

EGGen: Image Generation with Multi-entity Prior Learning through Entity Guidance

Anonymous Authors

1 DISCUSSIONS

Contribution and Novelty. To tackle the inconsistent multi-entity representation (IMR) challenge, we initially conducted thorough analyses of the diffusion process and attention mechanisms. Our findings suggest that the challenges associated with IMR predominantly arise from cross-attention mechanisms. Based on these analyses, we developed the entity guidance generation mechanism, which preserves the integrity of the original diffusion model parameters by integrating auxiliary plug-in networks. We have enhanced the stable diffusion model by decomposing complex prompts into separate entity-specific prompts, each defined within its bounding boxes. This approach facilitates a shift from multi-entity to single-entity generation within the cross-attention layers. Crucially, we have implemented entity-centric cross-attention layers that concentrate on individual entities to maintain their distinctiveness and precision. Alongside this, global entity alignment layers are utilized to refine the cross-attention maps, employing multi-entity priors for accurate positioning and attribute detail. Moreover, we incorporate a linear attenuation module that gradually diminishes the influence of these layers during inference to avoid oversaturation and maintain the fidelity of the generated images. Our extensive experiments show that this entity guidance generation improves the capability of existing text-to-image models to produce detailed, multi-entity images.

Limitations. While our approach is innovative and improves multi-entity generation, there remain several limitations, primarily due to the base model’s inherent capabilities. One challenge is generating detailed entities, including textures and finer details. This difficulty largely arises from our method’s focus on contour structure due to the integration of multi-centric prior information, which often overlooks these specific aspects. This reflects the base diffusion model’s limitations in accurately rendering such details. Perhaps opting for a more robust model like SDXL could address these shortcomings. Additionally, accurately capturing complex relationships, such as overlaps and interactions between entities, continues to pose challenges. For instance, generating intricate hand interactions during a dance sequence proves difficult. These complex and dynamic scenarios necessitate further enhancements and refinements in our approach to effectively depict and embody such complex relationships.

Future Work. Building on our current achievements, we identify two key areas for future development in text-based multi-entity image generation:

- **Enhancing Model Capacity:** To enhance our model’s capacity, we plan to integrate the SDXL-base model and enrich our dataset with a more diverse range of entities. This expansion will facilitate ongoing refinement, improving our ability to accurately represent multiple entities.

- **Complex Relationships:** We aim to incorporate additional multi-entity priors, such as depth perception and canny edge detection, to enhance the detail and realism of generated human figures. This initiative seeks to address the shortcomings of relying solely on contour information and will better capture the dynamic interactions between entities.
- **Enhancing Model Efficiency:** A challenge is integrating GLIGEN’s placement strategy into the entity-centric cross-attention layers, aiming to bypass the gated attention layers. This adjustment is expected to streamline the model’s processing efficiency and improve response times during image generation tasks.

2 ETHICAL AND SOCIAL IMPACTS

Our research into text-based multi-entity generation raises several ethical and social issues. The primary concern is the potential for misuse. This technology could enable the creation of more sophisticated and believable scenarios for spreading misinformation. Examples include fabricating images of crowds, events, or interactions that never happened, which could then be used to bolster false narratives or disseminate fake news. Additionally, generating images that involve multiple entities poses significant questions about consent and privacy. This is particularly problematic if the entities resemble real individuals or replicate their interactions. As a result, it is crucial to establish clear guidelines for ethically creating fictional scenarios, especially those that might depict controversial or damaging situations. In conclusion, while our method for multi-entity generation offers possibilities for innovative and inclusive image creation, it necessitates a firm commitment to ethical practices to mitigate potential harm.

3 DETAILED IMPLEMENTATION

Implementation Details. As described in the main text, during training, we use the AdamW optimizer [3] with a fixed learning rate of 0.00001 and weight decay of 0.01 for 10 epochs, and we set $\lambda = 10$ for loss control. In the inference stage, we adopt DDIM sampler [4] with 50 steps and set the guidance scale to 7.5. All experiments are performed on $8 \times$ Nvidia Tesla V100 GPUs. The detailed parameters are listed in the following Table 1.

