1 Algorithm

Here we first detail the mixed latent training strategy and then provide the pseudocode for a single loss computation step.

The mixed latent strategy contains two types of latents in the fine-tuning procedure, i.e., the latent starting from the pure noise and the noisy latent from the GT Images.

Latent starting from the pure noise This serves as the main branch in our pipeline. Our fine-tuning process shares the same procedure to generate an image as the diffusion model does in the inference time. We uniformly sample K steps from all the inference steps to enable the gradient. Therefore, the latent is sampled from the pure noise $\mathcal{N}(0, I)$. We iteratively denoise it to obtain the generated image. The image is then used to calculate the $\mathcal{L}_{i2t}$ and $\mathcal{L}_{adv}$ loss. It is also sent to the segmentation model to provide the object mask for computing the $\mathcal{L}_{pos}$ and $\mathcal{L}_{neg}$. The latent starting from the pure noise corresponds to the upper left part in Fig. **??**. Please refer to [34] for how to receive the gradient from the loss.

Noisy latent from the GT Images We also aim to inject information from the GT images to stabilize the fine-tuning process. We randomly sample a timestamp τ from a pre-defined range $[T_1, T_2]$. Then we obtain x_τ by adding the timestamped noise ϵ_τ on the latent of the GT Image x_0. We also iteratively denoise this noisy GT latent to get $\hat{x}_0$ as we do for the latents starting from the pure noise. This $\hat{x}_0$ is only used to calculate the $\mathcal{L}_{i2t}$ loss. The latent starting from the noisy GT corresponds to the bottom left part in Fig. **??**.

The pseudocode for a single loss computation step for the online T2I-Model is described below.

Algorithm 1 A single loss computation step for the online T2I-Model during fine-tuning

```
Input: Text prompt $\mathcal{P}$, GT Image $\mathcal{I}_g$, GT Prompt $\mathcal{P}_g$, original T2I-Model $\epsilon_{pre}$, online T2I-Model $\epsilon_\theta$, pre-trained I2T-Model $\mathcal{C}$, discriminator $\mathcal{D}_\phi$, segmentation model $\mathcal{S}$, timestep range $[T_1, T_2]$, timestep $\tau$, attention map $A$, scaler $\lambda$; [; ] denotes concatenate
$x_T, \xi \sim \mathcal{N}(0, I)$
$\tau \sim$ Uniform$[T_1, T_2]$
$x_\tau =$ AddNoise$(\mathcal{I}_g, \xi, \tau)$
$\hat{\mathcal{I}}, A =$ GenerateImage$(\epsilon_\theta, x_T, \mathcal{P})$
$\hat{\mathcal{I}}_g, \_ =$ GenerateImage$(\epsilon_\theta, x_\tau, \mathcal{P}_g)$
$\mathcal{L}_{i2t} =$ ComputeI2TLoss$(\mathcal{C}, \left[\hat{\mathcal{I}}; \hat{\mathcal{I}}_g\right], [\mathcal{P}; \mathcal{P}_g])$
$\hat{\mathcal{I}}_{pre}, \_ =$ GenerateImage$(\epsilon_{pre}, x_T, \mathcal{P})$
$\mathcal{L}_{adv} =$ ComputeAdvLoss$(\mathcal{D}_\phi, \hat{\mathcal{I}}, \hat{\mathcal{I}}_{pre})$
$\mathcal{L}_{pos}, \mathcal{L}_{neg} =$ ComputeAttrLoss$(\mathcal{S}, \hat{\mathcal{I}}, \mathcal{P}, A)$
$\mathcal{L} = \mathcal{L}_{i2t} + \mathcal{L}_{pos} + \mathcal{L}_{neg} + \lambda \mathcal{L}_{adv}$
Output: Training loss for the online T2I-Model $\mathcal{L}$
```

2 Additional Results and Analysis

2.1 User preference study

We randomly select 100 prompts from DSG1K [6] and use them to generate images with SDXL [26] and our method (CoMat-SDXL). We ask 5 participants to assess both the image quality and text-image alignment. Human raters are asked to select the superior respectively from the given two synthesized images, one from SDXL, and another from our CoMat-SDXL. For fairness, we use the same random seed for generating both images. The voting results are summarised in Fig. 1. Our CoMat-SDXL greatly enhances the alignment between the prompt and the image without sacrificing the image quality.

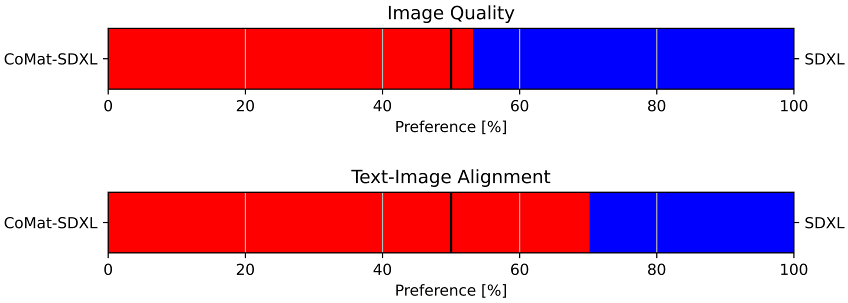


Figure 1: User preference study results.

