

TF-TI2I: Training-Free Text-and-Image-to-Image Generation via Multi-Modal Implicit-Context Learning in Text-to-Image Models

Supplementary Material

A. Experimental Setting

For the backbone T2I model, we mainly follow the provided default setting.

- Pre-trained T2I model: Stable Diffusion 3.5 Large¹
- Random seed: 0
- Number of inference steps: 28
- Classifier-free-guidance: 5
- Resolution: 1024x1024

Due to the instability of RCM in an early layer, as shown in Fig. 17a. (akin to the cross attention map in UNet-based T2I models), we following previous methods that only activate masking at the late layer [14, 52], we only activate RCM when $l > 25$. This strategy ensures that RCM is applied only when the masking becomes stable, balancing information retention and disentanglement.

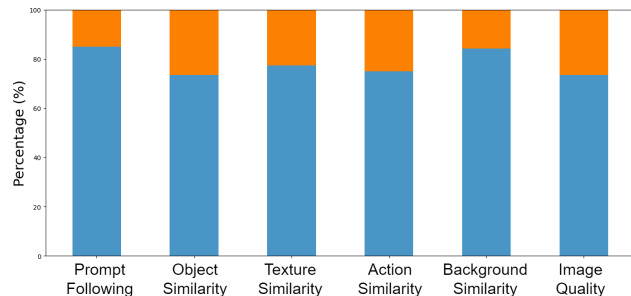


Figure 9. The user study comparison of TF-TI2I (denoted by blue) versus Emu2 (denoted by orange).

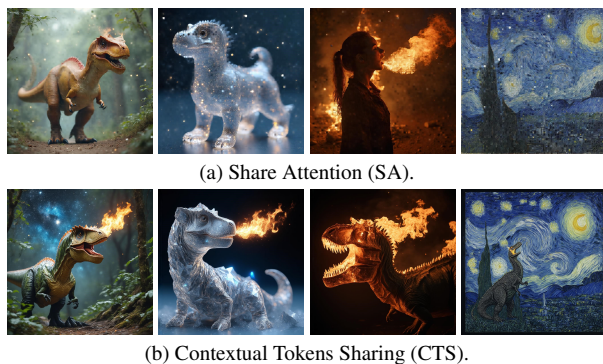


Figure 10. A single-reference comparison of CTS versus Share Attention [3, 13, 15, 22, 23], using example in Fig. 1.

¹<https://huggingface.co/stabilityai/stable-diffusion-3.5-large>

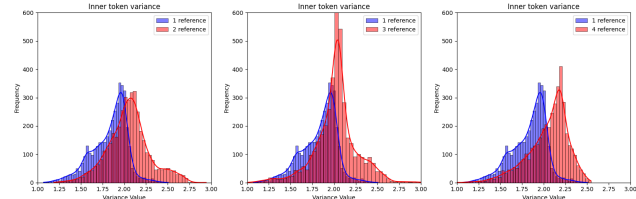


Figure 11. The visualization of distribution shift leading by additional visual references, computed in $t = 20, l = 37$.

B. Experimental Results

Due to space limitations in the main paper, we present additional experimental results here, including more qualitative results of TF-TI2I on the FG-TI2I Benchmark, incorporating dual-reference and trio-reference sub-tasks Fig. 13. We also report additional generation results on Dream-Bench and Wild-TI2I. Furthermore, we conduct an additional quantitative evaluation using a different set of metrics Tab. 4 to provide a more comprehensive assessment of TF-TI2I.

B.1. Qualitative Results

B.1.1. User Study

We invite 20 participants, each of whom is presented with 10 randomly sampled pairs of generated output from FG-TI2I for comparison (ours versus Emu2). They are asked to evaluate the images combined with input references based on six criteria: prompt following, object similarity, texture similarity, action similarity, background similarity, and overall image quality. The results in Fig. 9 demonstrate the superior performance of TF-TI2I over Emu2 across all metrics, which is further supported by our qualitative analysis.

B.1.2. Quad-Reference Sub-Tasks for FG-TI2I

We provide additional comparison of FG-TI2I with multiple references in Fig. 12. TF-TI2I achieves superior generation results compared to Emu2, effectively balancing information from multiple references in a more harmonious manner. Furthermore, benefiting from the high-quality generation capability of our pre-trained T2I backbone, the generated outputs exhibit greater visual fidelity than those of Emu2. These findings support our hypothesis that adapting T2I models for TI2I tasks can achieve higher performance with minimal cost.