Structure Details. The total trainable parameters are the three-layer MLP from the proposed ECA layers and the four-layer CNN network (Three convolution layers plus one SE Net layer) in GEA. At the same time, ECA and GEA modules are exclusively implemented at resolutions of 8×8 and 16×16 . The detailed implementations are shown in Table 2.

4 DETAILS OF LLMS

Given a complex multi-entity prompt, GPT-4-Vision can identify the entities and their attributes, leading to the generation of a

Table 1: List of implementation settings for both training and inference stages.

Implementation	Setting
Training Noise Scheduler	DDPM [1]
Epoch	10
Batch size per GPU	8
Optimizer	Adam
Learning rate	0.00001
weight decay	0.01
loss ratio λ	10
Training timestamp	1000
Training image size	512×512
Training HED image size	512×512
Sampling Noise Scheduler	DDIM [4]
Inference step	50
guidance scale	7.5
Inference image size	512×512

Table 2: Details of each network in ECA and GEA. "Linear layer: d_i " denotes a linear layer with d_i output channels; "Conv: 5×5 c128 s1" refers to a convolutional layer with a kernel size of 5×5 , 128 output channels, and a stride of 1.

MLP (ECA)	CNN (GEA)
Linear layer: d_i	Conv: 5×5 c128 s1
SiLU	BatchNorm + SiLU
Linear layer: d_i	Conv: 5×5 c128 s1
SiLU	BatchNorm + SiLU
Linear layer: d_i	SE Net [2]
	Conv: 5×5 c8 s1

reorganized global prompt and individual entity prompts. Moreover, it assigns bounding box attributes to each entity, predicted by the language model after evaluating the image's overall layout. Below, we outline the steps for effectively utilizing GPT-4-Vision with prompting engineering:

1). Task Specification: *You will be provided with a complex prompt that includes multiple entities. Your tasks include:*

- Describing each entity with a specific phrase.
- Predicting bounding boxes for each entity mentioned in the prompt.
- Using a phrase to describe the background scene.
- Combining entity prompts using "and" along with the background prompt to form the global prompt.
- You may make appropriate assumptions to supplement missing information.

2). Supporting details:

- *Image Dimensions:* The images are 512×512 pixels.
- *Coordinates:* The top-left corner is at $[0, 0]$, and the bottom-right corner is at $[511, 511]$.

3). Output formats:: *Format your outputs as follows:*

- Each entity should be in a dictionary format:
 $\{ \text{"Entity name": Entity prompt, "box": [top-left x, top-left y, width, height]} \}$
- The background scene should be formatted as:

$\{ \text{"background": Backgroundprompt} \}$

- The global prompt should be formatted as:

$\{ \text{"globalprompt": Globalprompt} \}$

- Combine all the returned information into a JSON file.

4). Example:

Caption: "A golden dog on the left and a black cat on the right are playing in the yard."

Returns:

```
{
  "object 1": {"dog": "a golden dog", \
    "box": [0, 150, 236, 250]},
  "object 2": {"cat": "a black dog", \
    "box": [280, 150, 232, 250]},
  "background": "a green yard",
  "global prompt": "a golden dog and a black dog
    in the green yard"
}
```

5). Input caption: *Provide your input caption here.*

5 DETAILS OF USED PROMPTS

We also include the complete prompts used for Figures 1 and 7.

Figure 1 (a):

```
{
  "object 1": {"rose": "a blue rose", \
    "box": [60, 260, 64, 64]},
  "object 2": {"rose": "a red rose", \
    "box": [210, 310, 64, 64]},
  "object 3": {"rose": "a yellow rose", \
    "box": [140, 410, 64, 64]},
  "object 4": {"rose": "a white rose", \
    "box": [280, 410, 64, 64]},
  "object 5": {"rose": "an orange rose", \
    "box": [360, 260, 64, 64]},
  "object 6": {"rose": "a pink rose", \
    "box": [360, 410, 64, 64]},
  "object 7": {"cloud": "a soft white cloud", \
    "box": [50, 50, 100, 100]},
  "object 8": {"cloud": "a soft white cloud", \
    "box": [350, 50, 100, 100]},
  "object 9": {"sky": "a clean blue sky", \
    "box": [0, 0, 511, 200]},
  "background": "A green Garden",
  "global prompt": "a blue rose and a red rose
    and a yellow rose and a white rose
    and an orange rose and a pink rose
    and a soft white cloud and a soft white cloud
    and a clean blue sky in a green garden."
}
```

Figure 1 (b):

Prompt: a red car and a pink elephant in a grey empty street.



