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

Teng-Fang Hsiao, Bo-Kai Ruan, Yi-Lun Wu, Tzu-Ling Lin, Hong-Han Shuai,
National Yang Ming Chiao Tung University
{tfhsiao.ee13, bkruan.ee11, yilun.ee08, tzulinglin.11, hhshuai}@nycu.edu.tw

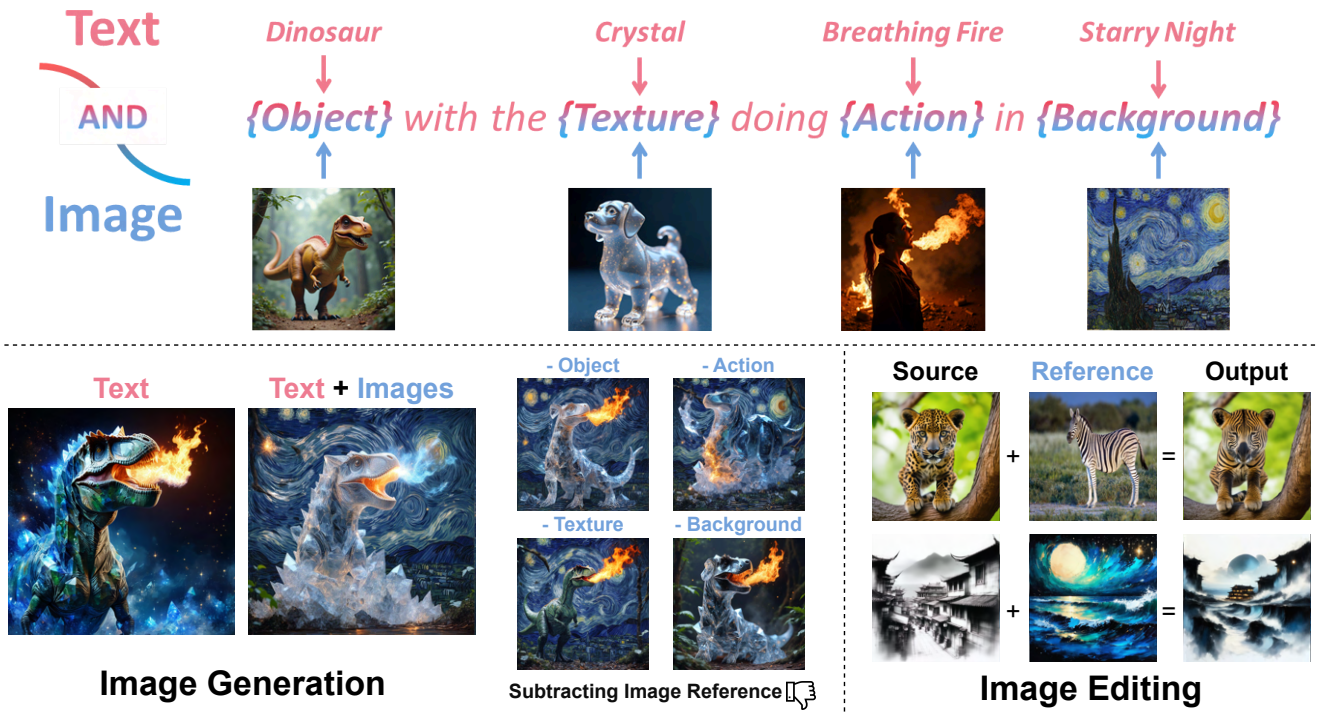


Figure 1. Illustration of our proposed TF-TI2I, designed to leverage multiple image references for generation and editing without additional training. The importance of image references is illustrated in the middle part, where excluding any reference significantly alters the output.

Abstract

Text-and-Image-To-Image (TI2I), an extension of Text-To-Image (T2I), integrates image inputs with textual instructions to enhance image generation. Existing methods often partially utilize image inputs, focusing on specific elements like objects or styles, or they experience a decline in generation quality with complex, multi-image instructions. To overcome these challenges, we introduce Training-Free Text-and-Image-to-Image (TF-TI2I), which adapts cutting-edge T2I models such as SD3 without the need for additional training. Our method capitalizes on the MM-DiT architecture, in which we point out that textual tokens can implicitly learn visual information from vision tokens. We enhance this interaction by extracting a condensed visual representation from reference images, facilitating selective information sharing through Reference Contextual Masking—this technique confines the usage of contextual tokens to instruction-relevant visual information. Additionally, our Winner-Takes-All module mitigates distribution shifts by prioritizing the most pertinent references for each vision token. Addressing the gap in TI2I evaluation, we also introduce the FG-TI2I Bench, a comprehensive benchmark tailored for TI2I and compatible with existing T2I methods. Our approach shows robust performance across various benchmarks, confirming its effectiveness in handling complex image-generation tasks.

1. Introduction

"Don't tell me the moon is shining; show me the glint of light on broken glass."

— Anton Chekhov (1860-1904)

Recent advances in text-to-image (T2I) generation empower users to produce compelling visual artwork using only textual prompts [27, 36, 42, 44, 58]. However, text descriptions alone are insufficient for conveying the nuanced spatial relationships, specific object characteristics, and stylistic details necessary to fulfill users' intentions. To address this, various approaches have integrated additional visual guidance into T2I pipelines, leveraging conditional-control methods [32, 62, 64], inpainting techniques [24, 25, 59, 65], style transfer [13, 15, 19, 23], and customized image synthesis [11, 18, 46, 55, 61].

Despite their effectiveness, current methods are typically constrained to a single functionality. For example, conditional-control, inpainting, and style-transfer approaches focus on one aspect of visual guidance. Similarly, customization methods excel at object-based references but struggle with other factors e.g., texture or background. One solution is the use of multimodal large language models (MLLMs) in image generation [40, 50], which can handle both visual and textual inputs. However, MLLM-based generation still lags behind specialized T2I models in terms of output quality and resolution, largely due to the limited availability of text-and-image-to-image training data. For instance, KOSMOS-G [40], utilizing BLIP2 [29] and CLIP-Seg [29] to autonomously identify objects as image inputs, remains restricted to object-level conditioning without providing finer control over texture or background details.

To enhance performance without larger training data, we explore MM-DiT—a fully transformer-based multimodal architecture used in state-of-the-art T2I models such as SD3 [17] and FLUX [27]. We contend that MM-DiT's multimodal attention demonstrates cross-modal understanding [16, 20, 51], which allows textual tokens to naturally integrate **implicit visual information** from image latent during inference. Unlike methods that require additional training to establish image-reference correlations, we propose **Training-Free Text-and-Image-to-Image (TF-TI2I)** by (1) harnessing the inherent ability of MM-DiT to infuse textual tokens with implicit visual context during generation, and (2) enabling these tokens to be shared across different inputs as contextual tokens. As our approach relies only on the model's intrinsic properties, we can extend T2I models with TI2I capabilities without fine-tuning.

However, challenges arise as the number of image references increases. Firstly, mutual interference between different references can lead to unintended visual blending. For instance, as shown in Fig. 1, when generating a background following reference 4 (Starry Night), elements from the backgrounds of references 1, 2, and 3 might inadvertently be incorporated due to insufficient differentiation among the references. To address the mutual interference among multiple references, we propose **References Contextual Masking (RCM)**. This technique limits the visual information learned from contextual tokens by focusing exclusively on vision tokens that are more related to the given instructions. This ensures that only the relevant visual information of a specified reference is utilized in the image generation process. By doing so, RCM effectively reduces the undesired blending of visual features from different references, maintaining clarity and fidelity to the original input conditions.

Secondly, as a training-free approach, TF-TI2I is also susceptible to distribution shifts that can degrade generation quality, a challenge noted in related techniques [13, 15, 35]. To counteract this, we introduce the **Winner-Takes-All** (WTA) module. This module assigns each vision token exclusively to one reference at a time, ensuring that each token robustly represents the visual characteristics of its assigned references. Concurrently, the WTA module facilitates the incorporation of visual information from other references through attention between the visual tokens. This dual mechanism not only maintains each reference's integrity within the output but also enriches the visual details, resulting in high-quality images that adhere to both intended references and textual instruction.

Finally, recognizing the need for a comprehensive evaluation framework for both TI2I models [40, 50] and visually guided T2I models [3, 8, 22, 57, 62], we introduce the Fine-Grained TI2I Benchmark (FG-TI2I Bench) tailored for general image generation scenarios. Drawing inspiration from EditBench [54], which assesses detailed object attributes like material, color, and shape, our benchmark structures prompt instructions around four key components and encompass 6 fundamental TI2I sub-tasks. These components are described using either text or images, making our framework compatible with visually guided T2I methods. Our approach achieves state-of-the-art performance on 12 out of 18 evaluation metrics across general-purpose TI2I tasks and remains highly competitive in task-specific benchmarks such as DreamBench [46] for customization and Wild-TI2I [53] for editing scenarios.