2.2 Composability with planning-based methods

Since our method is an end-to-end fine-tuning strategy, we demonstrate its flexibility in the integration with other planning-based methods, where combining our method also yields superior performance. RPG [35] is a planning-based method utilizing Large Language Model (LLM) to generate the description and subregion for each object in the prompt. We refer the reader to the original paper for details. We employ SDXL and our CoMat-SDXL as the base model used in [35] respectively. As shown in Fig. 2, even though the layout for the generated image is designed by LLM, SDXL still fails to faithfully generate the single object aligned with its description, e.g., the wrong mat color and the missing candle. Although the planning-based method generates the layout for each object, it is still bounded by the base model's condition following capability. Combining our method can therefore perfectly address this issue and further enhance alignment.

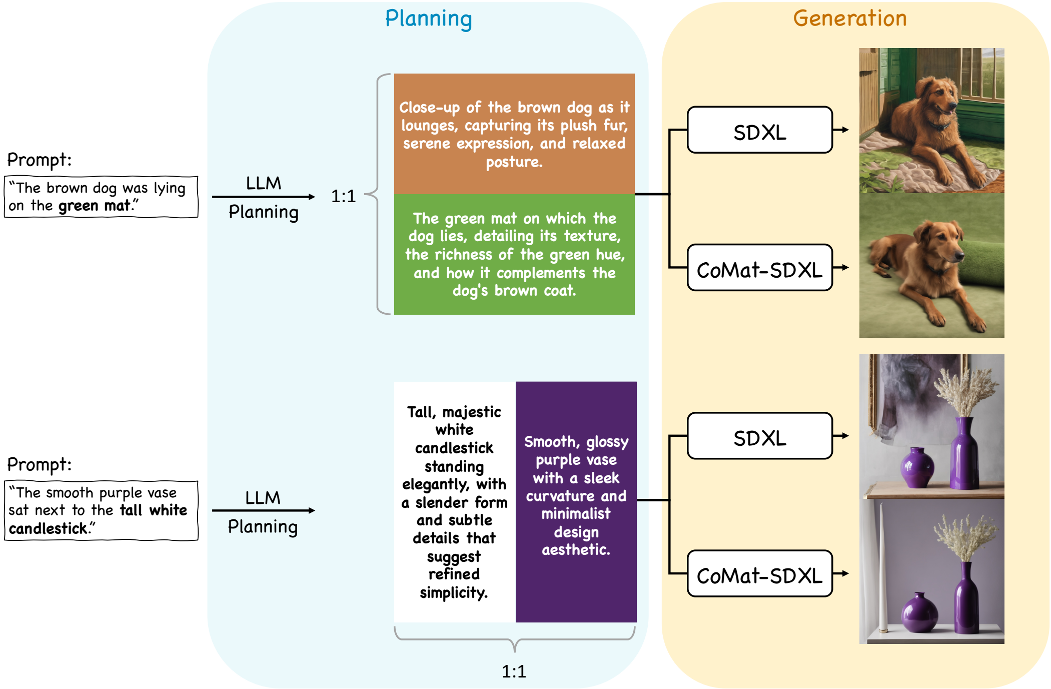

Planning
Generation
Prompt:
"The brown dog was lying on the green mat."
LLM
Planning
1:1
Close-up of the brown dog as it lounges, capturing its plush fur, serene expression, and relaxed posture.
The green mat on which the dog lies, detailing its texture, the richness of the green hue, and how it complements the dog's brown coat.
SDXL
CoMat-SDXL
Prompt:
"The smooth purple vase sat next to the tall white candlestick."
LLM
Planning
Tall, majestic white candlestick standing elegantly, with a slender form and subtle details that suggest refined simplicity.
Smooth, glossy purple vase with a sleek curvature and minimalist design aesthetic.
1:1
SDXL
CoMat-SDXL

Figure 2: Pipeline for integrating CoMat-SDXL with planning-based method. CoMat-SDXL correctly generates the green mat in the upper row and the tall white candle in the bottom row.

2.3 How to choose an image-to-text model?

We provide a further analysis of the varied performance improvements observed with different image-to-text models, as shown in Table 5 of the main text.