Figure 1: An example of failure case. While our model successfully generated the target entities with the correct color attributes, it failed to produce a realistic elephant in the intended pink color.

```
{
  "object 1": {"style": "a cartoon style", \
    "box": [0, 0, 511, 511]},
  "object 2": {"dog": "a golden dog", \
    "box": [0, 300, 200, 200]},
  "object 3": {"cat": "a black dog", \
    "box": [300, 300, 200, 200]},
  "object 4": {"house": "a red house", \
    "box": [100, 0, 310, 250]},
  "object 5": {"tree": "a green tree", \
    "box": [0, 0, 100, 250]},
  "object 6": {"tree": "a green tree", \
    "box": [410, 0, 100, 250]},
  "background": "a green grass scene",
  "global prompt": "a cartoon style
and a golden dog and a black cat
and a red house and a green tree
and a green tree in a green grass scene"
}
```

Figure 7 motorcycle:

```
{
  "object 1": {"motorcycle": "a blue motorcycle", \
    "box": [250, 100, 200, 250]},
  "object 2": {"car": "a red car", \
    "box": [0, 100, 200, 200]},
  "background": "a gray parking lot",
  "global prompt": "a blue motorcycle and a red car
in a gray parking lot"
}
```

Figure 7 truck:

```
{
  "object 1": {"car": "a red car", \
    "box": [250, 150, 200, 200]},
  "object 2": {"truck": "a black truck", \

```

```
"box": [0, 100, 200, 300]},
"background": "a gray parking lot",
"global prompt": "a red car and a black truck
in a gray parking lot"
}
```

6 ADDITIONAL RESULTS

Due to space constraints in the main paper, we only included a portion of the results. In this section, we will present further results and additional analyses.

6.1 An Example of Failure Case

In this subsection, we delve into a specific failure case encountered during our experiments to gain a deeper understanding of the limitations inherent in our methodology and to identify potential areas for improvement. The failure occurred while processing the prompt "a red car and a pink elephant in a grey empty street", as shown in Figure 1. This prompt was chosen to assess the system's capability to generate a pink elephant, which does not naturally exist. This incident highlights the critical need for diverse datasets and the model's capacity to generalize across complex prompts that blend various elements of reality.

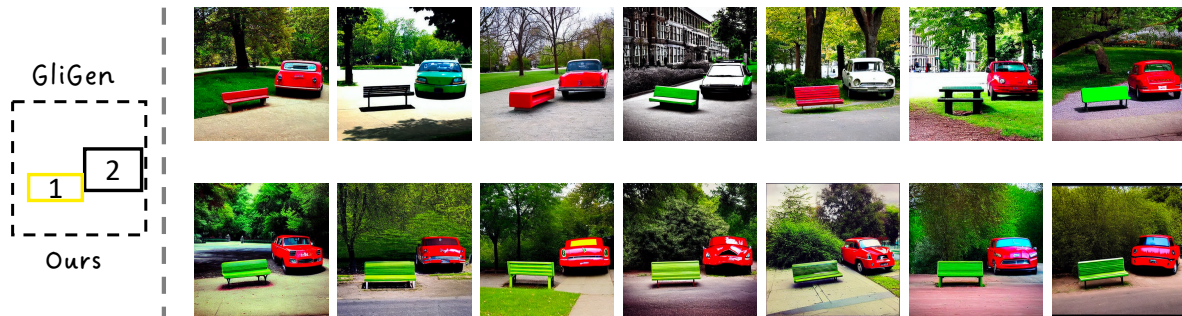
6.2 Other Visualizations in T2I Comp-Bench

In this subsection, we expand upon additional visualizations derived from the T2I Comp-Bench, which were not included in the main paper due to space constraints. The results are shown in Figure 2.