Our contributions are summarized as follows:

- We discover the implicit-context learning capability of MM-DiT and design TF-TI2I, a training-free approach, to generate images from different visual references and text prompts. This further augments the existing T2I model with TI2I capability.
- We develop *Reference Contextual Masking* and *Winner Takes All* modules to mitigate multi-reference problems, leading to more visually satisfactory results.
- We introduce FG-TI2I Bench, a comprehensive benchmark designed to evaluate a variety of visually guided T2I application *e.g.* customization and style transfer. Our

approach achieves state-of-the-art results in 12 out of 18 evaluation metrics across FG-TI2I and also remains competitive in DreamBench [46] and Wild-TI2I [53].

2. Related Work

2.1. TI2I with single aspect of reference

To expand the creativity of pretrained T2I models with TI2I capability, finetuning-based approaches [12, 18, 19, 41, 46, 61], such as DreamBooth [46], encode reference information into model weights through a finetuning process. While effective, these methods are often limited by computational costs and the need for a substantial number of training references. In contrast, training-free methods [8, 13, 15, 23, 34] provide efficiency but concentrate on a single aspect of reference images, such as style in StyleID [13] and ZSTAR [15], or objects in TF-ICON [34] and TF-GPH [23]. Our approach, which integrates the *Reference Context Control* and *Winner-Takes-All* modules, efficiently combines various aspects of references—including objects, textures, actions, and backgrounds—enabling a more comprehensive and flexible generation capability.

2.2. TI2I with multiple aspects of references

Models trained from scratch with both text and image inputs can better utilize diverse image references for TI2I tasks. These models can be broadly categorized into retrieval-based [1, 9, 48] and auto-regressive-based approaches [40, 50, 57]. Retrieval-based methods, such as KNN-D [48] and RDM [1], follow the concept of RAG [28] by retrieving relevant images from a database to provide visual prior for generation. However, their capability is dependent on the retrieved references, which limits the model's ability to synthesize novel content beyond what is provided. In contrast, auto-regressive-based methods like Emu2 [50] and Kosmos-G [40] face constraints in generation resolution and quality due to the limited availability of multimodal training pairs. Our proposed TF-TI2I leverages a pretrained T2I model, enhancing it with reference support for controllability, and achieves superior performance over existing methods without additional training costs.

3. Preliminary

Our work leverages the Multimodal Attention (MMA) blocks inherent in MM-DiT-based T2I models [17, 27]. Given a textual prompt P and a reference image I, the prompt and image are first encoded into latent features, denoted as $\tau_P^0 = \mathcal{E}_P(P) \in \mathbb{R}^{n_P \times d}$ and $\tau_I^0 = \mathcal{E}_I(I) \in \mathbb{R}^{n_I \times d}$, where $\mathcal{E}_P$ and $\mathcal{E}_I$ represent the text and image encoders, respectively. For the latent features within the MM-DiT backbone, we use the superscript $l \in 0, 1, \ldots, L$ to indicate the layer, e.g., τ_I^l and τ_P^l. For simplicity, we omit the time step notation, as it remains unchanged in all equations, i.e., $\tau_I^l = \tau_I^l(t)$. The multimodal attention mechanism, after disregarding the head dimension, is expressed as follows:

$$\{Q_I, K_I, V_I\} = \tau_I^l W_I^{\{Q,K,V\}}, \tag{1}$$

$$\{Q_P, K_P, V_P\} = \tau_P^l W_P^{\{Q,K,V\}}, \tag{2}$$

$$\mathrm{A}(\tau_I^l, \tau_P^l, M) = \mathrm{softmax}\big((QK^T + M)/\sqrt{d}\big)V, \tag{3}$$

where $Q = [Q_I; Q_P]$, $K = [K_I; K_P]$, and $V = [V_I; V_P]$, with $[\cdot;\cdot]$ denoting concatenation along the first dimension. Here, W represents the projection matrix, and M denotes the attention mask. A key distinction of MM-DiT from previous self- and cross-attention [42, 44] is that it updates not only the visual tokens τ_I^l but also refines the textual tokens τ_P^l. This dual-update process is formulated as:

$$[\tau_I^{l+1}; \tau_P^{l+1}] = [\tau_I^l; \tau_P^l] + \mathrm{A}(\tau_I^l, \tau_P^l, M). \tag{4}$$

Throughout this paper, we use SD3.5-large [17] as our default backbone, which consists of 38 MM-DiT layers, with $n_I = 4096$ and $n_P = 333$. Additionally, to facilitate discussions on the diffusion process in subsequent sections, we incorporate the noise-adding process, defined as

$$\tau_I^0(t) = \tau_I^0 + \epsilon(\tau_I^0, t), \tag{5}$$

where ϵ represents either standard Gaussian noise or inversion-based noise predictors [33, 45, 47, 49].

4. Method

In this section, we first examine the implicit-learning capability of textual tokens. Next, we demonstrate that sharing these textual tokens enables the manipulation of visual content through textual representations. Finally, we propose References Contextual Masking and Winner Takes All module to extract and assign instruction-relevant information from excessively rich visual content. The overall architecture can be found in Fig. 2.

4.1. Implicit-Context Learning from MM-DiT

The key distinction between MM-DiT and conventional T2I models [5–7, 30, 42, 44], lies in how textual information is treated. While traditional models use textual instruction as a fixed condition, MM-DiT updates the textual conditions τ_P^l at each layer. This characteristic motivates us to investigate the problem–**Is visual context learned implicitly by the textual tokens τ_P^l in the MM-DiT?** We use the term "implicit" to emphasize that textual tokens gradually aggregate visual context from visual tokens through multi-modal attention. Consequently, these textual tokens can implicitly represent the visual reference.

To empirically test whether textual tokens from different prompts encode similar features when conditioned on the same image, we sample 50 prompts structured as "{*Object*}

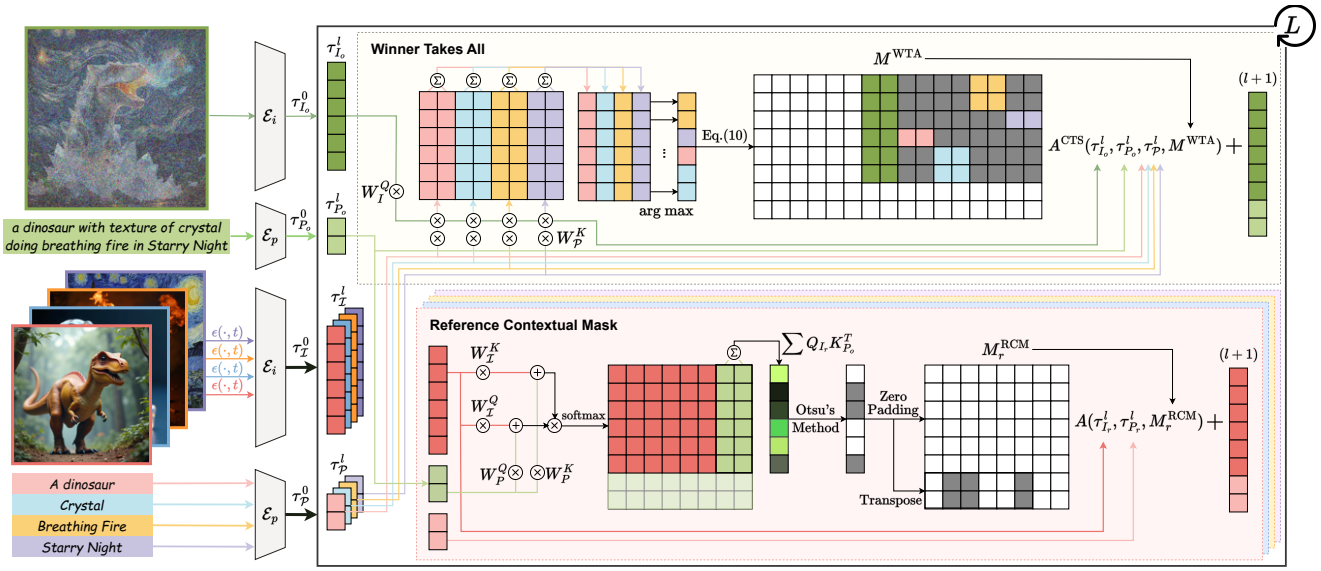


Figure 2. Illustration of the TF-TI2I pipeline. TF-TI2I leverages contextual visual information learned from textual tokens. Through sharing the contextual token by concatenating $\tau_{\mathcal{P}}^l$ to the upper block (Sec. 4.3.1), we achieve prompt-following while maintaining reference-aligned results. Additionally, we incorporate Reference Contextual Masking (Sec. 4.3.2) to mitigate mutual interference between references and employ the Winner-Takes-All module (Sec. 4.3.3) to minimize distribution shifts in multi-reference scenarios.