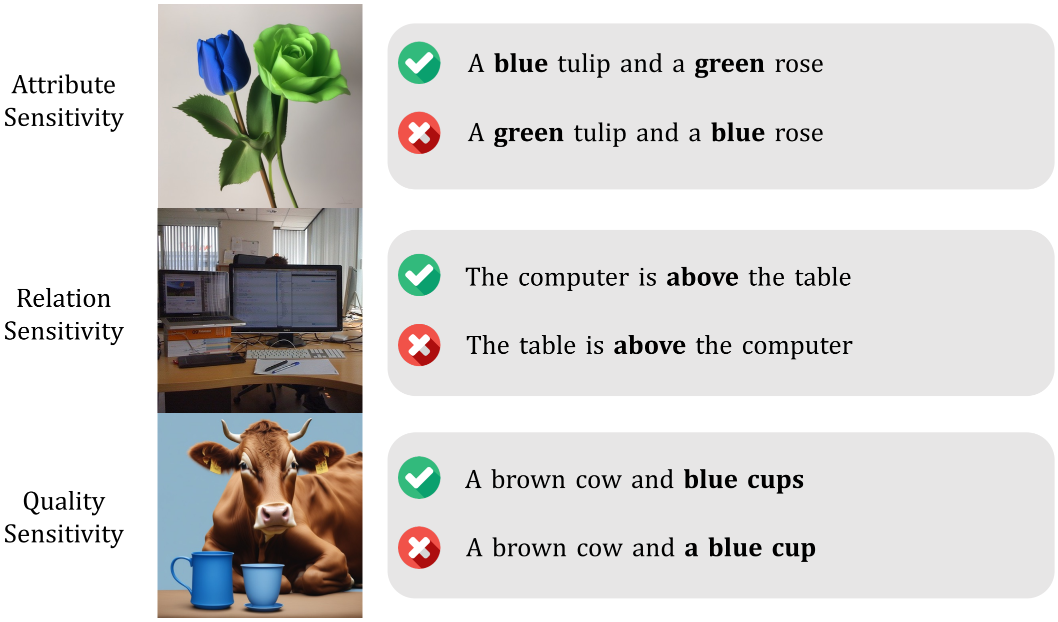

Attribute Sensitivity
✓ A **blue** tulip and a **green** rose
✗ A **green** tulip and a **blue** rose
Relation Sensitivity
✓ The computer is **above** the table
✗ The table is **above** the computer
Quality Sensitivity
✓ A brown cow and **blue cups**
✗ A brown cow and **a blue cup**

Figure 3: Examples for the three core sensitivities.

For an image-to-text model to be valid for the concept activation module, it should be able to tell whether each concept in the prompt appears and appears correctly. We construct a test set to evaluate this capability of the image-to-text model. Intuitively, given an image, a qualified image-to-text model should be sensitive enough to the prompts that faithfully describe it against those that are incorrect in certain aspects. We study three core demands for an image-to-text model:

- **Attribute sensitivity.** The image-to-text model should distinguish the noun and its corresponding attribute. The corrupted caption is constructed by switching the attributes of the two nouns in the prompt.
- **Relation sensitivity.** The image-to-text model should distinguish the subject and object of a relation. The corrupted caption is constructed by switching the subject and object.
- **Quantity sensitivity.** The image-to-text model should distinguish the quantity of an object. Here we only evaluate the model's ability to tell one from many. The corrupted caption is constructed by turning singular nouns into plural or otherwise.

We assume that they are the basic requirements for an image-to-text model model to provide valid guidance for the diffusion model. Besides, we also choose images from two domains: real-world images and synthetic images. For real-world images, we randomly sample 100 images from the ARO benchmark [37]. As for the synthetic images, we use the pre-trained SD1.5 [26] and SDXL [23] to generate 100 images according to the prompts in T2ICompBench [12]. These selections make up for the 200 images in our test data. We show the examples in Fig. 3.

For the sensitivity score, we compare the difference between the alignment score (i.e., log-likelihood) of the correct and corrupted captions for an image. Given the correct caption $\mathcal{P}$ and corrupted caption $\mathcal{P}'$ corresponding to image $\mathcal{I}$, we compute the sensitivity score $\mathcal{S}$ as follows:

$$\mathcal{S} = \frac{\log(p_{\mathcal{C}}(\mathcal{P}|\mathcal{I})) - \log(p_{\mathcal{C}}(\mathcal{P}'|\mathcal{I}))}{|\log(p_{\mathcal{C}}(\mathcal{P}|\mathcal{I}))|}. \tag{1}$$

Then we take the mean value of all the images in the test set. The result is shown in Table 1. The rank of the sensitivity score aligns with the rank of the gains brought by the image-to-text model model shown in the main text. Hence, except for the parameters, we argue that sensitivity is also a must for an image-to-text model to function in the concept activation module.

Table 1: Statistics of image-to-text models.

Image-to-text Model	Parameters	Sensitivity Score
BLIP [18]	469M	0.1987
GIT [31]	394M	0.1728
LLaVA [21]	7.2B	0.1483

3 More Related Work

The image-to-text model in the main text refers to the models capable of image captioning. Previous image captioning models are pre-trained on various vision and language tasks (e.g., image-text matching, (masked) language modeling) [19, 22, 15, 30], then fine-tuned with image captioning tasks [5]. Various model architectures have been proposed [32, 36, 17, 18, 31]. BLIP [18] takes a fused encoder architecture, while GIT [31] adopts a unified transformer architecture. Recently, multimodal large language models (MLLMs) have been flourishing [21, 43, 1, 44, 28, 14]. Empowered by the strong language ability of large language models (LLMs) [39], MLLMs are capable of various vision-language tasks like detailed image captioning [21, 43, 29, 2], visual question answering [20, 41, 42, 3, 13], etc. LLaVA [21, 16, 16] is one of the representative MLLMs. When prompted properly, it can generate elaborate image captions.

Similar to our work, [7] proposes to caption the generated images and optimize the coherence between the produced captions and text prompts. Although an image-to-text model is also involved, they fail to provide detailed guidance. It has been shown that the generated captions are prone to omit key concepts and involve undesired added features [11]. Besides, the method leverages a pre-trained text encoder to compute the similarity between the prompt and generated caption, which further causes information to be missed during text encoding. All these designs lead the optimization target to be vague and sub-optimal.

3.1 vs. Differentiable Reward Method

Similarity: Our method is inspired by the technique introduced in the differentiable reward method to perform gradient update.

Difference: (1) **Reward Model.** Our method is the first to leverage an image-to-text model to perform image captioning on the generated image and compute the loss on the caption. (2) **No fidelity preservation.** The current differentiable reward method ignores the aspect of preserving the generation capability if not training against a reward of image quality. Our method introduces a novel fidelity preservation module, which utilizes a discriminator with similar knowledge to preserve the generation capability. This greatly alleviates the reward hacking problem introduced by only training with the differentiable reward method. (3) **No guidance from real-world image.** The current differentiable reward method all starts from pure noise. Since our method is optimizing for alignment, we can incorporate real-world image-text pairs to guide the optimization process. With our mixed latent strategy, the latent starting from the noise is conditioned on the difficult prompt to promote alignment, while the latent starting from the noisy GT image is used to prohibit the diffusion model from overfitting to the image-to-text model.

