CTCAL: Rethinking Text-to-Image Diffusion Models via Cross-Timestep Self-Calibration

Supplementary Material

This appendix is structured as follows:

- In Appendix A, we provide a discussion of related work.
- In Appendix B, we provide additional implementation details regarding CTCAL, including the processing workflow for cross-attention maps, the architecture of the proposed lightweight autoencoder, and the training timestep sampling strategy for SD 3.
- In Appendix C, we provide more details of experimental settings, including tradeoff parameters and benchmarks.
- In Appendix D, we provide additional results and analyses, including further applications for inference acceleration, additional qualitative comparisons, and extended discussions.
- In Appendix E, we provide the limitations of CTCAL.
- In Appendix F, we provide details on the ethics statement.

A. Related Work

Text-to-image synthesis aims to generate visually realistic images that accurately reflect input text prompts. Early advances in this field were predominantly driven by Generative Adversarial Networks (GANs) [42, 48, 52, 53, 56] and Autoregressive Models (ARs) [4, 13, 35, 51]. Recently, Diffusion Models (DMs) [12, 19] have emerged as the dominant paradigm, demonstrating superior capabilities in generating high-fidelity and semantically coherent images.

Building upon the foundational framework of Denoising Diffusion Probabilistic Models (DDPMs), diffusion-based approaches have become the cornerstone of contemporary text-to-image synthesis. Representative models include GLIDE [31], which extends diffusion models for text-guided image generation and editing, DALL-E 2 [36], which leverages the CLIP embedding space for improved text-image generation, Imagen [40], which employs cascaded diffusion for high-resolution synthesis, and Stable Diffusion (SD) [39], which adopts latent diffusion for computational efficiency.

To further advance the quality and fidelity of generated images, researchers have proposed a range of architectural innovations. These include the integration of flow-based method [1, 27, 29, 30], the development of Diffusion Transformers (DiT) [7, 25, 32], and the introduction of Multi-Modal Diffusion Transformer (MM-DiT) [15, 22]. Notable recent advancements in this direction include PixArt-α [7], HunyuanDiT [25], Stable Diffusion 3 [15], FLUX.1 [22].

Beyond architectural improvements, fine-tuning strategies have been extensively explored to further enhance text-to-image diffusion models. Data augmentation-based methods [2, 9, 14, 20, 23, 41, 44] focus on modifying the training data distribution to improve both visual fidelity and textual alignment. Backpropagation-based approaches [8, 34, 45, 47] employ differentiable reward functions, enabling end-to-end optimization via gradient descent. Reinforcement learning-based methods [3, 6, 11, 16, 54] incorporate Reinforcement Learning from Human Feedback (RLHF), allowing models to iteratively refine their outputs based on reward signals. Furthermore, Direct Preference Optimization (DPO) based approaches [24, 26, 43, 49, 50, 55] circumvent the complexities of explicit reward modeling by directly optimizing for user preferences.

Despite these advances, existing methods largely overlook the explicit supervision of text-image correspondence learning, leading to suboptimal results. In this work, we address this gap by leveraging a fundamental characteristic of text-to-image diffusion models, utilizing the robust text-image alignment established at smaller timesteps to supervise and calibrate the learning at larger timesteps.

B. CTCAL

Processing workflow for cross-attention maps. Given an input triplet comprising a real image, a corresponding text prompt, and Gaussian noise, we first sample a specific timestep and subsequently extract the associated cross-attention maps generated during the forward process of the denoising network. These maps are subsequently averaged across all layers and attention heads to yield the final cross-attention map. The aggregated map $\mathbf{A} \in \mathbb{R}^{H \times W \times n}$ consists of n spatial attention maps, each corresponding to a token in the text prompt.

For Stable Diffusion 2.1, the attention maps are extracted at a spatial resolution of 16×16 pixels. Consistent with the methodology described in [5], we apply a Gaussian filter with a kernel size of 3 and a standard deviation of 0.5 to smooth the cross-attention maps. For Stable Diffusion 3, the attention maps are extracted at a spatial resolution of 64×64 pixels, and we apply a Gaussian filter with a kernel size of 5 and a standard deviation of 0.5 to smooth the cross-attention maps.

Regarding the *alignment terms* in CTCAL, it is important to note that the magnitude of the cross-attention response for each token-specific attention map varies with the sampled timestep, whereas CTCAL primarily emphasizes spatial alignment. Therefore, we normalize each to the range [0, 1] to ensure consistency in scaling. Regarding the *regularization term* in CTCAL, our approach concentrates

exclusively on attention maps corresponding to tokens that are semantically relevant to the given prompt. Specifically, we re-weight the attention values by excluding the specialized tokens *sot* and *eot*, and subsequently apply a softmax function over the remaining tokens.