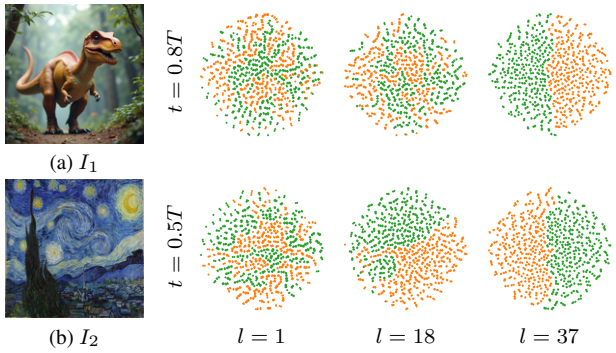


Figure 3. The t-SNE visualization of $\{\tau_{P_1}^l \mid P_1 \in \mathbb{P}\}$ and $\{\tau_{P_2}^l \mid P_2 \in \mathbb{P}\}$ at different timesteps and layers indices. The clusters form in deeper layers and later timesteps.

with the texture of Texture doing {Action} in the background of {Background}." We denote these prompts as $\mathbb{P}$ and consider two different images, I_1 (Fig. 3a) and I_2 (Fig. 3b). We then forward these prompt tokens, $\tau_{P_1}^0 = \mathcal{E}_P(P_1)$ and $\tau_{P_2}^0 = \mathcal{E}_P(P_2)$ for all $P_1, P_2 \in \mathbb{P}$, along with the vision tokens $\tau_{I_1}^l$ and $\tau_{I_2}^l$. Subsequently, we extract $\tau_{P_1}^l$ and $\tau_{P_2}^l$ from the T2I model to perform clustering visualization for different layer index l. Our hypothesis is that **if textual tokens learn from vision tokens during the forward process, the features distributions, $\{\tau_{P_1}^l \mid P_1 \in \mathbb{P}\}$ and $\{\tau_{P_2}^l \mid P_2 \in \mathbb{P}\}$, should form distinct clusters due to the different input visual tokens**. As illustrated in the t-SNE visualization in Fig. 3, we observe that τ_P^l exhibits stronger clustering effects in deeper layers and at later stages of generation, where visual features become more distinct from random noise. These findings highlight the implicit context-learning capability of the textual tokens τ_P.

(a) I_1 from seed 1 (b) I_2 from seed 2 (c) Replacing τ_{P_1}

Figure 4. The illustration of replacing contextual tokens between different images. As shown in the Fig. 4c, we can successfully transfer the visual information of Fig. 4b to Fig. 4a.

In the following, we refer to these textual tokens with implicit-context visual information as *contextual tokens* to distinguish them from the so-called textual tokens, which only contain textual information as described in prior work [17, 42, 44]. Our next question is: **Can the contextual tokens utilized by different set of vision tokens?** To test this possibility, we create a toy example where we use a single prompt, *"a princess in the dress"*, with two different seeds. Due to the difference in the initial noise of the vision tokens, the generated images are different in both details and layout, as shown in Fig. 4a and Fig. 4b. We then replace their contextual token by replacing $\tau_{P_1}^l$ with $\tau_{P_2}^l$ in each layer for all timesteps. The result is shown in Fig. 4c,

where the image structure still resembles I_1, but features like dress color shift toward I_2. This example supports our hypothesis that visual information in contextual tokens can be utilized by other set of vision tokens.

4.2. Training-Free Text-and-Image-to-Image

Building upon the previous findings that contextual tokens not only encapsulate visual information but can also be leveraged by other vision tokens, we propose TF-TI2I, a Training-Free Text-and-Image-to-Image pipeline, as shown in Fig. 2. By modifying pre-trained T2I models [17, 27], we enable the utilization of multiple images as reference inputs while maintaining high-quality generation results.

4.3. TF-TI2I Overview

The objective of TF-TI2I is to generate an image I_o based on the prompt instruction P_o and given visual references $\mathcal{I} \coloneqq \{I_1, I_2, \dots, I_R\}$. At each timestep t, we start with an initial vision token $\tau^0_{I_o}$, which can be either a random noise for image generation or a noisy image for editing. Along with this, we incorporate the textual token $\tau^0_{P_o}$ derived from the input prompt P_o. We then introduce reference vision tokens $\{\tau^0_{I_1}, \tau^0_{I_2}, \dots, \tau^0_{I_R}\}$, where $\tau^0_{I_r}(t) = \tau^0_{I_r} + \epsilon(\tau^0_{I_r}, t)$, with $\tau^0_{I_r} = \mathcal{E}_I(I_r)$ and $I_r \in \mathcal{I}$. Additionally, TF-TI2I requires reference prompts $\mathcal{P} \coloneqq \{P_1, P_2, \dots, P_R\}$, which are used to initialize the textual tokens $\{\tau^0_{P_1}, \tau^0_{P_2}, \dots, \tau^0_{P_R}\}$, where $\tau^0_{P_r} = \mathcal{E}_P(P_r)$ and $P_r \in \mathcal{P}$.

4.3.1. Contextual Tokens Sharing (CTS)

Extending from the findings that contextual tokens can be utilized by other vision tokens, we design a contextual token-sharing module to replace the original MMA layer and leverage the contextual information from image references. The idea of CTS is akin to the share-attention module [3, 13, 22, 23], which concatenates the vision tokens from different images together. By contrast, we concatenate the key and value of $\tau^l_{I_o}$ with contextual tokens $\tau^l_{P_r}$ to obtain the new input prompt tokens $\tau^l_{\mathcal{P}} \coloneqq [\tau^l_{P_o}; \tau^l_{P_1}; \tau^l_{P_2}; \dots; \tau^l_{P_R}]$. With $\tau^l_{\mathcal{P}}$, the CTS Attention replace Eq. (3) as follow:

$$\mathrm{A}^{\mathrm{CTS}}(\tau^l_{I_o}, \tau^l_{P_o}, \tau^l_{\mathcal{P}}, M) = \mathrm{softmax}\big((QK^T + M)/\sqrt{d}\big)V, \quad (6)$$

where $Q = [Q_{I_o}; Q_{P_o}]$, $K = [K_{I_o}; K_{P_o}; K_{\mathcal{P}}]$ and $V = [V_{I_P}; V_{P_o}; V_{\mathcal{P}}]$. By concatenating contextual tokens instead of vision tokens, we solve the main restriction of the share attention module — the computational overheads, R references require $(2R-1)$ times computation and memory sources for attention operation, while concatenating contextual tokens only require $((R-1) + (n_P/n_I)R)$ nearly half times of sources ($n_P/n_I = 0.08$ for SD3 [17]), allowing us to incorporate more references.

4.3.2. Reference Contextual Masking (RCM)

Along with the increase of references, the conflict of references grows severe. For instance, in Fig. 1, we aim to generate the background of a starry night instead of the jungle background from the *"dinosaur"* reference or the dark night background from the *"breathing fire"* reference. Thus, we design **References Contextual Masking** (RCM) to select only the vision tokens we need from the given reference instead of the whole image.

Inspired by prior works [2, 4, 14] that show cross-attention maps can already play a similar function as semantic matching, we extract the semantic connection between the instruction and each visual token. For the visual tokens $\tau^l_{I_r}$ from the given reference I_r, we utilize $\tau^l_{P_o}$ to measure the corresponding semantic connection of each token by computing the attention score. We then compare the summation of the attention scores between the query of $\tau^l_{I_r}$ and the key of $\tau^l_{P_o}$ on each visual token and retain only the salient tokens via a binarization operation. The resulting contextual mask for i^{th} vision token of r^{th} reference is:

$$M^{\mathrm{RCM}}_{i,r} = \mathrm{BI}\left(\sum_{j=1}^{n_P} \mathrm{softmax}\big(Q^l_{I_r}[K^l_{I_r}; K^l_{P_o}]^T\big)_{i,n_I+j-1}\right), \quad (7)$$

where "BI" denotes the binarize algorithm. Specifically, we utilized Otsu Algorithm [39] to detach the foreground from the background. We gather $M^{\mathrm{RCM}}_{i,r}$ for each vision token to form the reference contextual mask M^{RCM}_r as an attention mask during the forward process of $\tau^l_{I_r}$ and $\tau^l_{P_r}$:

$$[\tau^{l+1}_{I_r}; \tau^{l+1}_{P_r}] = [\tau^l_{I_r}; \tau^l_{P_r}] + \mathrm{A}(\tau^l_{I_r}, \tau^l_{P_r}, M^{\mathrm{RCM}}_r), \quad (8)$$

as shown in Fig. 2. By summing and binarizing the attention score into a mask, we can restrict the contextual tokens to only learn contextual information from specific vision tokens, reducing redundant information from reference.

4.3.3. Winner Takes All (WTA)

As a training-free modification applied to a pre-trained model, our approach inevitably faces distribution shifts, leading to unsatisfactory generation results, as suggested in [13, 15, 35]. A similar phenomenon is also observed in CTS as discussed in Appendix C.1. These issues become more pronounced as the number of reference images increases.