3.2 vs. TokenCompose [33]

Similarity: Both [33] and our method incorporates the object mask to guide the attention of the diffusion model.

Difference: (1) **Limited and inferior optimizing target.** [33] merely focuses on optimizing the consistency between the noun mask and the object mask. However, as shown in Fig. **??**, the attention mask of the noun (the 'bear' token) has already aligned well with the object mask. Optimizing for this consistency is inferior. On the other hand, our method focuses on a much broader area, i.e., entity tokens, which consist of nouns and their various associated attributes. We also find that the consistency between the attributes and the object mask bears very little similarity, which should be paid more attention. (2) **No negative concept mapping.** Since the training data of [33] is the real-world image-text pairs, all the nouns in the prompt show up in the image. However, this prohibits the model from learning in a negative way, i.e., if the entity is not on the image, none of the pixel should be activated by this token. Our method leverages images generated by the diffusion model. The entity missing is common. The model obtains the chance to learn in a negative way. (3) **No difficult training data.** Another issue caused by training with image-text pairs is that the training data may be of a common scenario, which is easier to learn. Since our method does not need real-world images and only starts from the noise and text prompt, this enables a more efficient training process.

3.3 vs. Class-specific Prior Preservation Loss [27]

Similarity: Both the class-specific prior preservation loss (CPP Loss) [27] and our proposed fidelity preservation module (FP) share the similar high-level idea of preserving the generation quality while fine-tuning the diffusion models.

Difference: (1) **Target task and preserve domain.** [27] seeks to personalize image generation for specific objects. While the introduced CPP Loss primarily maintains generative capabilities within a narrow domain—specifically, the object class present in the training data—our proposed FP module operates within the context of text-image alignment. FP aims to preserve general generative capabilities by computing adversarial loss across the entire training dataset, encompassing a diverse range of text prompts. (2) **Methodology.** Since the training data of [27] finetunes the diffusion model with the pretraining loss, i.e., the squared error denoising loss on a certain timestamp. CPP Loss follows its form. In contrast, our fine-tuning procedure simulates the inference process of the diffusion model to conduct a full-step inference. We aim to directly supervise the generated image to achieve the training-test alignment. Therefore, we propose the novel FP module to leverage a discriminator to adversarially preserve its quality. The applied discriminator is also updated along with the fine-tuning process, enabling finer control of the image quality.

4 Future Work

We believe our work can also be applied in the text-to-video diffusion models. With the introduction of various MLLMs handling videos [38, 4] and video segmentation models [24, 8], both our concept activation and attribute concentration modules could be used for text-video-alignment training.

5 Experimental Setup

5.1 Implementation Details

Training Details. In our method, we inject LoRA [10] layers into the UNet of the online training model and discriminator and keep all other components frozen. For both SDXL and SD1.5, we train 2,000 iters on 8 NVIDIA A100 GPUS. We use a local batch size of 6 for SDXL and 4 for SD1.5. We choose Grounded-SAM [25] from other open-vocabulary segmentation models [40, 45]. The DDPM [9] sampler with 50 steps is used to generate the image for both the online training model and the original model. In particular, we follow [34] and only enable gradients in 5 steps out of those 50 steps, where the attribute concentration module would also be operated. Besides, to speed up training, we use training prompts to generate and save the generated latents of the pre-trained model in advance, which are later input to the discriminator during fine-tuning.

Training Resolutions. We observe that training SDXL is very slow due to the large memory overhead at 1024×1024. However, SDXL is known to generate low-quality images at resolution 512×512. This largely affects the image understanding of the image-to-text model. So we first equip the training model with better image generation capability at 512×512. We use our training prompts to generate 1024×1024 images with pre-trained SDXL. Then we resize these images to 512×512 and use them to fine-tune the UNet of SDXL for 100 steps, after which the model can already generate high-quality 512×512 images. We continue to implement our method on the fine-tuned UNet.

Training Layers for Attribute Concentration. Following [33], only cross-attention maps in the middle blocks and decoder blocks are used to compute the loss.

Hyperparameters Settings. We provide the detailed training hyperparameters in Table 2.

Table 2: CoMat training hyperparameters for SD1.5 and SDXL.

Name	**SD1.5**	**SDXL**
Online training model		
Learning rate	5e-5	2e-5
Learning rate scheduler	Constant	Constant
LR warmup steps	0	0
Optimizer	AdamW	AdamW
AdamW - β_1	0.9	0.9
AdamW - β_2	0.999	0.999
Gradient clipping	0.1	0.1
Discriminator		
Learning rate	5e-5	5e-5
Optimizer	AdamW	AdamW
AdamW - β_1	0	0
AdamW - β_2	0.999	0.999
Gradient clipping	1.0	1.0
Token loss weight α	1e-3	1e-3
Pixel loss weight β	5e-5	5e-5
Adversarial loss weight λ	1	5e-1
Gradient enable steps	5	5
Attribute concentration steps r	2	2
LoRA rank	128	128
Classifier-free guidance scale	7.5	7.5
Resolution	512×512	512×512
Training steps	2,000	2,000
Local batch size	4	6
Local GT batch size	2	2
Mixed Precision	FP16	FP16
GPUs for Training	8 × NVIDIA A100	8 × NVIDIA A100
Training Time	$\sim$ 10 Hours	$\sim$ 24 Hours

