with visual comparisons on both short and long prompts using FLUX.1 Dev and SD3.5 Medium.

A. Code Implementation Details

Given that TACA necessitates modifications to the attention mechanism, and that the functions for computing attention are typically encapsulated within pre-compiled C/C++ binary libraries, directly reimplementing these attention computation functions using PyTorch would result in a significant performance degradation. To minimize the performance impact of modifying the attention mechanism while retaining the convenience of PyTorch, the following two implementation approaches for TACA can be adopted:

Flex Attention

```
from torch.nn.attention.flex_attention import flex_attention
gamma = 1.2
encoder_size = 512 # T5 encoder seq_len for FLUX

def score_mod(score, batch, head, token_q, token_kv):

    condition = (token_q >= encoder_size) & (token_kv < encoder_size)
    score = torch.where(condition, score * gamma, score)
    return score

hidden_states = flex_attention(query, key, value, score_mod=score_mod)
```

Listing 1. PyTorch Flex Attention

Selective Attention Recomposition

```
gamma = 1.2
encoder_size = 512 # T5 encoder seq_len for FLUX
key_scaled = key.clone()

# Shape of Q, K, V (B, H, N, D)
key_scaled[:, :, :encoder_size, :] *= gamma

# You can also change this into flash attention
hidden_states = F.scaled_dot_product_attention(
    query, key_scaled, value, attn_mask=attention_mask, dropout_p=0.0, is_causal=False
)

hidden_states_orig = F.scaled_dot_product_attention(
    query, key, value, attn_mask=attention_mask, dropout_p=0.0, is_causal=False
)

hidden_states[:, :, :encoder_size, :] = hidden_states_orig[:, :, :encoder_size, :]
```

Listing 2. Selective Attention Recomposition

We conducted empirical evaluations of the computational speed of both proposed methods, comparing them against PyTorch's native scaled dot-product attention implementation. All experiments employ a 30-step denoising process to generate 1024×1024 images via FLUX.1 Dev on a single Nvidia A100 80G GPU. We recorded the performance differential for both a single denoising step and for the complete 30-step denoising process (assuming temperature factor γ modification applied only to the initial 10% of steps). The results of this speed evaluation are presented in Table 6.

Method	Single Step	All 30 Steps	Speedup
Baseline	0.47 sec	14 sec	1.0x
Flex	2.13 sec	19 sec	0.74x
Selective	0.95 sec	16 sec	0.88x

Table 6. Speed Comparison of Different Approaches

B. Further Ablation Study on Text Alignment

B.1. The scale of CFG guidance

To investigate the text alignment improvements offered by our TACA method in comparison to increasing the CFG guidance scale (commonly employed in text-to-image models to enhance alignment, often at the cost of image quality), we conducted a series of ablation studies. These experiments aimed to determine whether TACA maintains its efficacy across varying CFG guidance scales and across different models. The results, presented in Table 7, reveal the effects of different CFG scales and the impact of TACA on both FLUX.1 Dev and SD3.5-Medium.

For FLUX.1 Dev, the default guidance scale of 3.5 appears to be a "sweet spot": further increases in CFG intensity beyond this point yield minimal gains in text alignment, and, notably, performance across several metrics degrades significantly. Concurrently, our TACA method demonstrated effectiveness across diverse guidance scales, suggesting its general applicability.

Model	Settings	Color ↑	Shape ↑	Texture ↑	Spatial ↑
FLUX.1 Dev	CFG = 3.5 (Default)	0.798	0.591	0.755	0.193
	CFG = 3.5 + TACA	**0.839**	0.634	**0.790**	0.207
	CFG = 5	0.787	0.553	0.756	0.175
	CFG = 5 + TACA	0.835	**0.635**	0.757	**0.224**
	CFG = 10	0.667	0.571	0.740	0.137
	CFG = 10 + TACA	0.751	0.633	0.699	0.191
SD3.5-Medium	CFG = 7 (Default)	0.812	0.730	0.850	0.159
	CFG = 7 + TACA	**0.843**	0.737	**0.864**	0.183
	CFG = 10	0.804	0.727	0.853	0.191
	CFG = 10 + TACA	0.820	**0.765**	0.863	**0.206**

Table 7. Ablation study on the effect of CFG scale with and without TACA (with $\gamma_0 = 1.2$) on FLUX.1 Dev and SD3.5-Medium. We randomly sampled 100 prompts for each attribute from the T2I-CompBench dataset to conduct the evaluation. For both models, **bold** indicates the best score and underline indicates the second-best score for each attribute.