To address this, we propose a novel **Winner Takes All** (WTA) strategy to mitigate conflicts between multiple references. We assume that, for each vision token, only one or two references are needed at a time. For instance, background references are not necessary at every step and layer when generating an object. Leveraging the semantic correlation capability of our T2I model, we first measure the attention score of references to a given vision token and selectively keep only the contextual tokens with highest attention score to minimize distribution disturbances. This is

formulated as:

$$s_{i,r} = \sum_{j=1}^{n_P} \text{softmax}\big(Q_{I_o} K_{P_r}^T\big)_{i,j}. \quad (9)$$

To perform the winner-takes-all criteria, we set $r_i^* = \arg\max_r s_{i,r}$. As such, the WTA mask is applied as:

$$M_{i,j}^{\text{WTA}} = \begin{cases} -\infty, & (n_I + n_P) < j \leq n_I + n_P \cdot r_i^* \\ -\infty, & n_I + n_P + n_P \cdot r_i^* \leq j \\ 0, & \text{otherwise} \end{cases}, \quad (10)$$

where i denotes the i^{th} vision token. We then incorporate WTA mask M^{WTA} into contextual sharing Eq. (6):

$$[\tau_{I_o}^{l+1}; \tau_{P_o}^{l+1}] = [\tau_{I_o}^{l}; \tau_{P_o}^{l}] + \text{A}^{\text{CTS}}(\tau_{I_o}^{l}, \tau_{P_o}^{l}, \tau_{\mathcal{P}}^{l}, M^{\text{WTA}}). \quad (11)$$

The illustration can be found in the upper part of Fig. 2. The WTA module may assign different contextual tokens to vision tokens across different layers, effectively mixing multiple references.

5. Experiment

As a unified TI2I framework, our proposed TF-TI2I shares similarities with the well-known Emu2 [50] and supports various downstream tasks, including object customization [9, 10, 18, 40, 46, 57], consistent synthesis [3, 22], style transfer [13, 15, 19, 23] and image editing [14, 21, 31, 37, 53]. To evaluate its customization and editing capabilities, we use DreamBench [46] (750 instances) and Wild-TI2I [53] (148 instances). Additionally, we introduce FG-TI2I for fine-grained evaluation of the TI2I task. Since our evaluation involves both textual and visual inputs, we adopt a diverse set of metrics for a comprehensive assessment:

- **Prompt following**: CLIP-Text [43] (CP), CLIP-Directional Score [26] (CDS), Image Reward [60] (IR), Human Preference Score [56] (HPS)
- **Reference alignment**: CLIP-Image [43] (CP-I), CLIP-Directional Score [26] (CDS), DINO Similarity [38] (DI), LPIPS [63] (LP)
- **Generation quality**: Aesthetic Score [44] (AS), Image Reward [60] (IR), Human Preference Score [56] (HPS)

More experimental results are provided in Appendix B.

5.1. Fine-Grained TI2I Benchmark

Although benchmarks [21, 23, 25, 34, 46, 54] exist for evaluating single aspect of image input, such as conditioning on object, background, and style, the evaluation of unified TI2I—where a single instruction incorporates multiple aspects of image references—remains relatively unexplored. To address this gap, we propose the Fine-Grained TI2I (FG-TI2I) benchmark, the first image generation benchmark that integrates multiple reference aspects within a single instruction. FG-TI2I adopts a unified representation for common image generation tasks, utilizing a single instruction template with four aspects of reference: *"{Object} with the texture of {Texture} doing {Action} in the background of {Background}"*. As illustrated in Fig. 5, we categorize the sub-tasks based on the number of image references. For each sub-task, we generate 100 prompts using ChatGPT-4o and create the corresponding image references using Flux 1.0 dev [27]. By introducing image references at different aspects, we can systematically evaluate a model's capability in various downstream tasks.

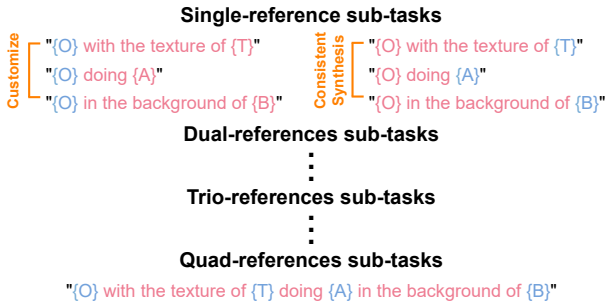


Figure 5. Illustration of sub-tasks in FG-TI2I, where we abbreviate Object, Texture, Action, and Background into O, T, A, and B, respectively. The text-only input is denoted with red, and image-support input is denoted with blue. The input sub-tasks are categorized by the number of image references.

5.2. Qualitative Results

For multiple-reference test cases in Fig. 6, TF-TI2I achieves better compliance with all given references and higher image quality compared to the previous TI2I method, Emu2 [50]. For instance, in row 1, although Emu2 produces a reference-like background, the other input components, *i.e.*, the given object, texture, and action, are missing in the output. Similarly, while Emu2 successfully generates the given object reference in cases 2, 3, and 4, the texture, action, and background are not preserved in the generated results. These findings confirm a key limitation of Emu2: constrained by its training design, Emu2 excels in object customization but struggles to generalize well to the more creative and complicated TI2I instructions introduced in our proposed FG-TI2I benchmark.

Notably, even when compared to OmniGen [57], StyleAlign [22], and MasaCtrl [3], which are designed for single-reference generation, our proposed method still demonstrates superior performance, especially in texture-related and action-related prompts. This is achieved through our proposed contextual token-sharing mechanism, which enables TF-TI2I to achieve a better balance between textual instructions and image references by leveraging image references as textual tokens. As a result, TF-TI2I generates more texture-aligned outputs, as shown in rows 1 and 4 on

	OBJ x TEX			OBJ x ACT			OBJ x BG			OBJ x TEX			OBJ x ACT			OBJ x BG		
	CP	DI	IR	CP	DI	IR	CP	DI	IR	CP	DI	IR	CP	DI	IR	CP	DI	IR
MasaCtrl [3]	26.8	**96.4**	-0.26	25.4	**96.5**	-0.92	26.6	**95.7**	-0.31	26.5	**91.7**	0.10	27.0	**88.7**	0.09	26.9	**97.3**	0.14
StyleAlign [22]	27.3	94.3	0.10	28.3	92.9	0.28	28.3	89.9	0.56	27.7	88.9	0.33	29.3	87.7	0.88	29.5	89.6	0.82
OmniGen [57]	28.2	84.8	0.36	28.6	82.7	0.62	29.2	81.2	1.09	28.4	75.8	0.40	28.7	76.1	0.85	29.2	85.5	1.12
Emu2 [50]	28.2	71.2	0.10	27.5	75.5	0.34	28.2	72.0	0.66	28.5	63.8	0.10	27.6	62.9	0.36	28.2	75.7	0.68
Ours	**29.3**	87.5	**0.70**	**30.3**	87.1	**1.21**	**29.3**	84.1	**1.14**	**30.1**	72.8	**0.67**	**30.7**	75.8	**1.30**	**30.1**	84.4	**1.26**

Table 1. Quantitative comparison over FG-TI2I single-entry, where we respectively abbreviate Object, Texture, Action, and Background into OBJ, TEX, ACT, and BG. **Bold** highlights the best result, and underlines mark the second-best.