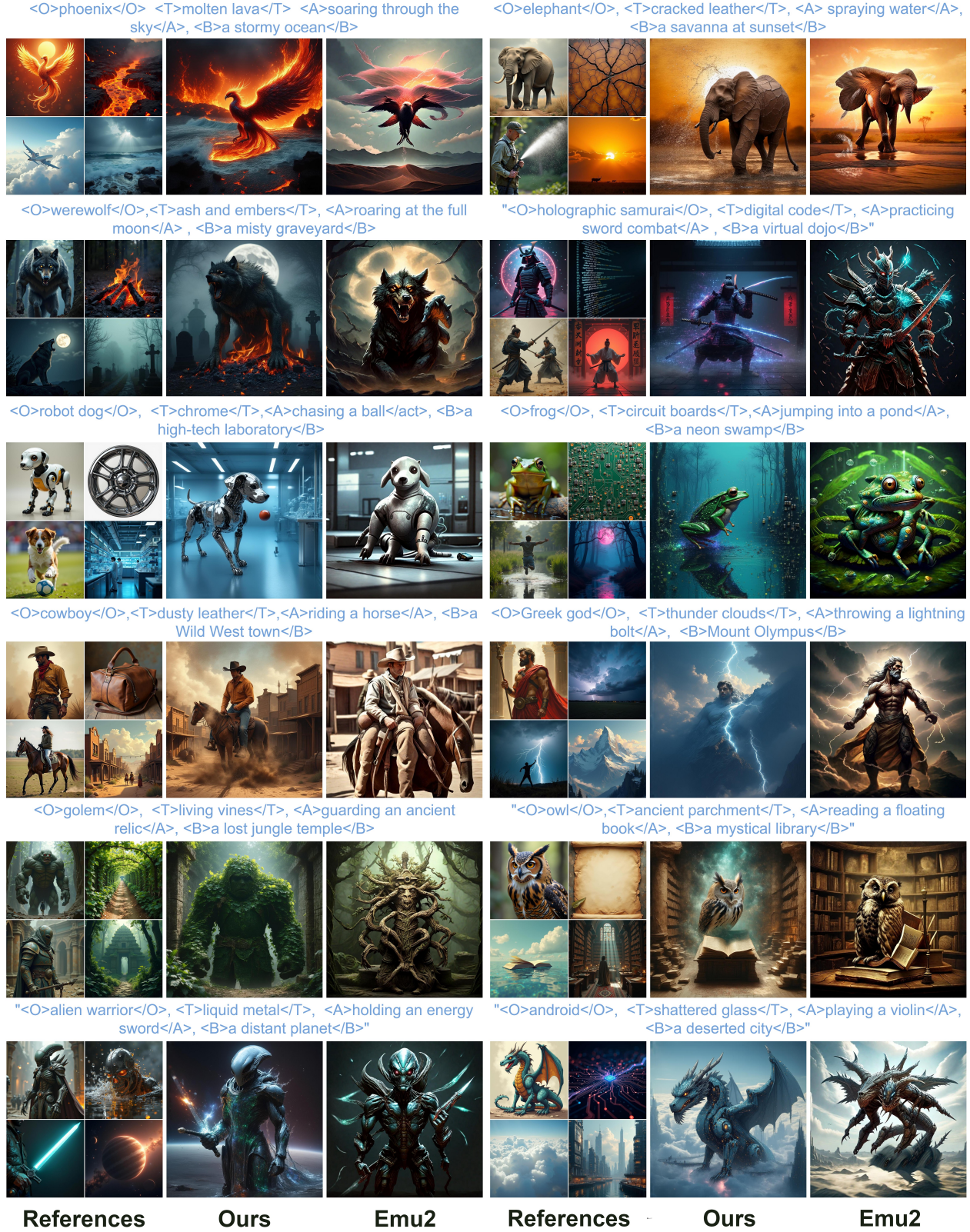


Figure 12. Qualitative comparison of Quad-references sub-tasks. The input Object, Texture, Action, and Background—are denoted as O, T, A, and B. We use red for text-only input and blue for reference-supported input.