For SD3.5-Medium, increasing the CFG scale also enhances text-image alignment but tends to degrade visual fidelity, resulting in reduced metrics (e.g., Color score drops at CFG=10 compared to CFG=7). Our TACA method, however, directly reinforces the dependence of image tokens on textual tokens, improving alignment without such adverse effects. TACA consistently improves results across different CFG scales on SD3.5-Medium, showing both generalization and complementarity.

Overall, the combined results across both models indicate that while increasing CFG can improve alignment to some extent, it often comes at the cost of overall performance. TACA, on the other hand, offers a more targeted and effective approach to enhancing text-image alignment, being beneficial and complementary across different CFG scales and diffusion models.

B.2. The content of the prompt

We have identified several prevalent issues regarding text alignment in state-of-the-art text-to-image models. Our TACA can mitigate these issues to a certain extent.

- Difficulty in handling unrealistic scenarios, such as *"a blue sun and a yellow sea"*.
- Difficulty in handling spatial relationships, such as with the prompt *"a squirrel to the left of the man"*. Models frequently interpret the left side of the image as the left side specified in the text, rather than the left side relative to the man's frame of reference within the image.
- Difficulty in handling specific numerical quantities. For instance, when prompted for four vases, the model may generate images containing five or three vases.

B.3. The length of the prompt

We also observe that models are more prone to omitting details from longer prompts, particularly when the prompt's token count exceeds the maximum token limit supported by the CLIP text encoder.

Our proposed TACA method demonstrates comparatively more widespread effectiveness for mitigating the attribute missing issues often found in longer prompt, rather than the shorter ones. Currently, a mature benchmark for evaluating the text-image alignment capabilities of text-to-image models with long prompts is lacking, despite the practical prevalence of longer prompts in real-world applications. Therefore, we have manually curated a set of authentic, long prompts from the internet to assess our method's performance, and the corresponding results are presented in Fig. **??**.

C. Full parameter fine-tuning vs LoRA

In addition to LoRA training, we also experimented with full parameter fine-tuning as an alternative approach. However, we found that this method required significantly more computational resources and storage, especially for large models like FLUX.1 Dev. Moreover, our experiments revealed that full parameter fine-tuning is highly sensitive to learning rate settings. If the learning rate is set too high, the generated images tend to appear blurry or overly stylized, resembling oil paintings. On the other hand, if the learning rate is too low, the model struggles to learn the original data distribution effectively. These challenges, combined with the lack of superior artifact reduction compared to LoRA, led us to conclude that LoRA training is a more robust, efficient, and practical solution.