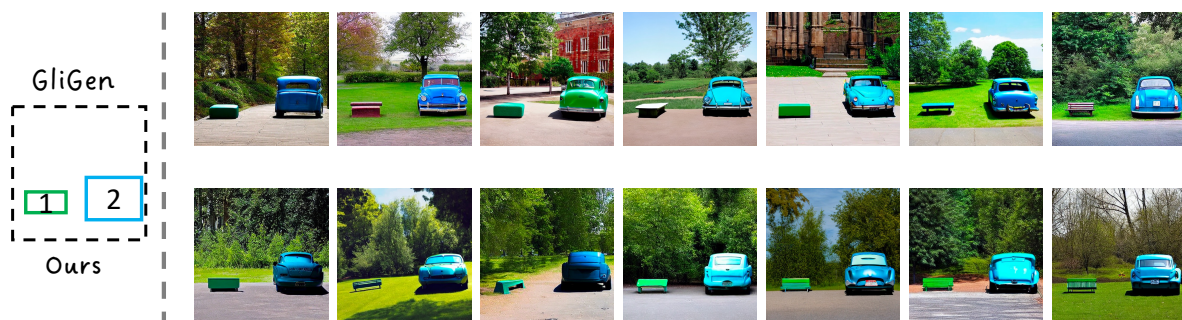
REFERENCES

- [1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [2] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [3] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- [4] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

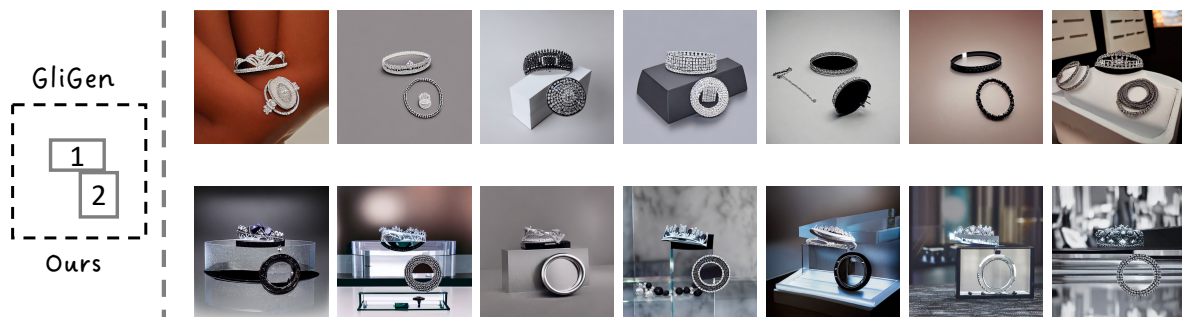
Prompt: a green bench and a red car in a various urban park.



Prompt: a green bench and a blue car in a varies outdoor park.



Prompt: a silver diamond tiara and a silver round bracelet in a dark display scene.



Prompt: a yellow and black warning sign and a gold or silver trophy in a tint plain background.

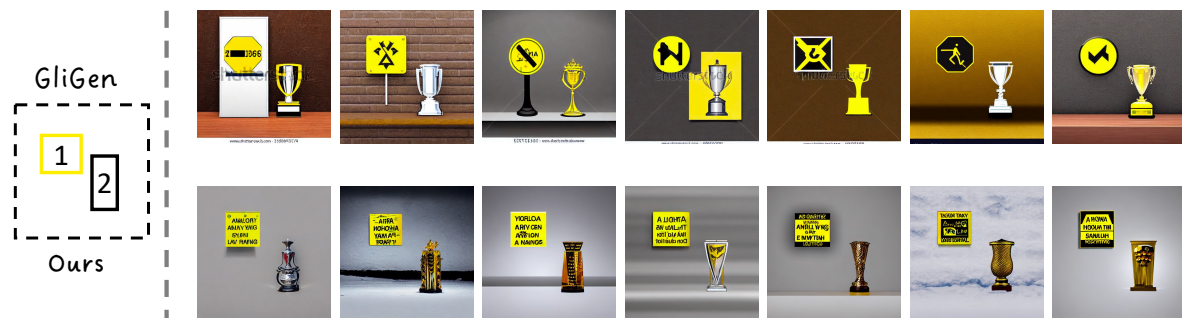


Figure 2: Additional qualitative comparison with baseline method in T2I Comp-Bench. Our method demonstrates a substantial enhancement in handling complex multi-entity prompts, providing a promising avenue for future research in advanced image synthesis.