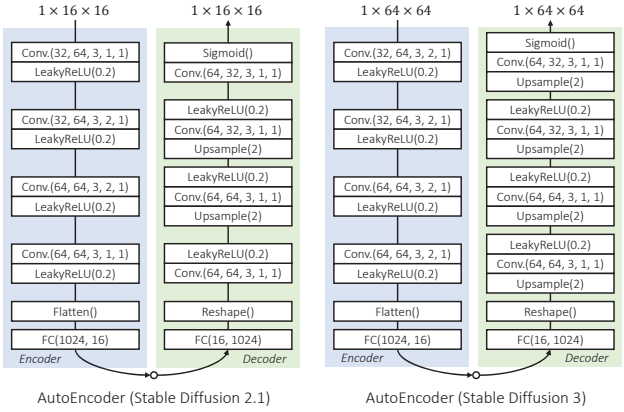

Figure 1. **The detailed architecture of the autoencoder.**

Lightweight autoencoder. The detailed architecture of the proposed lightweight autoencoder is illustrated in Fig. 1. We design dedicated autoencoders for Stable Diffusion 2.1 and Stable Diffusion 3, respectively, with the latter exhibiting a higher compression ratio. The latent code form extracted is the same in both cases.

Training strategy for Stable Diffusion 3. As outlined in the main paper, recent advancements in text-to-image diffusion models, such as Stable Diffusion 3, have shifted from traditional training paradigms that employ uniform timestep sampling to more sophisticated strategies incorporating non-uniform timestep samplers, notably logit-normal sampler [15]. This methodological evolution necessitates a reevaluation of timestep prioritization, particularly when determining t_{tea}.

Given timestep t, we define its priority as $w_t = \frac{p_t}{t}$, where p_t represents the probability of sampling timestep t under the employed distribution. Intuitively, timesteps characterized by lower noise levels (*i.e.*, smaller values of t) and higher sampling probabilities p_t are assigned greater priority values w_t. Consequently, we select the timestep with the highest priority as t_{tea}.

We also provide the following PyTorch code for calculating p_t in the case of a logit-normal sampler:

```
from scipy.stats import norm

def logistic_interval_prob(a, b, mu,
    sigma):
    """
    Computes the probability that
       Y = sigmoid(X) falls within the
           interval [a, b],
       where X ~ N(mu, sigma^2).

    Parameters:
       a, b : Endpoints of the interval
           for Y (0 < a < b < 1).
       mu, sigma : Parameters of the
           Gaussian distribution.

    Returns:
       P(a <= Y <= b)
    """
    # Ensure numerical stability (avoid
        division by zero)
    a_clip = np.clip(a, 1e-7, 1 - 1e-7)
    b_clip = np.clip(b, 1e-7, 1 - 1e-7)

    # Compute the corresponding endpoints
        for X
    x_a = np.log(a_clip / (1 - a_clip))
    x_b = np.log(b_clip / (1 - b_clip))

    # Calculate the Gaussian CDF
    prob = norm.cdf(x_b, loc=mu, scale=
        sigma) - \
           norm.cdf(x_a, loc=mu, scale=
               sigma)
    return prob
```

C. Experimental settings

Training details. Here we expand on training details and provide hyperparameters:

For SD 2.1, we utilize the AdamW optimizer, configuring the learning rate to 5×10^{-5} for the denoising network and autoencoder, 5×10^{-6} for the text encoder. The optimizer hyperparameters are set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, with a weight decay of 0.01. Training is conducted with a per-device batch size of 16 across 4 NVIDIA A800 GPUs, resulting in an effective batch size of 64. The fine-tuning is completed with 20,000 training steps.

For SD 3, we similarly employ the AdamW optimizer. The denoising network and autoencoder are trained with a learning rate of 5×10^{-5}, while the text encoder utilizes a learning rate of 1×10^{-5}. The corresponding optimizer parameters are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, with a weight decay of 1×10^{-4} for the denoising network and 1×10^{-3} for the text encoder. Training is performed with a per-device batch size of 4 on 4 NVIDIA A800 GPUs, the gradient accumulation step is set to 4, yielding a total batch size of 64. Fine-tuning is performed for a total of 20,000 optimization steps.