D. More Qualitative Results

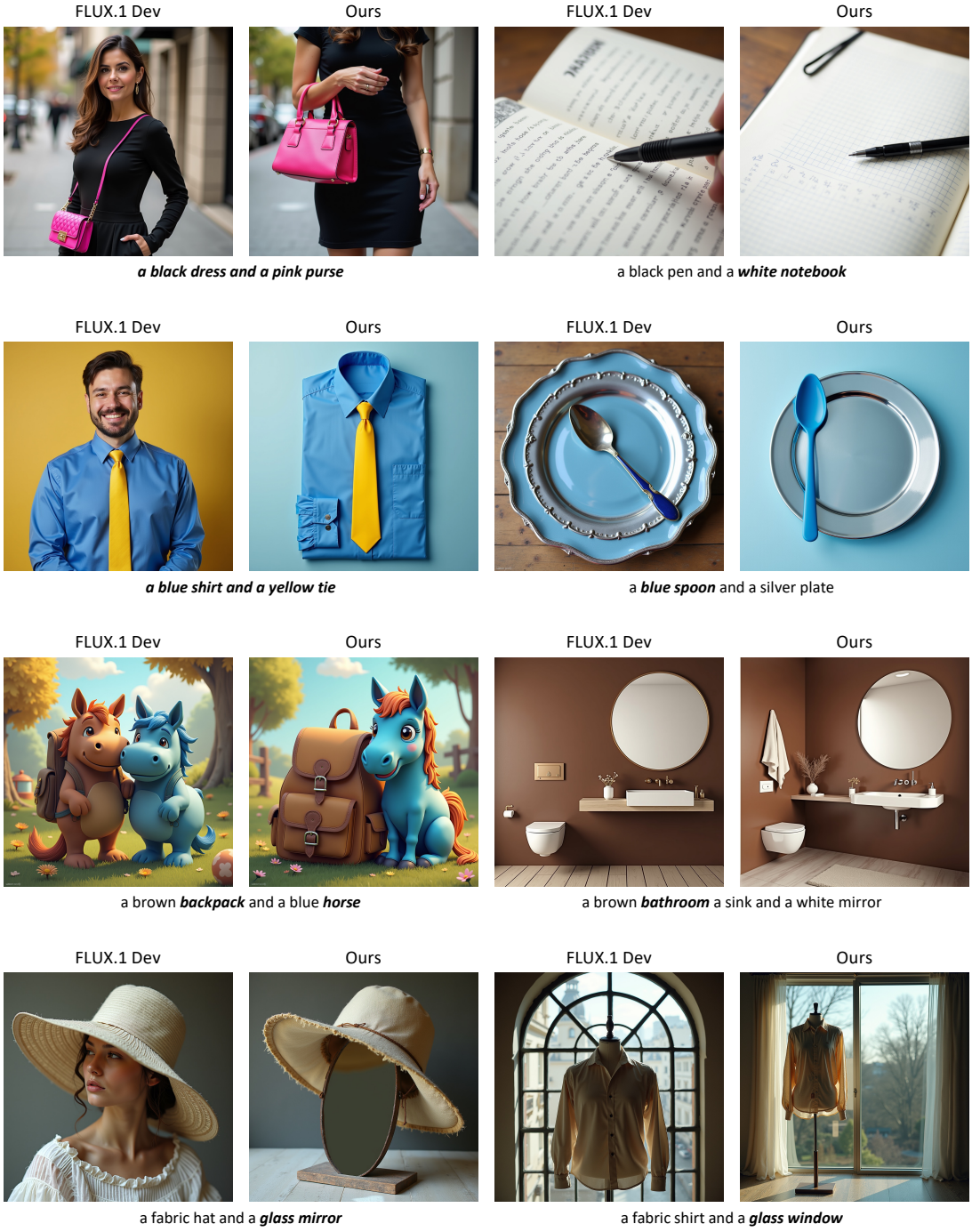


Figure 9. Visual comparisons on text-image alignment (FLUX.1 Dev, short prompts)