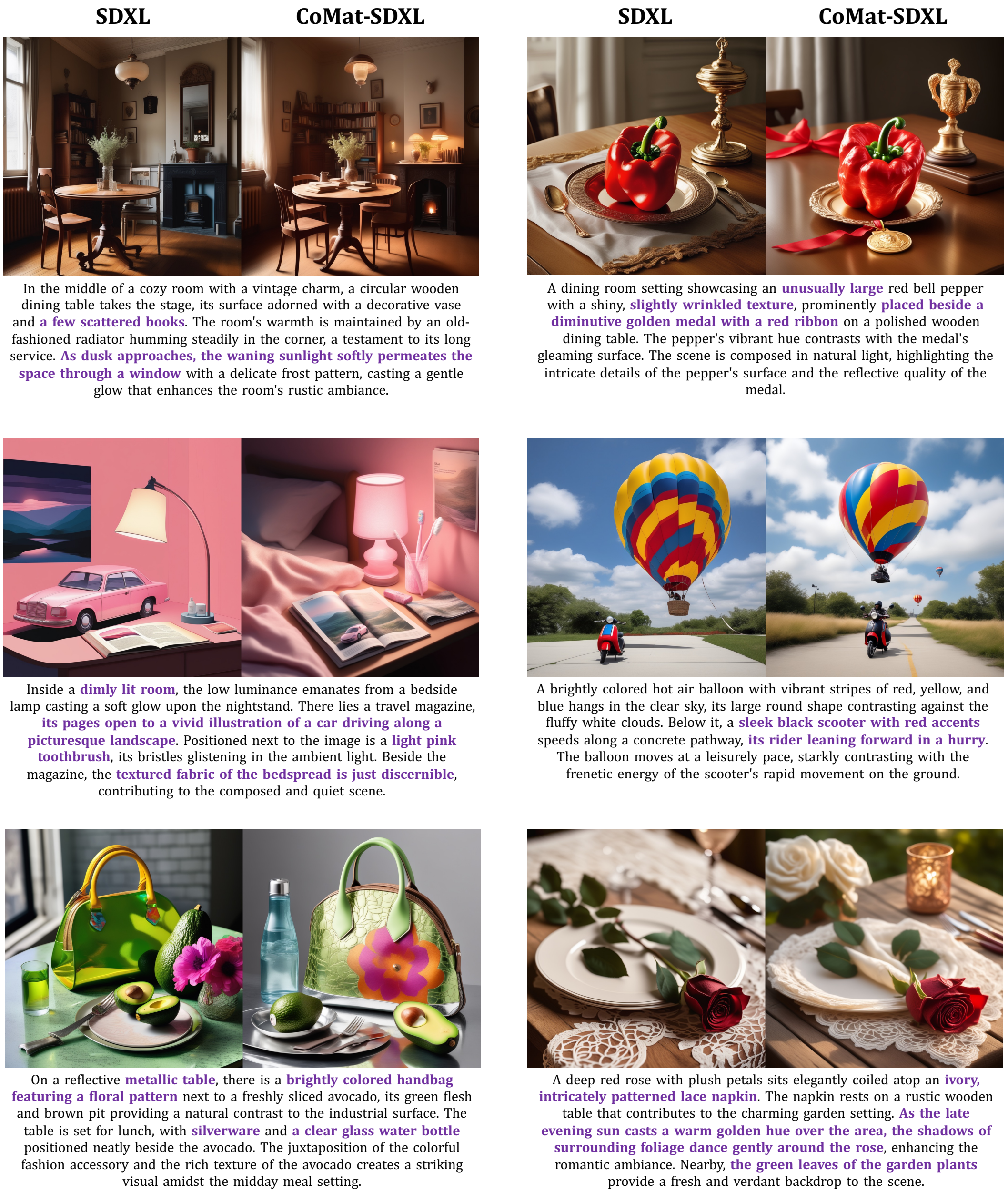
6 More Qualitative Results

6.1 Effectiveness of the Fidelity Preservation module (FP) and Mixed Latent (ML) strategy

We visualize the effectiveness of how we preserve the generation capability of the diffusion model in Fig. **??**. As shown in the figure, without any preservation technique, the diffusion model generates misshaped envelopes and swans. With the FP and ML applied, the diffusion model generates images aligned with the prompt and without artifacts.

6.2 Comparison with the baseline model

We showcase more comparison results between our method with the baseline model in Fig. 4 to 7. Fig. 4 shows the generation results with long and complex prompts. Fig. 5 to 7 shows that our method solves various problems of misalignment, including object missing, incorrect attribute binding, incorrect relationship, inferior prompt understanding.



In the middle of a cozy room with a vintage charm, a circular wooden dining table takes the stage, its surface adorned with a decorative vase and **a few scattered books**. The room's warmth is maintained by an old-fashioned radiator humming steadily in the corner, a testament to its long service. **As dusk approaches, the waning sunlight softly permeates the space through a window** with a delicate frost pattern, casting a gentle glow that enhances the room's rustic ambiance.

A dining room setting showcasing an **unusually large** red bell pepper with a shiny, **slightly wrinkled texture**, prominently **placed beside a diminutive golden medal with a red ribbon** on a polished wooden dining table. The pepper's vibrant hue contrasts with the medal's gleaming surface. The scene is composed in natural light, highlighting the intricate details of the pepper's surface and the reflective quality of the medal.