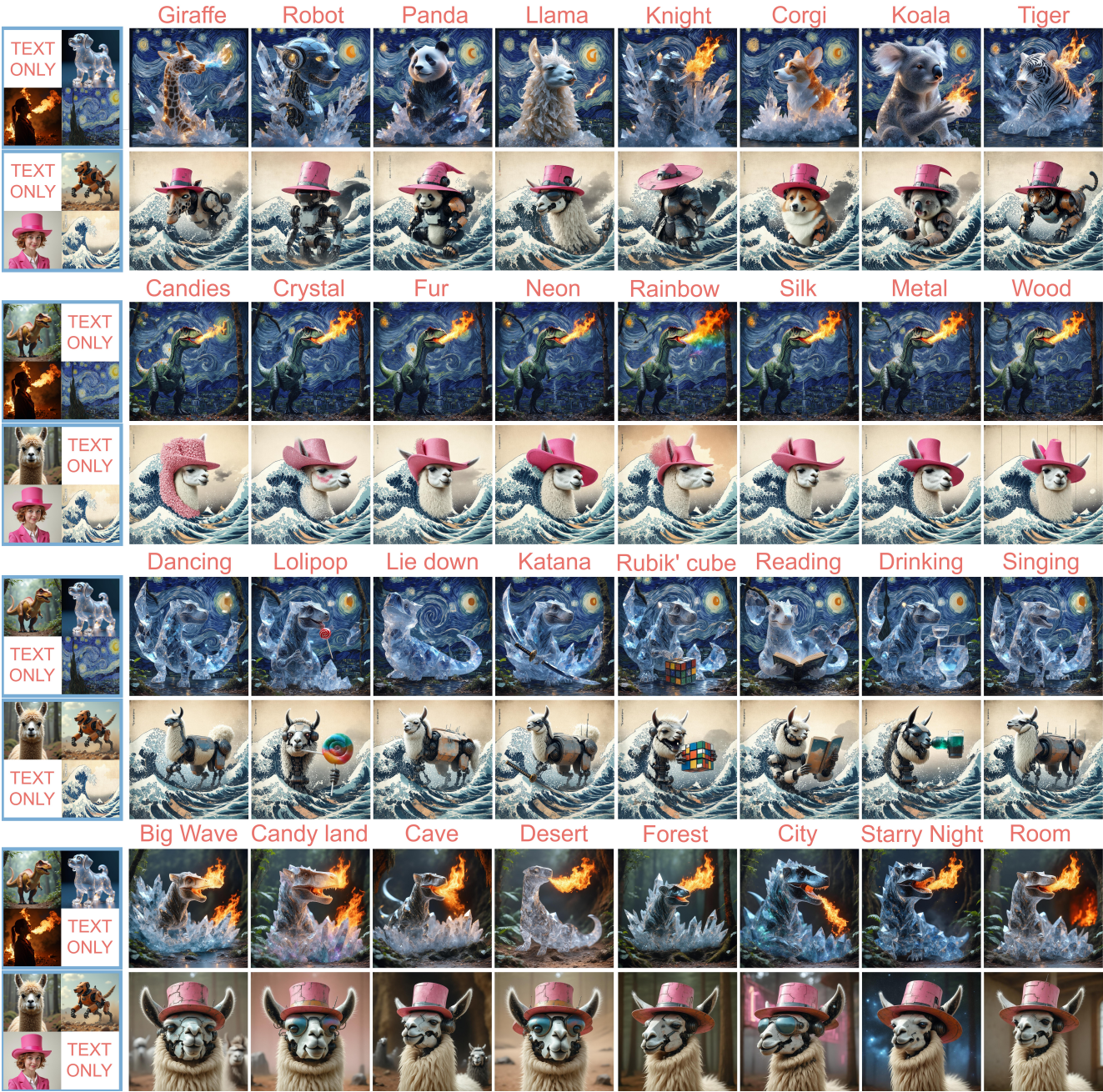


Figure 13. Qualitative results of Trio-references sub-tasks on FG-TI2I, the first template of the prompt is *a dinosaur with the texture of crystal doing breathing fire in the starry night*, and the second one is *a llama with the texture of robot doing wearing pink high gat in the great wave*.. In this illustration, we remove a single image and replacing with other text each time.