Tradeoff parameters. CTCAL is a hyperparameter-robust method. Instead of meticulously tuning the tradeoff parameters, we empirically determine them by aligning the magnitudes of various loss terms. This straightforward strategy enables stable training convergence and performance gains. For SD 2.1, we empirically set $\lambda_1 = 0.01$, $\lambda_2 = 1$, $\lambda_3 = 0.01$, and $\lambda_4 = 0.1$. For SD 3, which we found to be less

Methods	Overall	Single object	Two object	Counting	Colors	Position	Color attribution
SD 1.5 [39]	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD 2.1 [39]	0.50	0.98	0.51	0.44	0.85	0.07	0.17
SD XL [33]	0.55	0.98	0.74	0.39	0.85	0.15	0.23
Pixart-α [7]	0.48	0.98	0.50	0.44	0.80	0.08	0.07
Hunyuan-DiT [25]	0.63	0.97	0.77	0.71	0.88	0.13	0.30
DALL-E 2 [36]	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALL-E 3 [2]	0.67	0.96	0.87	0.47	0.83	0.43	0.45
FLUX.1-dev [22]	0.67	0.99	0.81	0.79	0.74	0.20	0.47
Sana (1.6B) [46]	0.66	0.99	0.77	0.62	0.88	0.21	0.47
SD 3 (2B) [15]	0.62	0.98	0.74	0.63	0.67	0.34	0.36
SD 3 (2B) + CTCAL	**0.69**	**0.99**	**0.85**	**0.70**	**0.79**	**0.38**	**0.42**

Table 1. **Objective evaluation on GenEval.** CTCAL improves performance across all categories.

susceptible to the effects of imbalanced cross-attention responses among subjects on the T2I-CompBench++ benchmark, we empirically set $\lambda_1 = 0.01$, $\lambda_2 = 1$, $\lambda_3 = 0.01$, and $\lambda_4 = 0$.

To further enhance performance through extensive hyperparameter tuning, we suggest considering a strategy based on the gradient alignment of each loss term. Regarding concerns about performance degradation or training instability, we recommend appropriately increasing the weight of the diffusion loss to mitigate such issues.

Benchmark. We conduct a comprehensive evaluation of CTCAL utilizing two widely recognized benchmarks: T2I-CompBench++ [21] and GenEval [17]. T2I-CompBench++ comprises 8,000 compositional prompts across eight categories, and serves to assess text-to-image generation models through a multifaceted evaluation framework, including visual question answering (VQA), object detection, image-text matching scores, and assessments by multimodal large language models (MLLMs). GenEval consists of 553 object-centric prompts and focuses on evaluating the compositional reasoning capabilities of text-to-image generation models, which utilizes object detection and color classification tasks to rigorously assess the ability to capture and reproduce fine-grained object properties as specified in the text prompts.

Baseline (GORS). Generative model finetuning with Reward-driven Sample selection (GORS) presents a straightforward yet effective methodology to enhance the compositional capabilities of pre-trained text-to-image diffusion models. The core idea is to fine-tune a pre-trained model, *e.g.*, Stable Diffusion [39], using a curated set of generated images that exhibit a high degree of alignment with given text prompts, and the fine-tuning loss is modulated by a reward signal.

Formally, let θ denotes the text-to-image diffusion model and $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$ represents a collection of text prompts. For each prompt $\mathbf{y}_i$, GORS generates k candidate images, yielding a total of kn images $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{kn}\}$. Each image $\mathbf{x}_j$ is assigned an alignment score s_j, serving as the reward metric. GORS then selects those images whose corresponding reward scores surpass a predefined threshold, forming a subset of samples $\mathcal{D}_s$ to be used for fine-tuning.

During the fine-tuning phase, the loss associated with each sample is weighted according to its reward score. The overall fine-tuning objective is expressed as:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x},\mathbf{y},s)\in\mathcal{D}_s}\left[s \cdot \|\epsilon - \epsilon_\theta(\mathbf{x}, \mathbf{y}, t)\|_2^2\right], \quad (1)$$

where $(\mathbf{x}, \mathbf{y}, s)$ is the triplet of the image, text prompt, and reward.

Methods	SD 2.1	INITNO	SynGen	VSC	CTCAL
Color	0.50	0.69	0.71	0.74	**0.78**
Texture	0.49	0.61	0.61	0.64	**0.70**

Table 2. **More quantitative comparison.**

D. Additional experimental results

D.1. More objective evaluation on GenEval.

As an extension of Table 2 in the main paper, we show all objective evaluation results on GenEval in Table 1.

D.2. More quantitative comparison

Table 2 provides a quantitative comparison with additional methods. Following the VSC [10] protocol and leveraging higher-quality training data synthesized by advanced models, specifically FLUX [22], Stable Diffusion 3.5 [15], and SynGen [37], the comparative analysis in Table 2 demonstrates that our approach consistently outperforms VSC [10], SynGen [37], and INITNO [18].