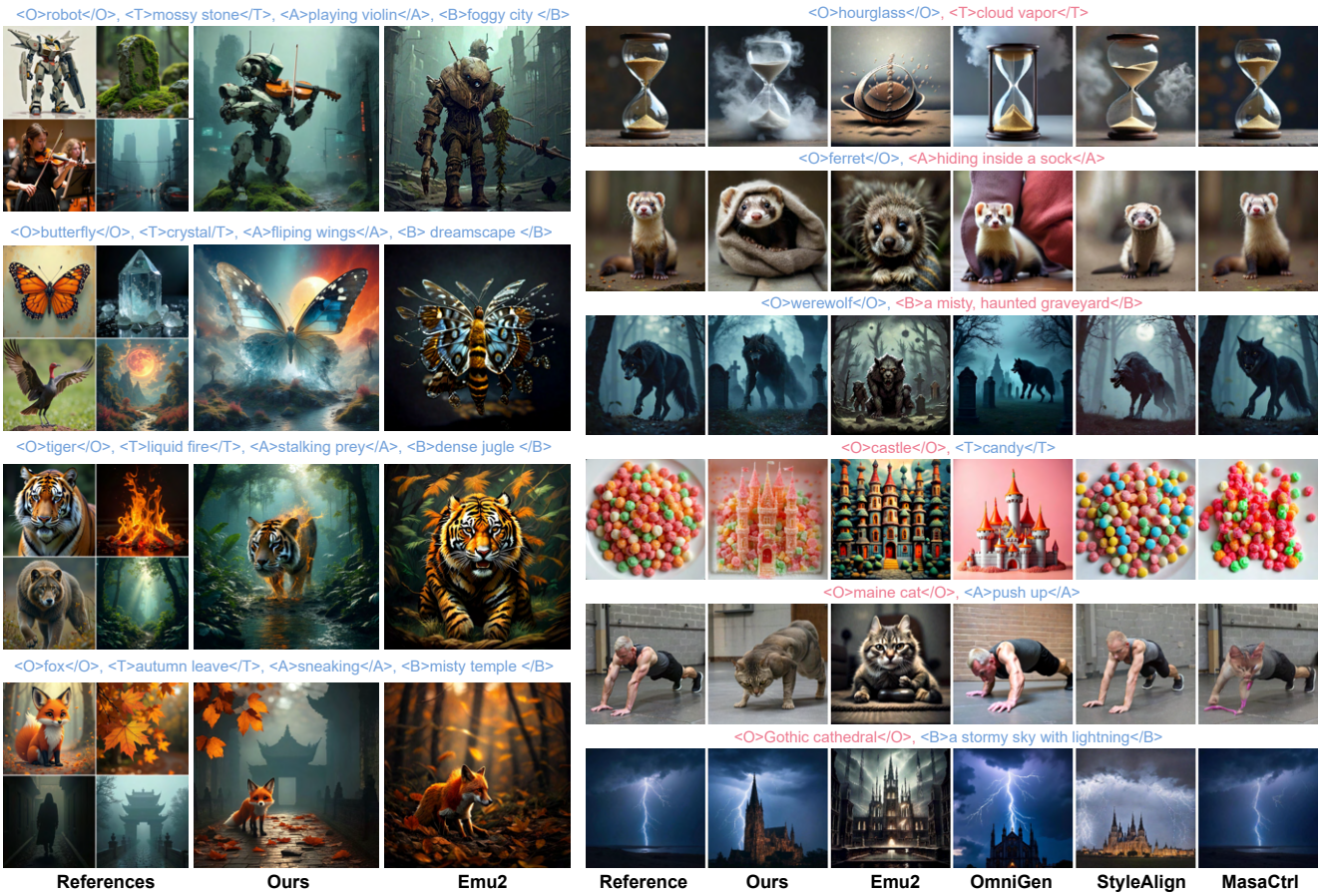


Figure 6. Qualitative comparison of Quad-references sub-tasks (left) and Single-reference sub-tasks (right) in FG-TI2I. The input Object, Texture, Action, and Background—are denoted as O, T, A, and B. We use red for text-only input and blue for reference-supported input.

	CP-T	CP-I
Tex-Inv [18]	25.5	78.0
DB [46]	30.5	80.3
ReImagen [9]	27.0	74.0
SuTI [10]	30.4	81.9
Kosmos-G [40]	28.7	**84.7**
OmniGen [57]	31.5	80.1
Ours	**33.1**	79.1

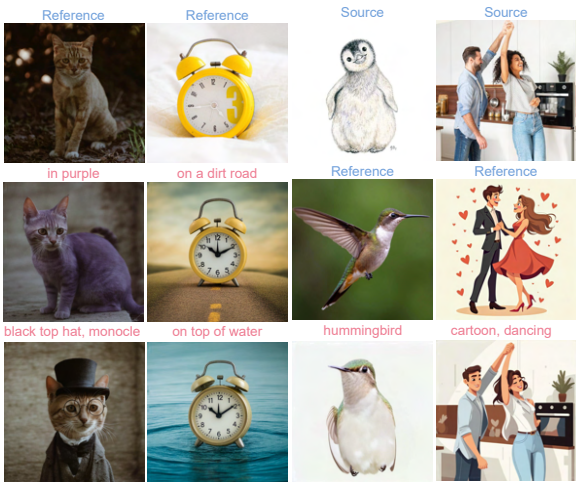
Table 2. Quantitative comparison on DreamBench.

	CP	CDS
SDEdit [37]	27.5	0.122
DiffEdit [14]	26.3	0.088
P2P [21]	28.4	0.193
PnP [53]	28.5	0.202
MasaCtrl [3]	29.3	0.209
TIS [31]	29.0	**0.210**
Ours	**30.4**	0.115

Table 3. Quantitative comparison on Wild-TI2I

the right side of Fig. 6, successfully synthesizing the cloud vapor hourglass based on textual instructions and the candy castle based on visual references. Additionally, in rows 2 and 5, TF-TI2I better aligns with action-related instructions, effectively following both textual and visual guidance.

Furthermore, TF-TI2I is readily compatible with customization and editing tasks without requiring any modifications, as shown in Fig. 7. It seamlessly enables object modifications (*i.e.*, changing the color and outfit of a cat) and object transfer to different backgrounds.



(a) DreamBench [46] (b) Wild-TI2I [53]

Figure 7. Qualitative results of TF-TI2I on Customization [46] and Editing [53]. For Wild-TI2I, we first generate a reference for each sample to support text-only editing instructions.

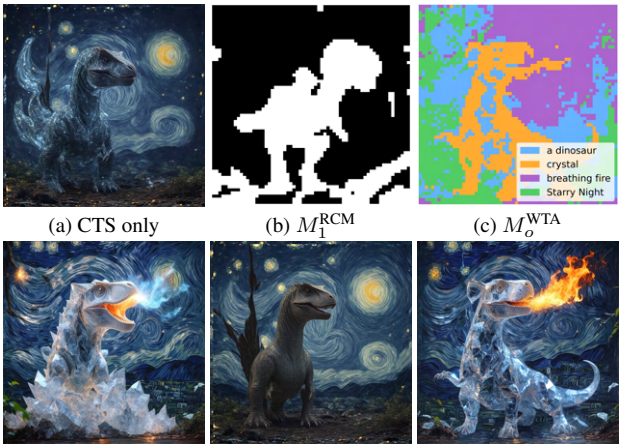
5.3. Quantitative Results

Our quantitative comparison in Tab. 1 primarily focuses on single-reference evaluations, as most baseline methods are designed for such settings. Our proposed TF-TI2I framework demonstrates superior performance in both prompt-following and image quality metrics. This advantage stems from our proposed CTS strategy, which integrates additional contextual tokens into the generation pipeline. These contextual tokens, functionally similar to textual tokens but enriched with visual information, introduce minimal disturbance to the original T2I model. Consequently, TF-TI2I achieves higher prompt alignment and better image quality.

However, as a trade-off, our method exhibits slightly lower performance in DINO Similarity (DI) compared to training-based approaches like OmniGen [57] and inversion-based methods such as StyleAlign [22] and MasaCtrl [3]. This limitation arises because TF-TI2I is not designed for direct reference copying but instead encodes references into high-level semantics, thereby sacrificing low-level details but ensuring instruction compliance.

A similar trend can be observed in the reported performance in Tab. 2 and Tab. 3 with TF-TI2I achieves state-of-the-art (SOTA) results in CLIP-Text (CP) but does not surpass all the existing methods in terms of CLIP-Image (CP-I) and CLIP-Directional Score (CDS) as our approach might alter the reference to make the output compatible with the new scene. Besides, TF-TI2I maintains competitive reference-alignment performance—outperforming the TI2I model EMU2 across all FG-TI2I sub-tasks, surpassing Textual Inversion [18] by 1%, and trailing the SOTA customization method OmniGen by only 7%.

5.4. Ablation study



(a) CTS only (b) M_1^{RCM} (c) M_o^{WTA}

(d) +RCM+WTA (e) +RCM (f) +WTA

Figure 8. Ablation study of TF-TI2I, using Fig. 1 as example.

RCM. By restricting the learning of contextual tokens to only instruction-related vision tokens using M_r^{RCM}, we can ease mutual interference between different references, as shown in Fig. 8b, where M_1^{RCM} is the contextual mask for *the dinosaur*. Comparing Fig. 8e to Fig. 8a, both the generated background *Starry Night* and the generated object *a dinosaur* exhibit greater similarity to their respective input references, demonstrating the effectiveness of RCM in reducing mutual disturbance between references.

WTA. Designed to mitigate distribution shifts caused by additional reference inputs, it achieves this by measuring the saliency score of each reference and generating corresponding reference assignments for each vision token, as shown in Fig. 8c. This strategy ensures both reference adherence and visually satisfactory results. As demonstrated in Fig. 8d and Fig. 8f, the generated images exhibit richer details and colors compared to Fig. 8a and Fig. 8e.

6. Conclusion

In this paper, we investigate the implcit-context learning capability of MM-DiT and leverage this insight to develop TF-TI2I, a training-free text-and-image-to-image generation pipeline. Our approach surpasses previous methods in effectively integrating multiple types of image references. To overcome the limitations of existing benchmarks, we introduce FG-TI2I, a more comprehensive evaluation framework for TI2I tasks. Experimental results across various benchmarks demonstrate the effectiveness of our method. However, mutual interference between image references remains significant due to the limited precision of RCM, as MM-DiT is not explicitly optimized for semantic segmentation. In future work, we aim to mitigate this interference to achieve finer control over generation.