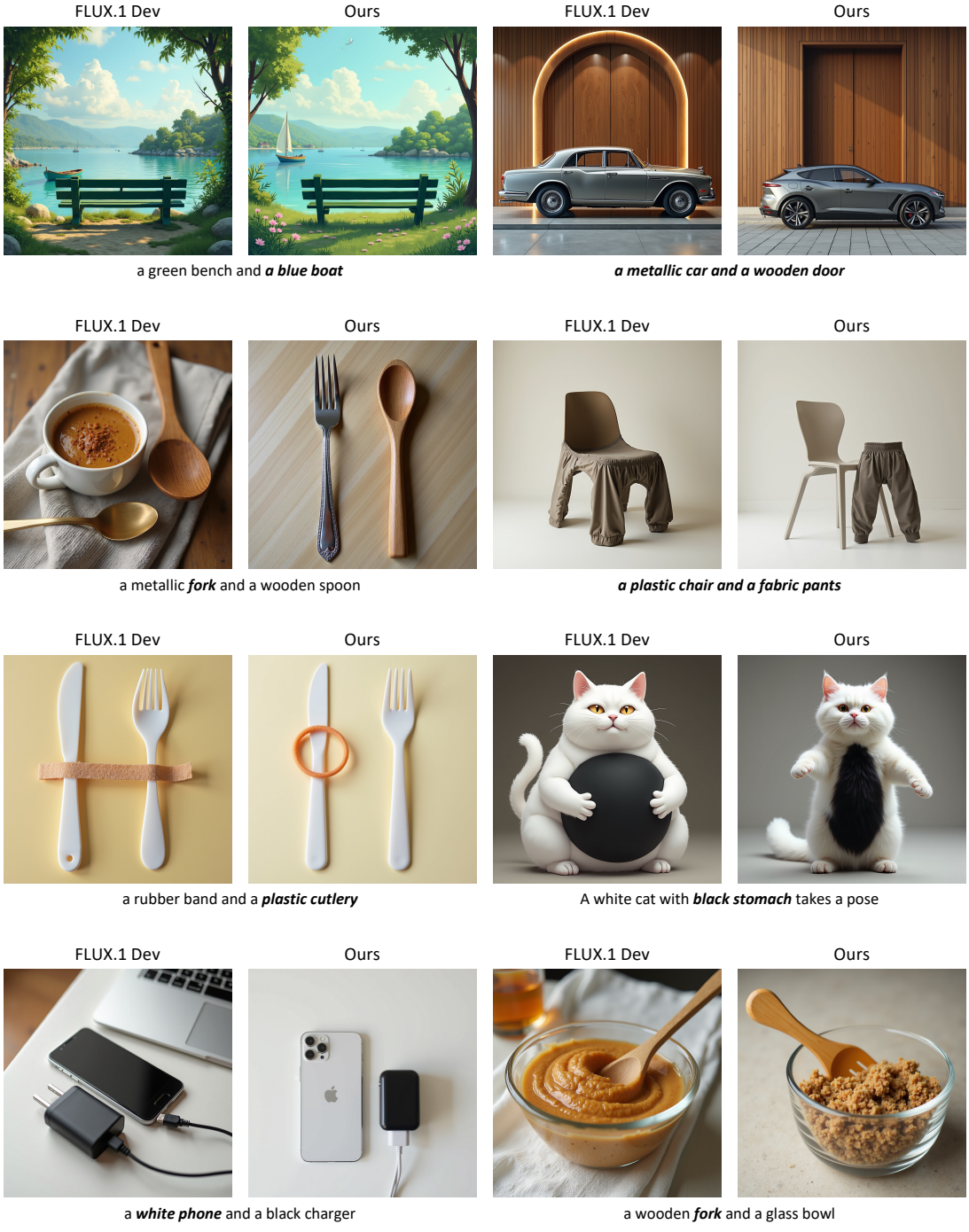


Figure 10. Visual comparisons on text-image alignment (FLUX.1 Dev, short prompts)