B.1.3. Trio-References Sub-Tasks for FG-TI2I.

The qualitative result is shown in Fig. 13, whereby keeping the image references while changing a single textual prompt, we can reach an effect similar to consistency image generation, for example, in rows 1 and 3, we can change the object or adding new object while keeping the consistency throughout different prompts. However, for abstract tex-

tual instruction, such as changing texture (row 2) or changing the background (row 4), The effect of textual guidance becomes weaker. This is due to the overly rich information from image features, and the abstract concept from textual instruction is heavily interfered with by the overly rich visual prior provided by the contextual token. This phenomenon is also discussed in Appendix C that our cur-

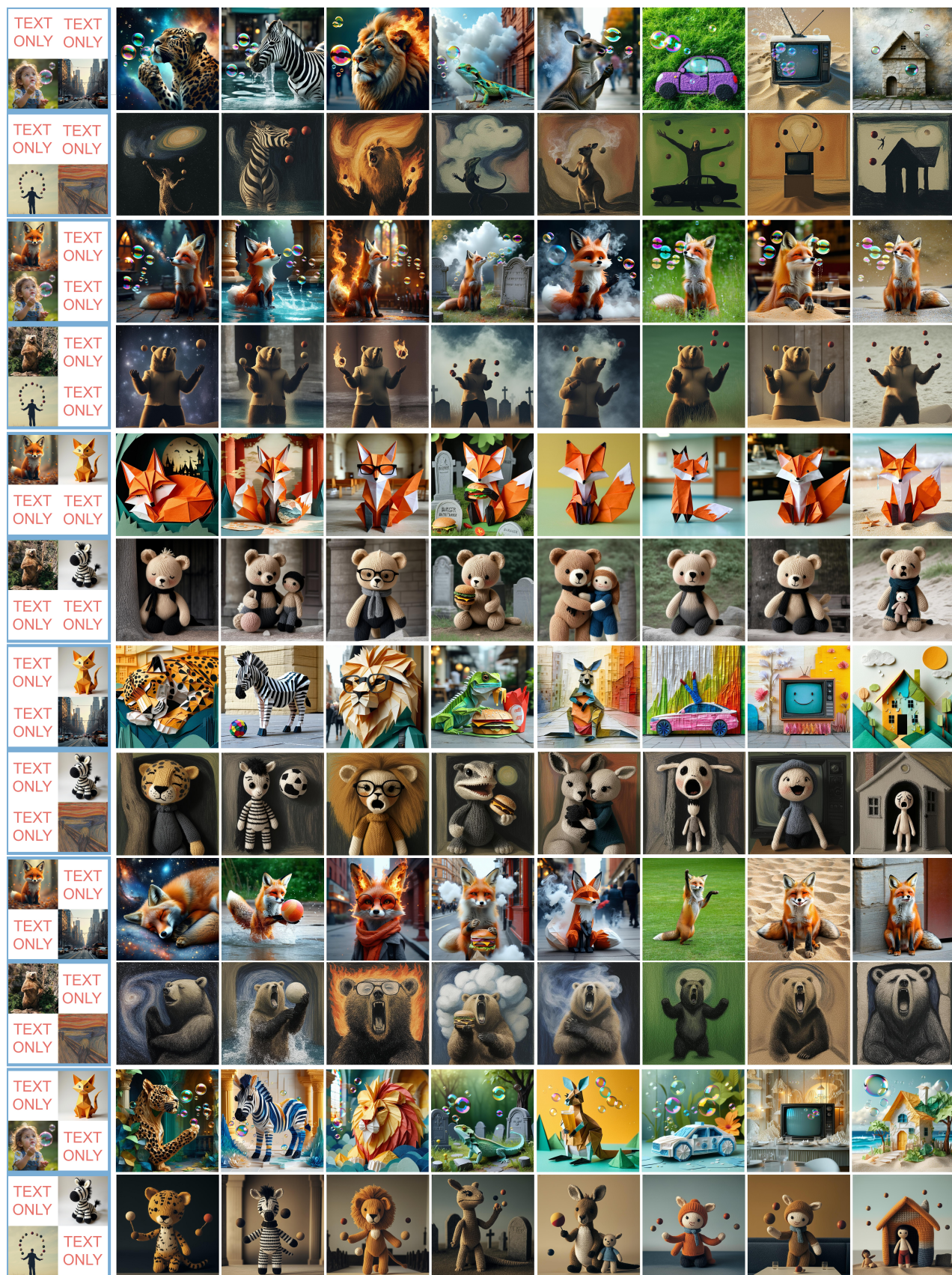


Figure 14. Qualitative results of Dual-references sub-tasks on FG-TI2I, we leave the textual prompt in Appendix B.1.4