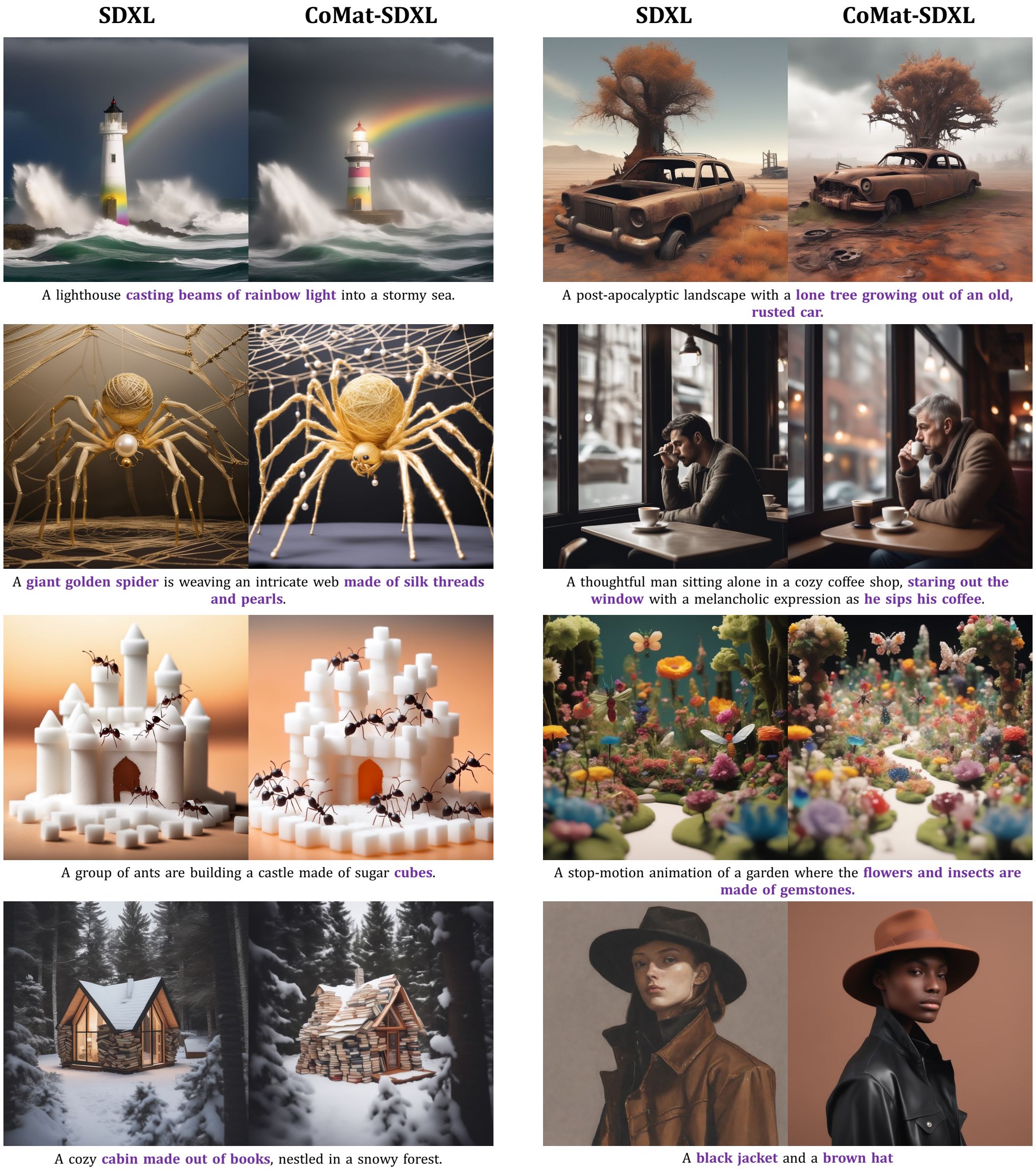
Inside a **dimly lit room**, the low luminance emanates from a bedside lamp casting a soft glow upon the nightstand. There lies a travel magazine, **its pages open to a vivid illustration of a car driving along a picturesque landscape**. Positioned next to the image is a **light pink toothbrush**, its bristles glistening in the ambient light. Beside the magazine, the **textured fabric of the bedspread is just discernible**, contributing to the composed and quiet scene.

A brightly colored hot air balloon with vibrant stripes of red, yellow, and blue hangs in the clear sky, its large round shape contrasting against the fluffy white clouds. Below it, a **sleek black scooter with red accents** speeds along a concrete pathway, **its rider leaning forward in a hurry**. The balloon moves at a leisurely pace, starkly contrasting with the frenetic energy of the scooter's rapid movement on the ground.

On a reflective **metallic table**, there is a **brightly colored handbag featuring a floral pattern** next to a freshly sliced avocado, its green flesh and brown pit providing a natural contrast to the industrial surface. The table is set for lunch, with **silverware** and **a clear glass water bottle** positioned neatly beside the avocado. The juxtaposition of the colorful fashion accessory and the rich texture of the avocado creates a striking visual amidst the midday meal setting.

A deep red rose with plush petals sits elegantly coiled atop an **ivory, intricately patterned lace napkin**. The napkin rests on a rustic wooden table that contributes to the charming garden setting. **As the late evening sun casts a warm golden hue over the area, the shadows of surrounding foliage dance gently around the rose**, enhancing the romantic ambiance. Nearby, **the green leaves of the garden plants** provide a fresh and verdant backdrop to the scene.