References

[1] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Muller, and Bjorn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. 3

[2] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8861–8870, 2024. 5

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2, 5, 6, 7, 8, 1

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 5, 6

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3

[6] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[7] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-delta: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 3

[8] Nan Chen, Mengqi Huang, Zhuowei Chen, Yang Zheng, Lei Zhang, and Zhendong Mao. Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization. *arXiv preprint arXiv:2409.05606*, 2024. 2, 3, 7

[9] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 3, 6, 7

[10] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36:30286–30305, 2023. 6, 7

[11] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 2

[12] Bin Cheng, Zuhao Liu, Yunbo Peng, and Yue Lin. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22736–22746, 2023. 3

[13] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024. 2, 3, 5, 6, 1

[14] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 5, 6, 7, 1

[15] Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z-star: Zero-shot style transfer via attention rearrangement. *arXiv preprint arXiv:2311.16491*, 2023. 2, 3, 5, 6, 1

[16] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out!, 2025. 2

[17] 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. 2, 3, 4, 5, 6

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 6, 7, 8

[19] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024. 2, 3, 6

[20] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 6, 7

[22] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 2, 5, 6, 7, 8, 1

[23] Teng-Fang Hsiao, Bo-Kai Ruan, and Hong-Han Shuai. Training-and-prompt-free general painterly harmonization via zero-shot disentenglement on style and content references, 2024. 2, 3, 5, 6, 1

[24] Teng-Fang Hsiao, Bo-Kai Ruan, Sung-Lin Tsai, Yi-Lun Wu, and Hong-Han Shuai. Freecond: Free lunch in the input conditions of text-guided inpainting. *arXiv preprint arXiv:2412.00427*, 2024. 2

[25] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European*

Conference on Computer Vision, pages 150–168. Springer, 2024. 2, 6

[26] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 6

[27] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 5, 6

[28] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 3

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742, 2023. 2

[30] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 3

[31] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7817–7826, 2024. 6, 7

[32] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 2

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 3

[34] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 3, 6

[35] Yang Luo, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Zhineng Chen, Yu-Gang Jiang, and Tao Mei. Freeenhance: Tuning-free image enhancement via content-consistent noising-and-denoising process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7075–7084, 2024. 2, 5, 6

[36] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024. 2

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*. 6, 7

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[39] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 5

[40] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 6, 7

[41] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. lambda-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. *arXiv preprint arXiv:2402.05195*, 2024. 3

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2, 3, 4, 7

[43] 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. 6

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2, 3, 4, 6, 7

[45] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024. 3, 7

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 3, 6, 7, 8

[47] Dvir Samuel, Barak Meiri, Haggai Maron, Yoad Tewel, Nir Darshan, Shai Avidan, Gal Chechik, and Rami Ben-Ari. Lightning-fast image inversion and editing for text-to-image diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[48] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 3

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[50] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 14398–14409, 2024. 2, 3, 6, 7

[51] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[52] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Trans. Graph.*, 43(4), 2024. 1

[53] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2, 3, 6, 7, 8

[54] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023. 2, 6

[55] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 2

[56] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *CoRR*, 2023. 6

[57] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 2, 3, 6, 7, 8

[58] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[59] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 2

[60] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 6

[61] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[64] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[65] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer, 2024. 2

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 Large [1]
- 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. 18a. (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.

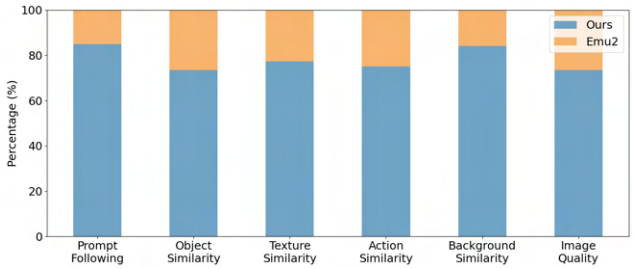


Figure 9. The user study comparison of TF-TI2I (denoted by blue) versus Emu2 (denoted by orange).

(a) Share Attention (SA).

(b) Contextual Tokens Sharing (CTS).

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

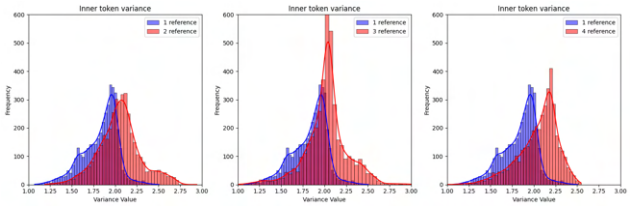


Figure 11. The visualization of distribution shift leading by additional visual references, computed in $t = 20, l = 37$.

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 DreamBench 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.

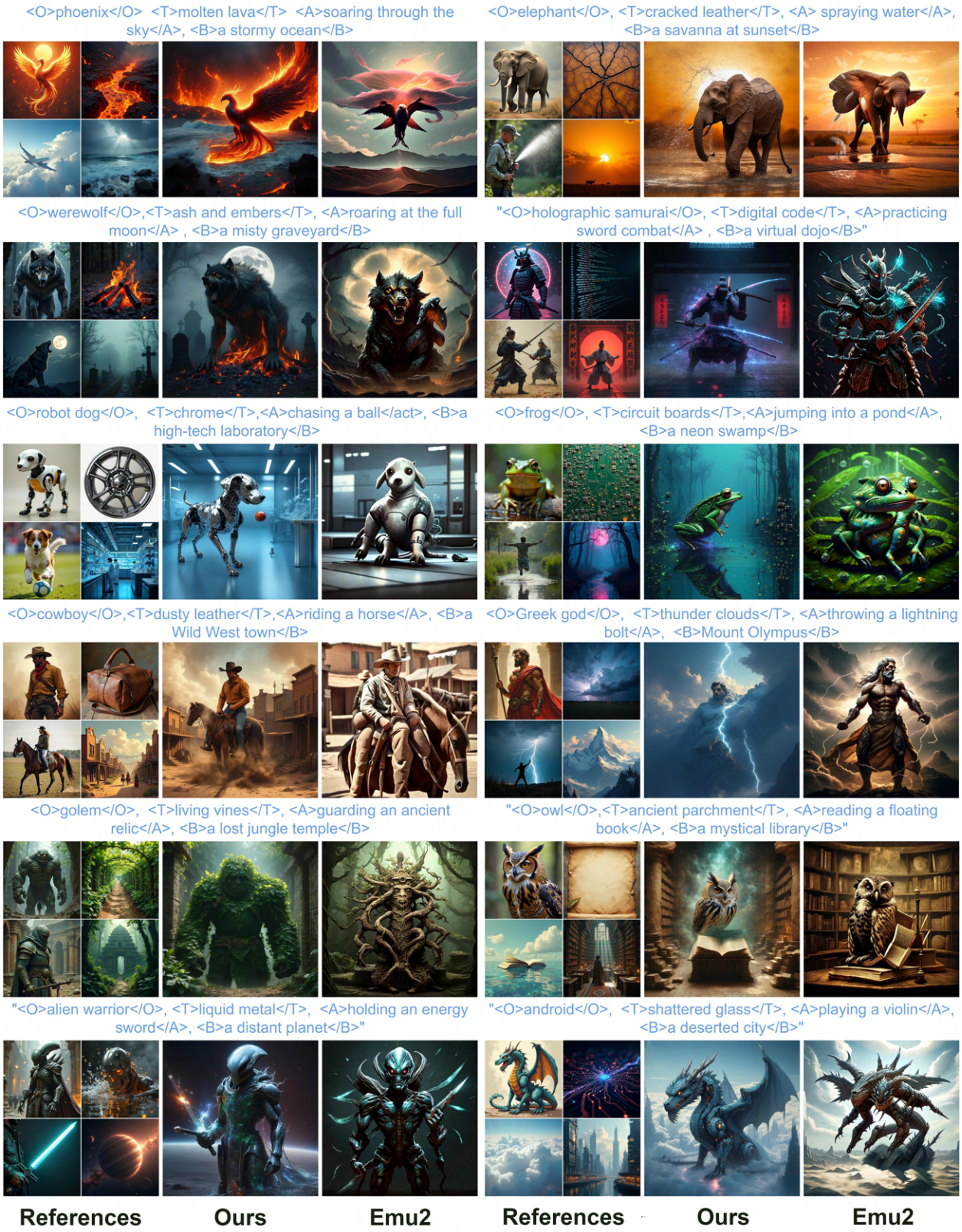


Figure 12. Qualitative comparison of Quad-references sub-tasks. The input Object, Texture, Action, and Background—are denoted as O, T, A, and B. We use red for text-only input and blue for reference-supported input.