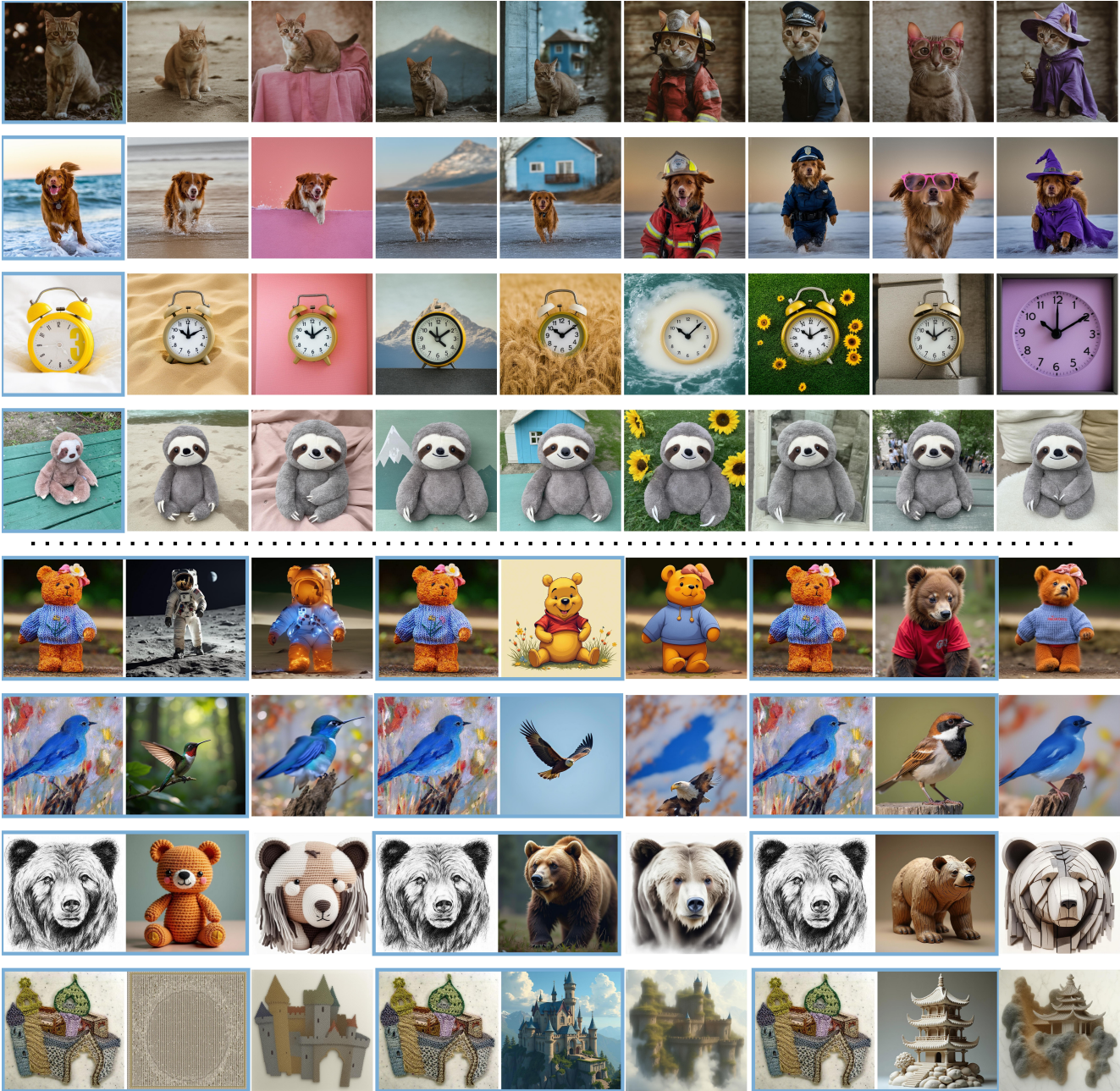


Figure 15. Qualitative results of TF-TI2I on DreamBench (upper part) and Wild-TI2I (lower part), where the image input is denoted with blue, the textual prompts are abbreviated for conciseness.

rently proposed RCM and WTA are more sensitive to object prompts instead of another concept.