Moreover, in contrast to VSC, which is similarly based on fine-tuning, the proposed method necessitates no archi-

tectural modifications and imposes no restrictions on the input text templates.

D.3. More application on inference acceleration

The text-to-image diffusion model fine-tuned with CTCAL exhibits better consistency between cross-attention maps at larger timesteps and those at smaller timesteps. This phenomenon can be leveraged for inference acceleration.

To further validate this approach, we utilize TGATE [28], a simple and training-free inference acceleration method that efficiently generates images by caching and reusing attention outputs from the predefined timestep. Specifically, we adopt its design for cross-attention layers: during inference, cross-attention computation is performed only for the first T_c timesteps, and the output from the T_c-th timestep is reused in subsequent inference steps.

The validation experiment is shown in Fig. 2, SD 2.1 fine-tuned with CTCAL significantly improves the quality of the generated images while showing excellent alignment with text prompts under smaller T_c settings, for both hard samples (top) and simple samples (bottom).

D.4. Additional qualitative comparison

Fig. 3 presents examples of complex prompts, and Fig. 4 presents additional qualitative comparisons. It can be observed that our method generates semantically more plausible and photorealistic results than its counterparts, successfully capturing all input concepts.

D.5. Further discussion on external model reliance

In contrast to the self-calibration paradigm of the proposed method, an alternative approach involves leveraging external detectors (*e.g.*, Grounded SAM [38]) to supervise the learning of image-text correspondences. The relevant discussion regarding the limitations of external model reliance is detailed below: *(1) Scalability*: Relying on superior external encoders can, in some cases, be impractical. While visual encoders, *e.g.*, DINO, support ImageNet-scale training for DiT/SiT, they struggle to support T2I training (*e.g.*, FLUX, SD) on billion-scale datasets. Grounded SAM also faces the limitations. *(2) Domain Shift*: External models (*e.g.*, Grounded SAM) pre-trained on generic real images often generalize poorly to specialized domains (*e.g.*, medical) or synthetic data, the latter being indispensable for modern large-scale generative training, as leveraged in this work. *(3) Representational disparity*: Native attention maps provide continuous soft masks that retain rich spatial confidence. Conversely, binary hard masks or non-native signals from Grounded SAM lack this precision, resulting in sub-optimal alignment. *(4) Experiments*: Results in Table 3 show that our self-calibration paradigm outperforms Grounded SAM-based alternatives.

Methods	Color B-VQA	2D-Spatial UniDet
CTCAL w/ Grounded SAM	0.6988	0.2017
CTCAL (Ours)	**0.7233**	**0.2142**

Table 3. **More quantitative comparison** with Grounded SAM-based strategy.

E. Limitations

In this study, we utilize *Stanza* for part-of-speech analysis to extract nouns from text prompts, subsequently prioritizing the attention maps corresponding to these nouns for utilization in CTCAL. However, it is noteworthy that not every noun extracted via *Stanza* encapsulates meaningful spatial semantics. For instance, directional nouns such as *top* and *left* may be extracted; while a blacklist-based filtering mechanism can be employed to exclude such distractors, this method lacks generalizability and may not robustly address the issue across diverse prompts. The inadvertent inclusion of these nouns can consequently compromises the overall performance of CTCAL. To address this limitation, incorporating large language models to discern nouns that possess explicit physical spatial semantics emerges as a promising direction for enhancing the robustness and effectiveness of noun selection.

F. Ethics statement

We uphold the highest ethical standards in our research, which includes complying with legal frameworks, respecting privacy rights, and encouraging the generation of positive content. Text-to-image models have a wide range of applications across diverse scenarios. Although these models may be misused for harmful purposes, it is essential to employ safety checkers or filters during actual deployment to prevent the generation of NSFW content. In this work, we rely on existing pre-trained models and therefore inherit their inherent biases. However, our training framework is highly flexible and, by utilizing customized, safe, and trustworthy datasets for fine-tuning, can effectively mitigate the potential harms associated with current models.