Figure 4: More Comparisons between SDXL and CoMat-SDXL on complex prompts. All pairs are generated with the same random seed.



A lighthouse **casting beams of rainbow light** into a stormy sea.

A post-apocalyptic landscape with a **lone tree growing out of an old, rusted car.**

A **giant golden spider** is weaving an intricate web **made of silk threads and pearls**.

A thoughtful man sitting alone in a cozy coffee shop, **staring out the window** with a melancholic expression as **he sips his coffee**.

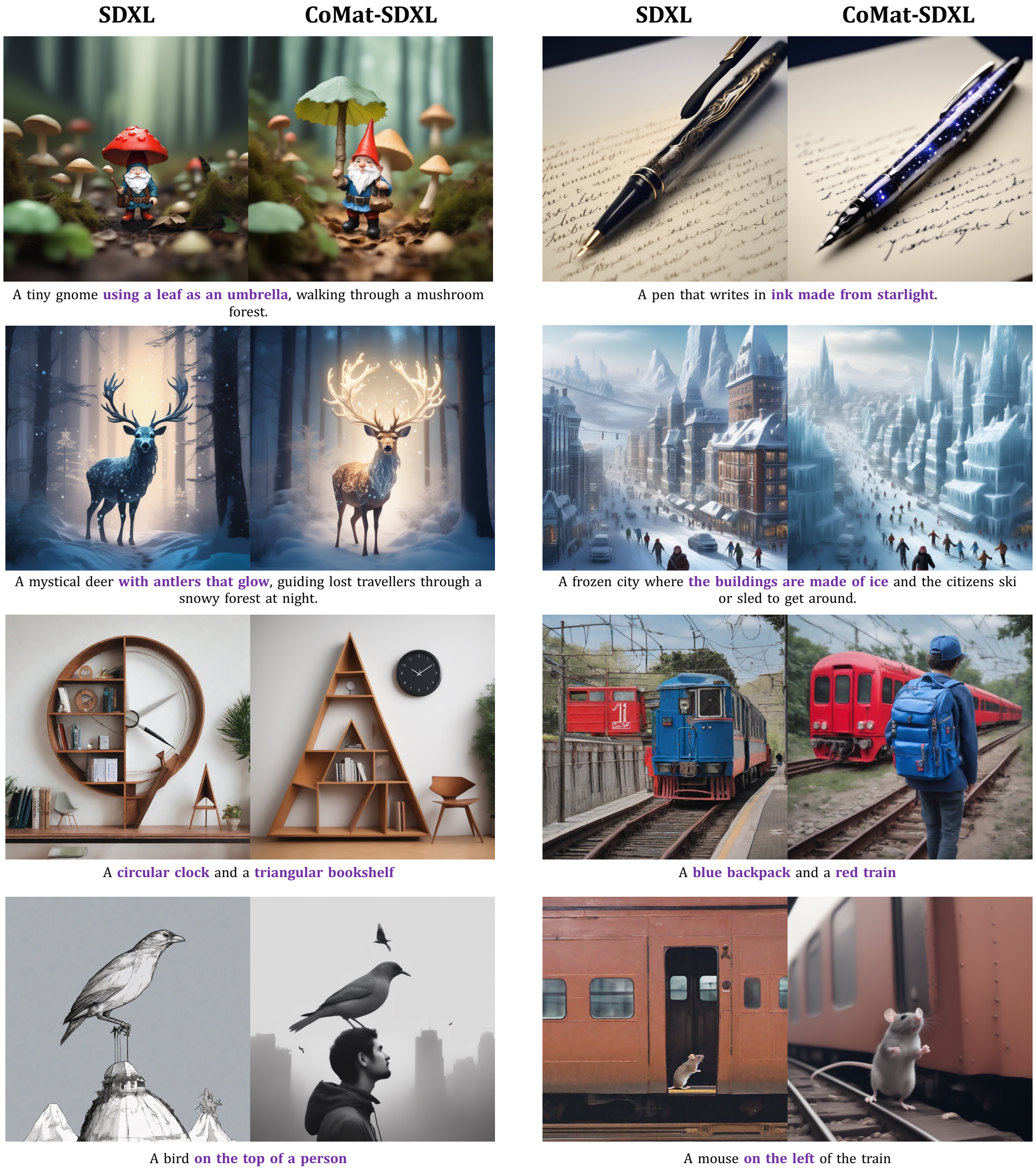
A group of ants are building a castle made of sugar **cubes**.

A stop-motion animation of a garden where the **flowers and insects are made of gemstones.**

A cozy **cabin made out of books**, nestled in a snowy forest.

A **black jacket** and a **brown hat**

Figure 5: More Comparisons between SDXL and CoMat-SDXL. All pairs are generated with the same random seed.



A tiny gnome **using a leaf as an umbrella**, walking through a mushroom forest.

A pen that writes in **ink made from starlight**.

A mystical deer **with antlers that glow**, guiding lost travellers through a snowy forest at night.

A frozen city where **the buildings are made of ice** and the citizens ski or sled to get around.