B.1.4. Dual-Reference Sub-Tasks for FG-TI2I

For this demonstration, we design the instructions following the Quad-Reference template but modify two prompt entries at a time. The prompts are formulated as follows:

- Object: [leopard, zebra, lion, lizard, kangaroo, car, television, house]

- Texture: [stars and galaxy, water, fire, cloud, smoke, grass, sand, stone, ice]
- Action: [sleeping, playing ball, wearing glasses, eating a burger, hugging, handstand, smiling, crying]
- Background: [haunted mansion, palace, church, graveyard, school, hospital, restaurant, beach]

The generation results are shown in Fig. 14. Due to the reduced number of image references, the generated outputs

exhibit greater creativity. However, the influence of image references also appears to diminish, validating that while increasing the number of references enables more precise control, it also restricts generation diversity.

B.1.5. DreamBench and Wild-TI2I

As a versatile TI2I model, TF-TI2I effectively generates coherent objects in DreamBench while also producing prompt-following yet reference-aligned outputs in Wild-TI2I, as shown in Fig. 15. For instance, it seamlessly transfers objects across different visual styles, including realistic photos, paintings, and pencil sketches. These results further validate the efficacy of TF-TI2I across various TI2I evaluation settings.

B.2. Quantitative Results

While numerical metrics alone may not be ideal for TI2I evaluation—*i.e.* directly copying the reference can achieve the highest reference-alignment score while maintaining decent image quality—methods focused on inversion and editing, such as StyleAlign and MasaCtrl, tend to dominate these two metrics. However, as shown in Fig. 6, their generated results are often suboptimal compared to other approaches. To provide a more comprehensive evaluation, we present additional quantitative results using alternative metrics, as shown in Tab. 4. Notably, TF-TI2I still achieves competitive performance in prompt-following and image quality, as measured by HPS and AS.

C. More Analysis

C.1. Contextual Token Sharing

Introducing additional key and value tokens can lead to distribution shifts, as discussed in prior works [13, 15, 35]. Similarly, TF-TI2I exhibits this phenomenon, as shown in Fig. 11. As the number of reference images increases, the variance within each token grows, causing different attention heads to become inconsistent and more uncertain. This uncertainty propagates to the generated output, resulting in objects and overall image styles that appear inconsistent and inharmonious, ultimately leading to sub-optimal results.

C.2. Reference Contextual Masking

The visualization of Reference Contextual Masking (RCM) is presented in Fig. 17a. Instruction-related features, particularly those associated with object references, can be identified based on the input instruction. However, in early steps or shallow layers, these features may be less distinct, especially for non-object instructions. Prior research [4, 14, 31] also suggests that attention maps for objects are generally more prominent than those for abstract concepts. As the process progresses, feature extraction stabilizes in deeper

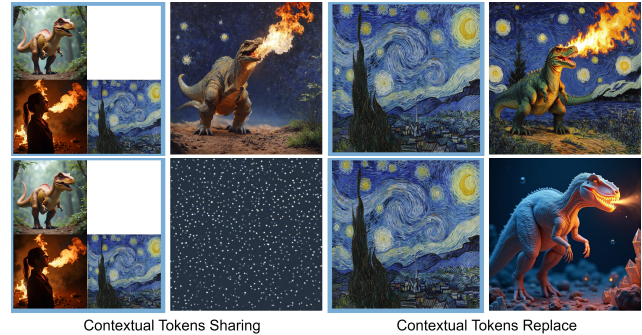


Figure 16. Illustration of introducing TF-TI2I to two other backbone model Stable Diffusion 3 medium [17] (upper) and Flux [27] (lower part), the input references is denoted by blue.

layers and later steps. To ensure precise feature extraction, we activate RCM only in the late steps.