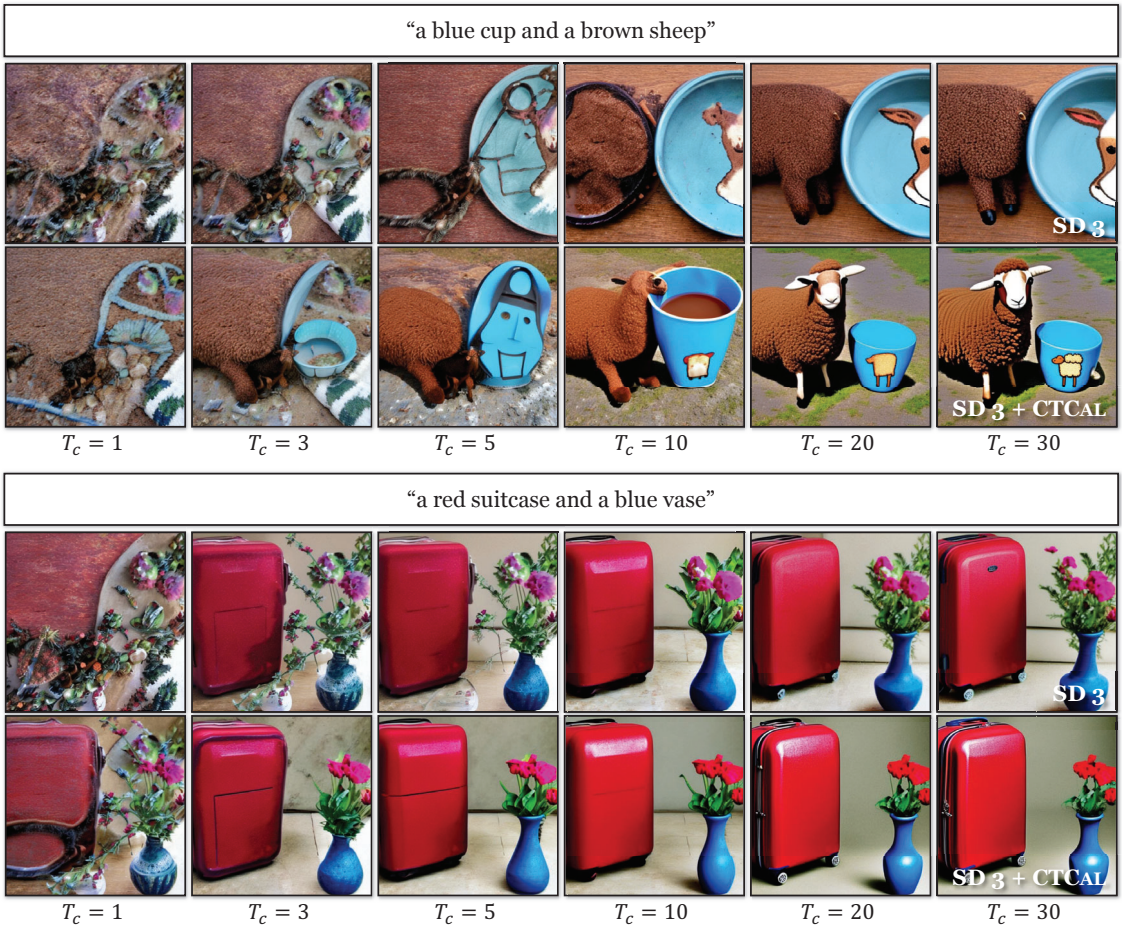

Figure 2. **More application on inference acceleration.**

Figure 3. **More results** synthesized by the proposed CTCAL using complex text prompts.

Figure 4. **More results** synthesized by the proposed CTCAL.

References

[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2(3):8, 2023. 1, 3

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 1

[4] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning (ICML)*, 2023. 1

[5] 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. 1

[6] Chaofeng Chen, Annan Wang, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Enhancing diffusion models with text-encoder reinforcement learning. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 3

[8] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024. 1

[9] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 1

[10] Do Huu Dat, Nam Hyeon-Woo, Po-Yuan Mao, and Tae-Hyun Oh. Vsc: Visual search compositional text-to-image diffusion model. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 3

[11] Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1

[13] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1

[14] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research (TMLR)*, 2023. 1

[15] 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 *International Conference on Machine Learning (ICML)*, 2024. 1, 2, 3

[16] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1

[17] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[18] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1

[20] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1

[21] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025. 3

[22] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 1, 3

[23] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 1

[24] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. In *NeurIPS*, 2024. 1

[25] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1, 3

[26] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024. 1

[27] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[28] Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion via temporal attention decomposition. *Transactions on Machine Learning Research (TMLR)*, 2025. 4

[29] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[30] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[31] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, 2022. 1

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[33] 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. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[34] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 1

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021. 1

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3

[37] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[38] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[41] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023. 1

[42] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[43] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[44] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[45] Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[46] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[47] 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. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1

[48] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[49] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[50] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense

reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024. 1

[51] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research (TMLR)*, 2022. 1

[52] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[53] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[54] Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[55] Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. In *International Conference on Learning Representations (ICLR)*, 2025. 1

[56] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1