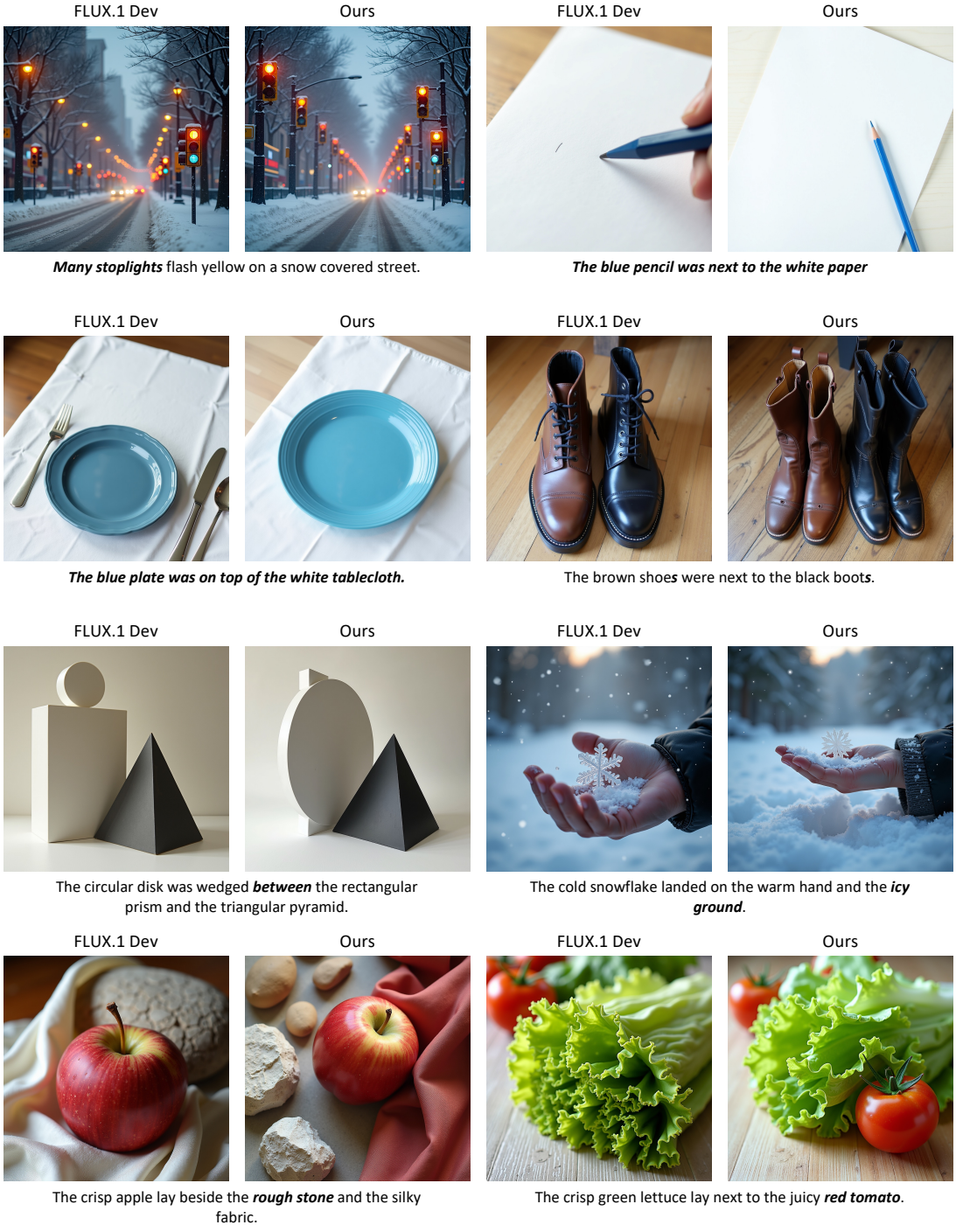


Figure 11. Visual comparisons on text-image alignment (FLUX.1 Dev, short prompts)

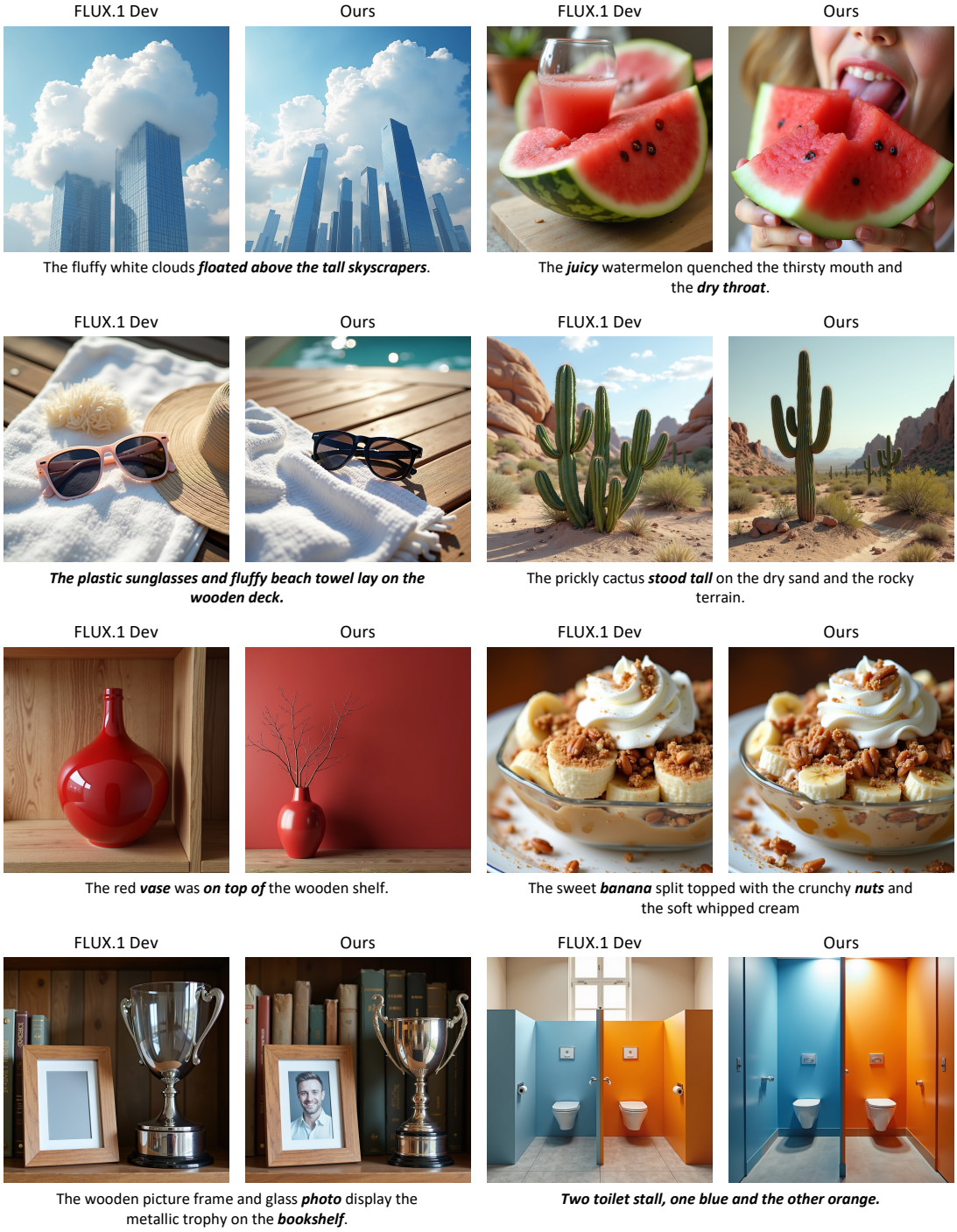


Figure 12. Visual comparisons on text-image alignment (FLUX.1 Dev, short prompts)