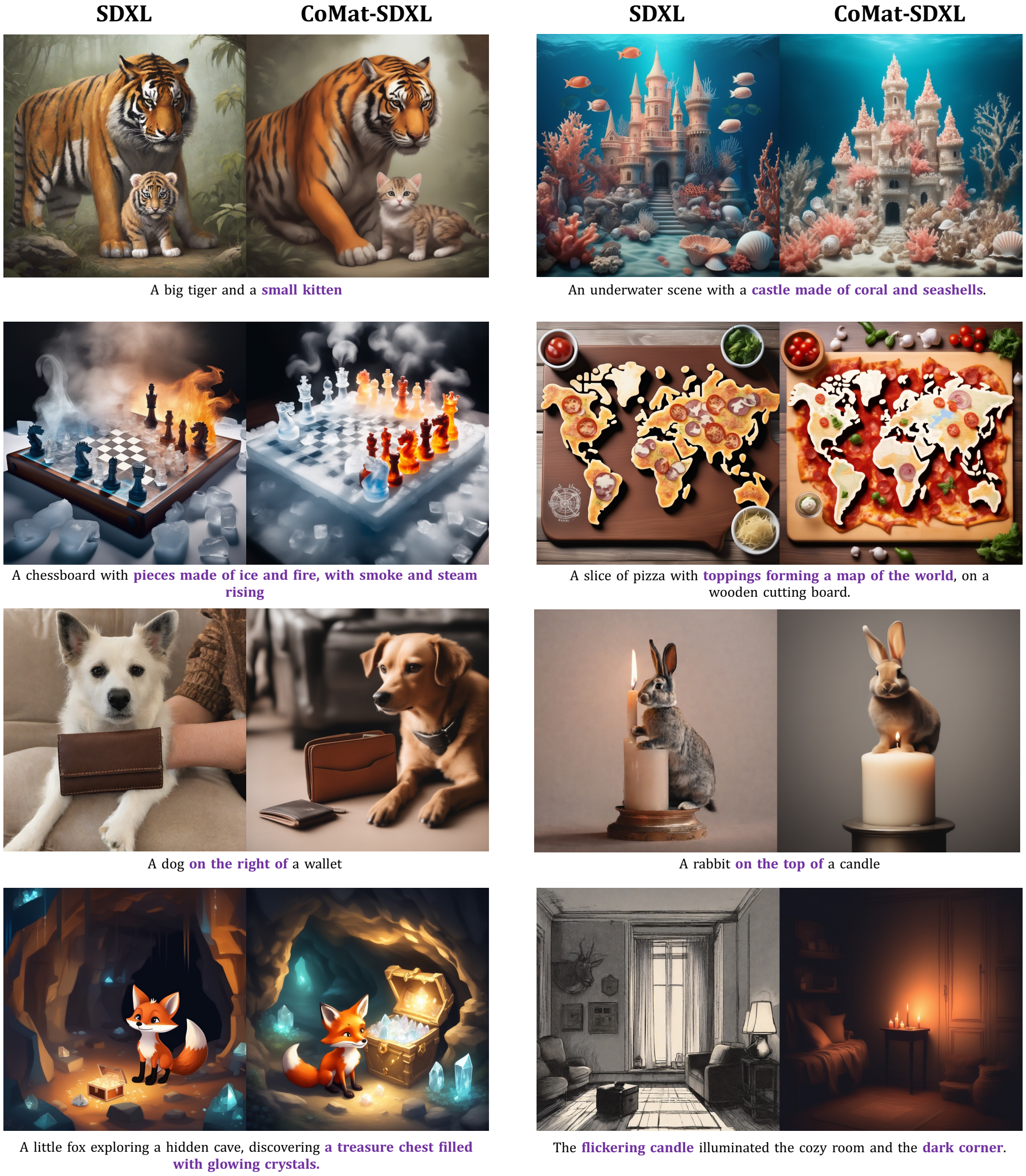
A **circular clock** and a **triangular bookshelf**

A **blue backpack** and a **red train**

A bird **on the top of a person**

A mouse **on the left** of the train

Figure 6: More Comparisons between SDXL and CoMat-SDXL. All pairs are generated with the same random seed.



A big tiger and a **small kitten**

An underwater scene with a **castle made of coral and seashells**.

A chessboard with **pieces made of ice and fire, with smoke and steam rising**

A slice of pizza with **toppings forming a map of the world**, on a wooden cutting board.

A dog **on the right of** a wallet

A rabbit **on the top of** a candle

A little fox exploring a hidden cave, discovering **a treasure chest filled with glowing crystals.**

The **flickering candle** illuminated the cozy room and the **dark corner**.

Figure 7: More Comparisons between SDXL and CoMat-SDXL. All pairs are generated with the same random seed.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024.

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[6] Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.

[7] Guian Fang, Zutao Jiang, Jianhua Han, Guangsong Lu, Hang Xu, and Xiaodan Liang. Boosting text-to-image diffusion models with fine-grained semantic rewards. *arXiv preprint arXiv:2305.19599*, 2023.

[8] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768*, 2024.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[11] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

[12] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[13] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.

[14] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. *arXiv preprint arXiv:2403.07304*, 2024.

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[16] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[20] Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, and Yu-Gang Jiang. Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *arXiv preprint arXiv:2404.12966*, 2024.

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[23] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[25] 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.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[27] 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.

[28] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

[29] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[30] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[31] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[32] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[33] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Grounding diffusion with token-level supervision. *arXiv preprint arXiv:2312.03626*, 2023.

[34] Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. 2024.

[35] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*, 2024.

[36] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[37] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

[38] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

[39] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *ICLR 2024*, 2023.

[40] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *ICLR 2024*, 2023.

[41] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

[42] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024.

[43] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[44] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.

[45] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.