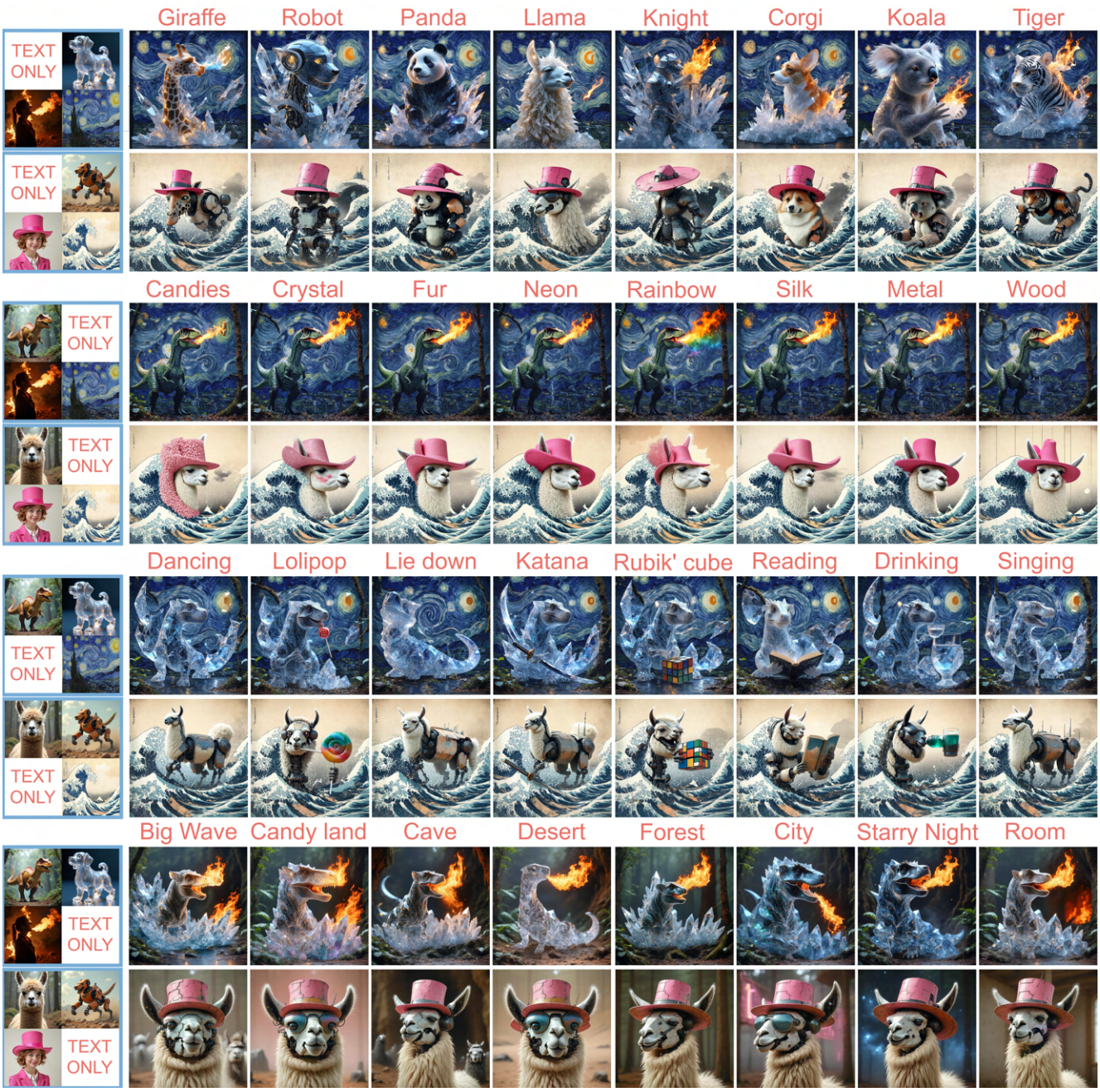


Figure 13. Qualitative results of Trio-references sub-tasks on FG-TI2I, the first template of the prompt is *a dinosaur with the texture of crystal doing breathing fire in the starry night*, and the second one is *a llama with the texture of robot doing wearing pink high gat in the great wave.*. In this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

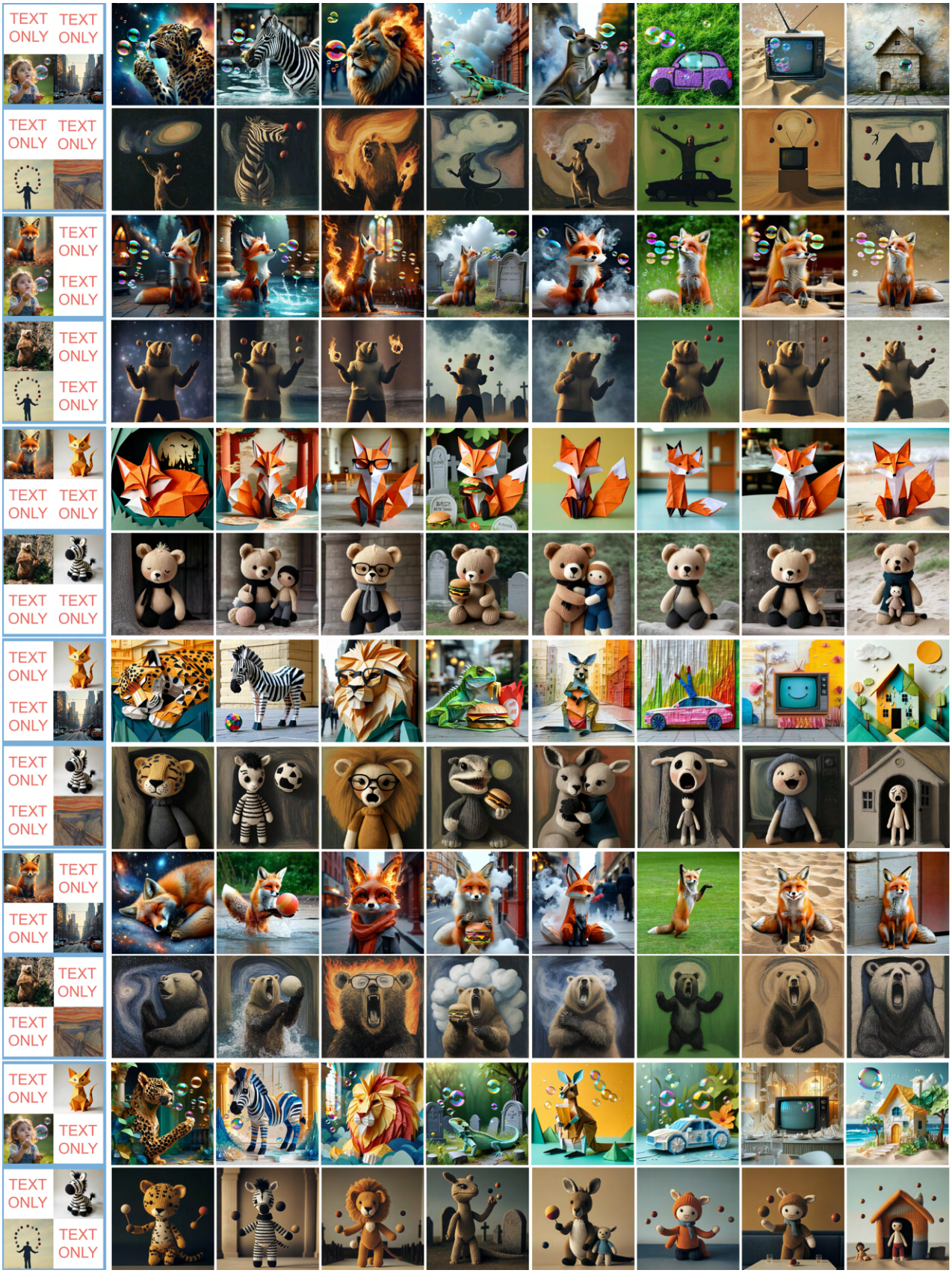


Figure 14. Qualitative results of Dual-references sub-tasks on FG-TI2I, we leave the textual prompt in Appendix B.1.4