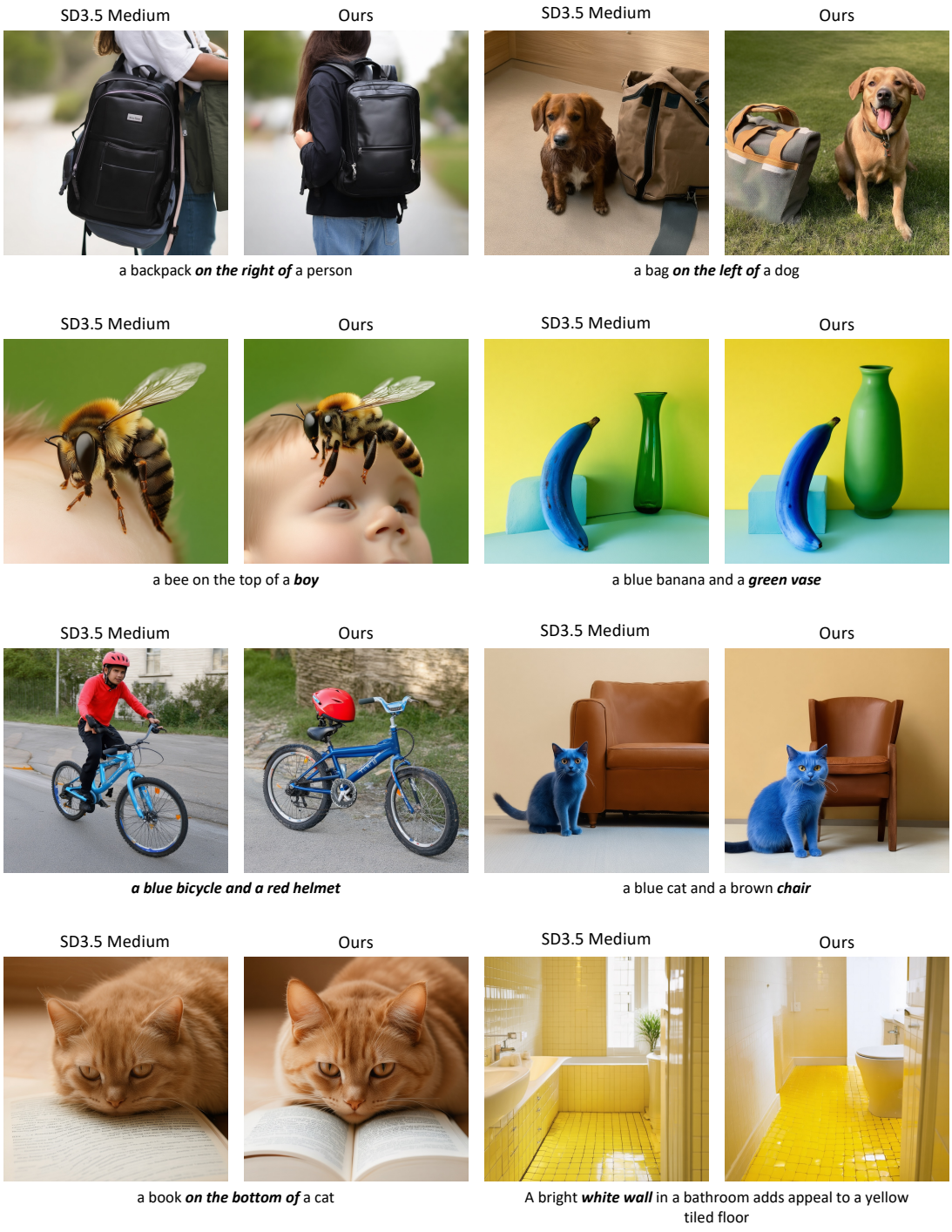


Figure 13. Visual comparisons on text-image alignment (SD3.5 Medium, short prompts)

Figure 14. Visual comparisons on text-image alignment (SD3.5 Medium, short prompts)

FLUX.1 Dev

Ours

in the style of Jed-clrfl, highly detailed 8K image, This is a super high-quality 8K photo. This is a high-quality photo of the cars from the Pixar Cars series, just like real cars. It shows Lightning McQueen racing against other cars from the Pixar Cars series in a grand racing stadium. ***There are a lot of spectator cars in the stadium***. Lightning McQueen is in the lead, Jackson Storm is in second place, and Cheek Hicks is in third place. Several cars collide, creating huge flames and smoke, but Lightning McQueen, Jackson Storm and Chick Hicks luckily escape the crash and remain in the lead.