C.3. Winner-Takes-All

The visualization of WTA across different layers and steps is presented in Fig. 17b. The reference assignment closely resembles the cross-attention map, as discussed in prior works [3, 4, 14, 21]. These findings suggest that by encoding visual features from the image, the contextual token can function similarly to a textual instruction. This visualization validates our design of the WTA module, which serves a role akin to traditional cross-attention layers while mitigating distribution shifts caused by an increasing number of references.

C.4. Changing Backbone Model

As TF-TI2I is designed for seamless integration with other MM-DiT-based TI2I models, we evaluate its functionality on two different models: Stable Diffusion 3 Medium (SD3m) [17] and Flux 1.0-dev [27]. Toy examples are illustrated in Fig. 16. The results demonstrate that TF-TI2I can be readily integrated into the SD3-medium model either by sharing contextual tokens or replacing them.

On the other hand, in Flux (the lower part of Fig. 16), sharing additional contextual tokens produces out-of-distribution and visually disturbing images. While replacing contextual tokens with another reference results in a visually harmonized image, the reference effect remains negligible. Given that replacing contextual tokens with unrelated tokens does not alter Flux’s output, we contend that the pooled condition from the input already guides the generation process, making contextual tokens unnecessary. Additionally, Flux’s architecture is not fully MM-DiT-based; only half of its layers utilize multi-modal attention, while the remaining layers rely solely on single-modal attention for vision tokens.

	OBJ x TEX			OBJ x ACT			OBJ x BG			OBJ x TEX			OBJ x ACT			OBJ x BG		
	HPS↑	LP↓	AS↑	HPS↑	LP↓	AS↑	HPS↑	LP↓	AS↑	HPS↑	LP↓	AS↑	HPS↑	LP↓	AS↑	HPS↑	LP↓	AS↑
MasaCtrl [3]	0.25	0.17	6.67	0.23	0.18	6.05	0.25	0.19	6.90	0.23	0.22	5.83	0.23	0.24	5.77	0.23	0.15	6.91
StyleAlign [22]	0.25	0.41	6.16	0.25	0.43	5.84	0.27	0.47	6.70	0.25	0.42	5.98	0.26	0.46	5.75	0.26	0.44	6.51
OmniGen [57]	<u>0.26</u>	<u>0.57</u>	6.32	<u>0.27</u>	0.61	6.01	0.30	<u>0.57</u>	7.01	<u>0.26</u>	<u>0.53</u>	<u>6.02</u>	0.27	<u>0.59</u>	<u>6.02</u>	0.29	0.48	<u>6.89</u>
Emu2 [50]	<u>0.26</u>	0.62	5.86	0.25	0.59	5.66	<u>0.28</u>	0.59	6.67	0.26	0.62	5.84	0.26	0.62	5.69	<u>0.28</u>	0.55	6.70
Ours	0.28	0.58	<u>6.37</u>	0.28	0.54	6.07	<u>0.28</u>	0.51	6.78	0.28	0.54	6.30	0.28	0.55	6.08	<u>0.28</u>	<u>0.42</u>	6.79

Table 4. Quantitative comparison over FG-TI2I single-entry, with another set of quantitative metrics, where we abbreviate Object, Texture, Action, Background into OBJ, TEX, ACT, BG.

C.5. Inversion ϵ for References.

Since the CTS module relies on learning visual information at the same timestep, we introduce noise into clean reference images to match the noise level of the initial latent throughout the generation process. By default, we set ϵ to standard Gaussian noise, following a process similar to SDEdit [37], due to its computational efficiency. As shown in ??, replacing ϵ with an inversion algorithm, such as RF-inversion [45], does not significantly alter the generation outcome.

C.6. Empty String for References.

Following the setup of previous reference-supported T2I models [8, 18, 46, 50, 53, 57], which utilize an additional reference prompt to stabilize the generation process, we adopt the same approach in our proposed TF-TI2I and FG-TI2I. However, as shown in ??, this design may not be necessary for TF-TI2I. When using an empty string as the reference prompt, the generated output still adheres to both the textual instruction and the reference image.

C.7. CTS versus Shared Attention.

The Shared Attention module is a widely used modification for UNet-based T2I models [42, 44], where the self-attention keys and values from the reference image are concatenated into the generation process. This enables the generated output to share visual elements with the reference image. However, due to the fully transformer-based architecture of MM-DiT, concatenating vision tokens in this manner dilute the influence of textual tokens, causing the output to disregard textual instructions and adhere primarily to the reference image, as shown in Fig. 10a. In contrast, CTS mitigates this issue by encoding reference images into contextual tokens, enabling the output to align with both the textual prompt and the reference image, as demonstrated in Fig. 10b.