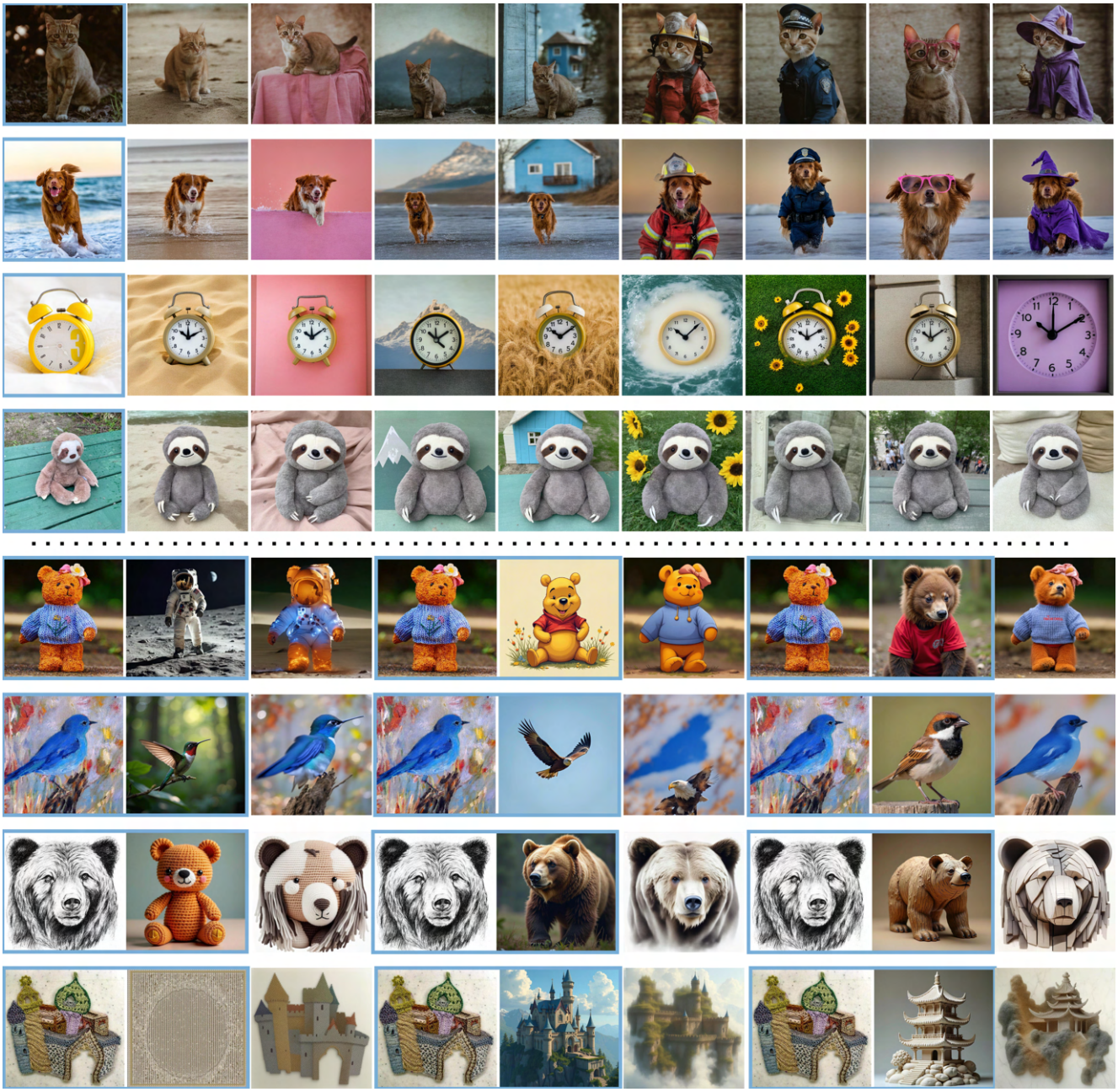


Figure 15. Qualitative results of TF-TI2I on DreamBench (upper part) and Wild-TI2I (lower part), where the image input is denoted with blue, the textual prompts are abbreviated for conciseness.

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 Anaylsis

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. 18a. 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 layers and later steps. To ensure precise feature extraction, we activate RCM only in the late steps.

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.

C.3. Winner-Takes-All

The visualization of WTA across different layers and steps is presented in Fig. 18b. 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 T2I 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]	0.26	0.57	6.32	0.27	0.61	6.01	**0.30**	0.57	**7.01**	0.26	0.53	6.02	0.27	0.59	6.02	**0.29**	0.48	6.89
Emu2 [50]	0.26	0.62	5.86	0.25	0.59	5.66	0.28	0.59	6.67	0.26	0.62	5.84	0.26	0.62	5.69	0.28	0.55	6.70
Ours	**0.28**	0.58	6.37	**0.28**	0.54	**6.07**	0.28	0.51	6.78	**0.28**	0.54	**6.30**	**0.28**	0.55	**6.08**	0.28	0.42	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 Fig. 17a, replacing ϵ with an inversion algorithm, such as RF-inversion [45], does not significantly alter the generation outcome.

C.6. Empty String for References.

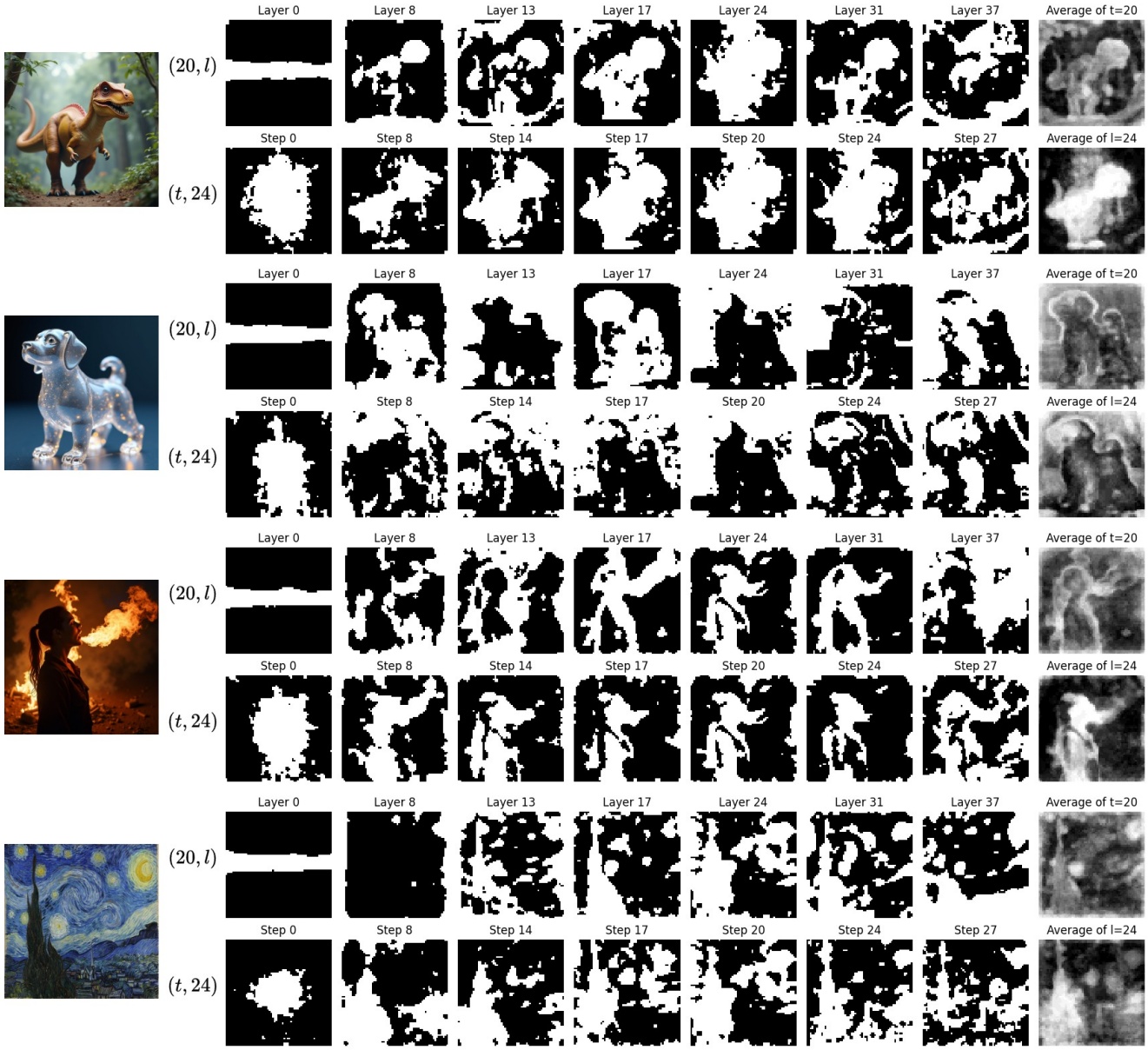
Following the setup of previous reference-supported T2I models [8, 18, 46, 50, 53, 57], which utilize an additional reference prompt to stabilize the generation process, we adopt the same approach in our proposed TF-TI2I and FG-TI2I. However, as shown in Fig. 17b, this design may not be necessary for 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.

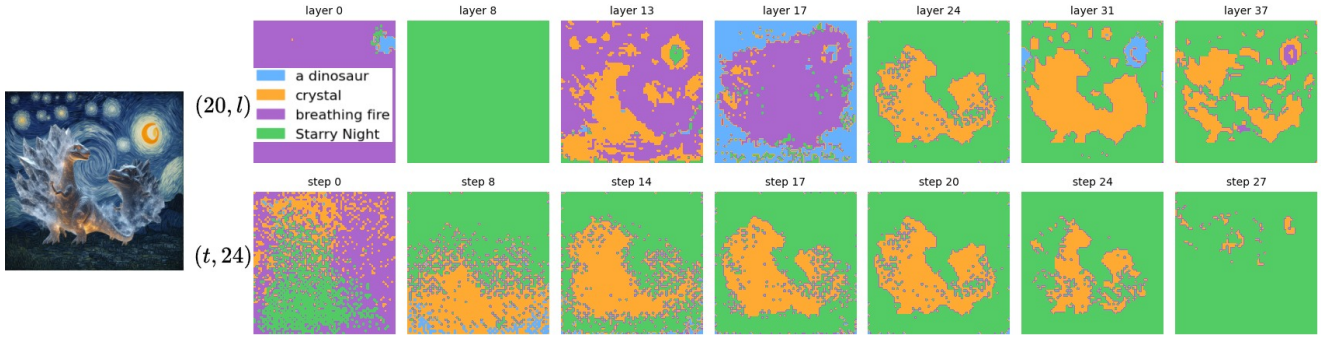
The Shared Attention module is a widely used modification for UNet-based T2I models [42, 44], where the self-attention keys and values from the reference image are concatenated into the generation process. This enables the generated output to share visual elements with the reference image. However, due 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.

(a) Inversion noise (b) Empty string for references

Figure 17. An illustration of using different noise function and reference prompt for TF-TI2I.



(a) Reference Contextual Masking



(b) Winner-Takes-All

Figure 18. 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)