FLUX.1 Dev

Ours

A captivating, cinematic shot of a sun-drenched coastal city, nestled between ***towering, dramatic cliffs***. The crystal-clear ocean waters lap gently against the pristine sandy shores, where colorful sailboats dot the horizon. The bustling marketplace is filled with the sounds of spirited haggling and laughter, while the tantalizing aroma of fresh seafood permeates the air. A mysterious, dark-cloaked figure ***stands on a cliff overlooking the city***, their gaze lost in the vast expanse of the sea, as if contemplating the mysteries of the world beyond. The painting captures the essence of an enchanting, picturesque landscape, with the dark fantasy element adding a touch of intrigue and depth.

FLUX.1 Dev

Ours

High-angle close-up view of a collection of fallen leaves. The leaves are various shades of dark brown, deep purple, and muted gray tones, indicating the leaf fall's late autumn or early winter state. A single, distinctly lighter-colored leaf, almost white, stands out amid the darker leaves, appearing to be a different type of plant. The light leaf has a maple-like shape. The leaves are densely packed together, creating a textured surface. The overall impression is one of a forest floor covered in ***decaying leaves***. The lighting is diffused, not overly bright, and creates ***an overall subdued and muted atmosphere***.

FLUX.1 Dev

Ours

A cozy oil painting of a pebble path going up towards a beautiful, blood-orange sunset with blue hue. ***On the left of the path are olive trees, on the right of the path is an old and worn out wall.*** The path is full of fireflies. The colors are intense, conveying a sense of isolation and vastness. The brushstrokes are bold and sweeping, while the fireflies are the focus of detail.

Figure 15. Visual comparisons on text-image alignment (FLUX.1 Dev, long prompts)

FLUX.1 Dev

Ours

A very smug-looking chicken stands on a stage in a farmyard, wearing a glittery, oversized tuxedo and ***holding a tiny microphone***. Behind him is a homemade banner that says, “Chicken Idol: Sponsored by Buzz.” To his left, a panel of farm animals—an unimpressed cow, a sheep with headphones, and a pig with sunglasses—sits at a long judging table with little buzz coins as score paddles. The chicken strikes an overly dramatic pose, one wing outstretched, the other clutching the microphone, as if he’s about to belt out a power ballad. His beak is wide open, mid-squawk, with musical notes floating comically out of it

FLUX.1 Dev

Ours

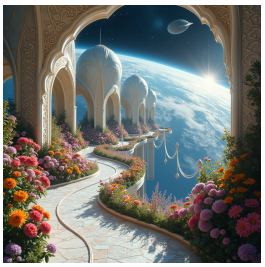
A Cinematic Photography. Black-and-white photograph captures a young woman seated at a small round table in a cozy café setting. She is positioned slightly off-center to the right, with her left elbow resting on the table and her head propped up on her hand, ***giving her a contemplative, pensive expression***. Her hair is straight and falls just below her shoulders, framing her face. She is wearing a form-fitting, long-sleeved black dress that accentuates her slender physique, paired with ***opaque black tights and glossy black high-heeled shoes***. The table in front of her is simple, with a small white cup or saucer, likely for a beverage, placed on it. The café's interior features a rustic charm, with ***metal folding chairs*** and a wooden table with a worn finish. The background shows the glass front of the café, revealing a blurred street scene with parked cars and other indistinct objects, suggesting a bustling urban environment. The lighting is soft, likely natural, creating gentle shadows and highlights that enhance the textures of her clothing and the café's surfaces. The overall atmosphere is intimate and reflective, with a touch of melancholy due to the monochrome palette.