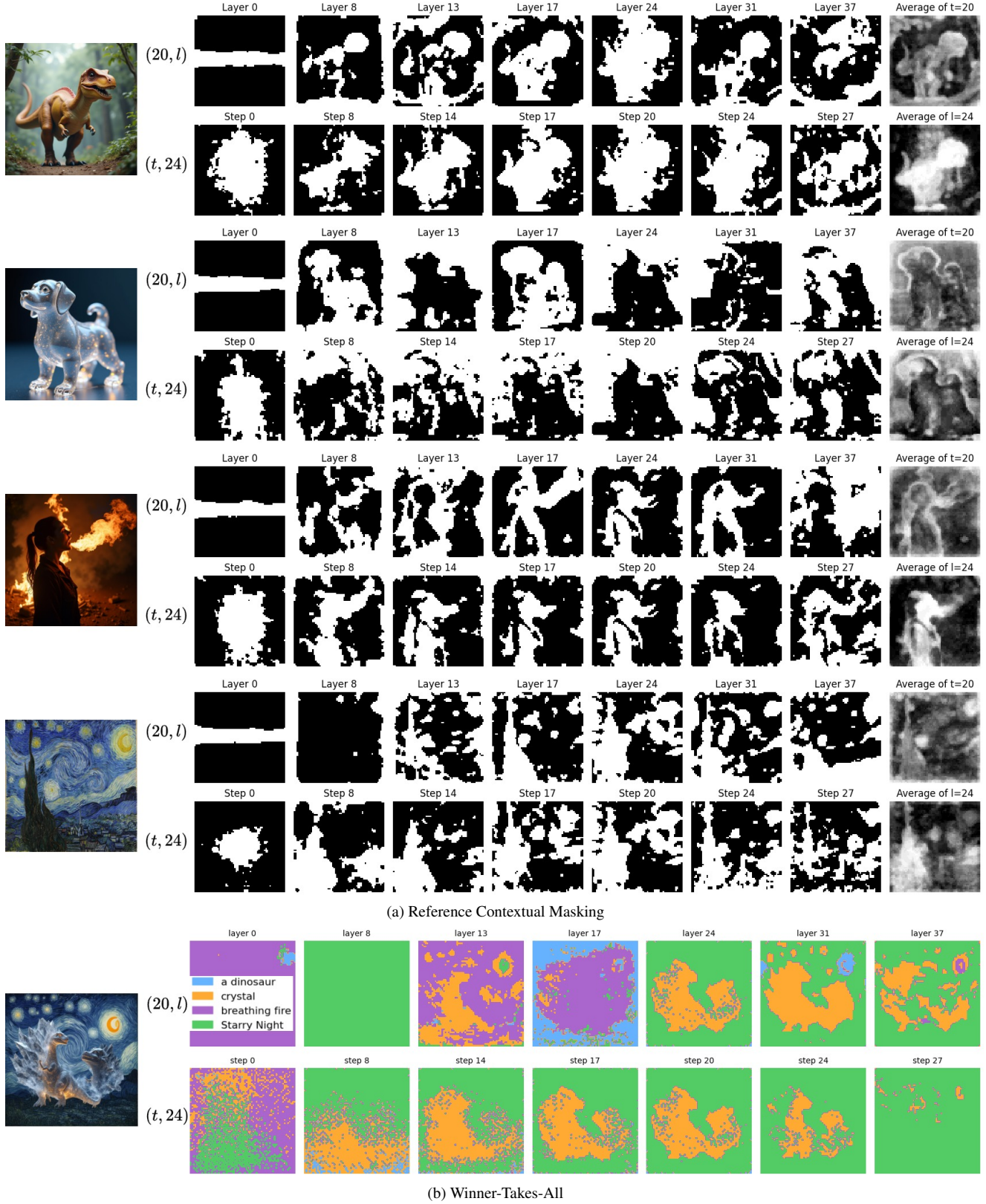


Figure 17. Illustration of M_r^{RCM} and M^{WTA} at different layers and timesteps (Note that during generation, we only activate RCM at the late layer instead of all layers shown here)