FLUX.1 Dev

Ours

A striking eco-brutalist living room with a tall ceiling, defined by its clean lines, raw textures, and a harmonious connection to nature. A floor-to-ceiling window offers an expansive view of a lush garden filled with swaying palm trees, allowing natural light to flood the space and create a serene, open atmosphere. The centerpiece of the room is a pair of sleek, white, ***modular sofas arranged to foster conversation***. Their minimalist design contrasts beautifully with the raw concrete walls and polished concrete floor, which are signature elements of brutalist architecture. A soft, woven area rug in neutral tones grounds the seating area, adding warmth and texture. ***Above the sofas, contemporary art pieces in bold, abstract shapes and earthy hues are mounted on the wall, providing a dynamic focal point.*** The artwork juxtaposes the rugged materials of the room with creative, modern energy. A low, natural wood coffee table with an organic, irregular shape sits in the center, its surface adorned with a few carefully placed itemsâ ceramic vases, books, and a touch of greenery in a glass terrarium. Nearby, a statement plant, such as a tall monstera or bird-of-paradise, adds to the room's eco-friendly vibe. The lighting includes a sculptural pendant light hanging from the tall ceiling, its design inspired by nature with a modern twist. Subtle recessed lighting emphasizes the raw textures of the walls without overpowering the natural light streaming through the window. The lush garden outside, with its vibrant greenery and tall palm trees, feels like an extension of the room itself, framed perfectly by the expansive window. The interplay of natural elements, rugged architecture, and minimalist design creates a tranquil yet visually compelling space.

FLUX.1 Dev



Ours

A breathtaking view from a magical fairy space station in outer space, gazing towards Earth. Iridescent ivory walkways lead between ***hybrid fantasy/sci-fi housing pods***, moored together by sparkling gossamer ropes, stretch between each colossal pod. Along the path, elegant white lattice archways rise at regular intervals, each intricately detailed with swirling patterns. These arches are adorned with an explosion of vivid flowers in every imaginable color, from crimson roses to golden sunflowers and delicate lavender blooms, their petals seemingly untouched by gravity. The Earth below is illuminated, showing vibrant blues and greens, while the vast blackness of space is dotted with twinkling stars, adding a magical, dreamlike quality.

Figure 16. Visual comparisons on text-image alignment (FLUX.1 Dev, long prompts)

FLUX.1 Dev

Ours

In the heart of a verdant Swiss village, nestled between towering, alpine mountains, a subtle yet profound moment of connection unfolds. The colorful timber-framed houses, their facades painted in bright yellow and red, stand as a tapestry of harmony against the serene greens and blues of nature. The air is crisp, carrying the scent of pine and the faint, distant sound of ***a rushing river***. ***A charming pedestrian bridge***, its wooden planks worn smooth by countless footsteps, ***arcs gracefully over the water***, connecting two sides of the village. Under the soft, golden light of the setting sun, a middle-aged couple, both with the weathered yet gentle faces of those who have lived and loved deeply, stand at the edge of the bridge. The man, a robust figure with a thick beard and a knitted cap, leans against the railing, his eyes locked on the woman standing a few steps away. She, with long, auburn hair that cascades down her back and a dress of deep blue, turns to meet his gaze. Her cheeks are flushed, and a gentle smile plays on her lips. The bridge, a symbol of connection and passage, amplifies the sense of mutual understanding and shared history between them. The wooden beams and slate roofs of the houses around them echo the steadfast nature of their bond, a bond that has weathered the seasons and the trials of time. ***The river, with its constant, soothing flow, mirrors the deep, enduring current of their love.*** In this moment, the world around them seems to pause, as if acknowledging the profound and timeless connection that exists between two hearts.

FLUX.1 Dev

Ours

a woman with long blonde hair, wearing a black top and ***a ring on her finger***, standing in a city square with a fountain in the background, looking up ***and to the left***, ***thoughtful pose***, realistic city life scene, bright daylight

FLUX.1 Dev

Ours

a man with fair skin, wearing a white vest and a purple shirt, standing in front of a futuristic blue and black machine ***with various buttons and screens***, looking to his left with a surprised expression, dark and moody background with a hint of a spaceship in the distance, science fiction style

FLUX.1 Dev

Ours

a man with dark hair, wearing a blue apron, and a woman with shoulder-length hair, wearing a black shirt, engaging in conversation in a rustic kitchen, wooden wall background with various objects including ***a deer head, a clock***, and a shelf with items, warm and cozy atmosphere

FLUX.1 Dev

Ours

a young man with short dark hair, wearing a black cap and a blue life jacket, standing ***on a boat*** moving through water, ***looking to the side*** with a slight smile, ***rocky shore and water in the background***, casual and candid capture

Figure 17. Visual comparisons on text-image alignment (FLUX.1 Dev, long prompts)