Supplementary Material
MultiGen: Zero-shot Image Generation from Multi-modal Prompts

Zhi-Fan Wu, Lianghua Huang, Wei Wang, Yanheng Wei, and Yu Liu

Alibaba Group

1 More Qualitative Results

In this section, we present the complete qualitative results of MultiGen in comparison to other methods. It can be observed that MultiGen continues to maintain its advantage in most of cases.

When given a single image and a text prompt, MultiGen consistently outperforms InstructPix2Pix [1], BLIP-Diffusion [2], and Kosmos-G [3], with results illustrated in Fig. 1. As we mentioned in the main paper, compared to other methods, MultiGen is capable of better integrating the object with the background described in the text prompt without being disrupted by the original image's background. Secondly, MultiGen is more adept at making reasonable alterations to the object in accordance with the description of the text prompt. Lastly, MultiGen also has an advantage in ensuring the fidelity of the object.

Next, we showcase the full qualitative comparison results of image generation given multiple images and text prompt, as presented in Fig. 2. It is evident that, in comparison to Kosmos-G [3] and Emu2 [4], MultiGen continues to demonstrate a clear advantage. Firstly, MultiGen does not omit any objects when combining multiple objects. Secondly, MultiGen can more effectively blend the backgrounds of the two images together. For instance, with the prompt "A horse running on the sea", MultiGen successfully integrates the horse and the sea, whereas the images generated by Kosmos-G and Emu2 contain some grassland from the original image of the horse. Lastly, the results produced by MultiGen are more visually coherent.

Overall, compared to existing methods, MultiGen possesses a significant advantage in comprehending multi-modal prompts. Beyond the image and text modalities, as demonstrated in the main paper, MultiGen also supports coordinate modality. This endows MultiGen with powerful and flexible image generation capabilities.

2 More Ablation Studies

In this section, we validate the generation results of the feature model and the coordinate model, demonstrating their effectiveness in compensating for missing modalities when the corresponding modalities are not provided in the prompt.

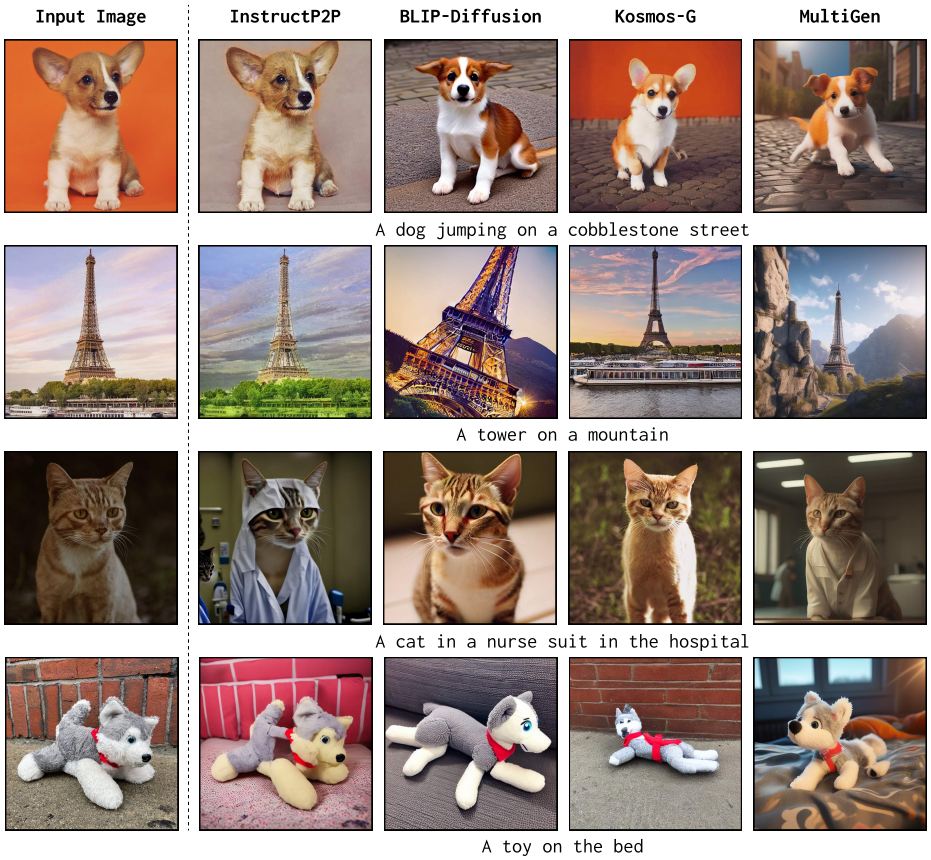

Fig. 1: Full qualitative comparison results of zero-shot image generation results from *single-object* multi-modal prompts. We compare MultiGen (fifth column) with InstructP2P [1] (second column), BLIP-Diffusion [2] (third column) and Kosmos-G [3] (fourth column). While ensuring object fidelity, MultiGen can make changes to objects and backgrounds according to the requirements of the text prompt.

In Fig. 3, we conduct the ablation study on the feature model. In row (c), we present images generated by MultiGen based on the object-level text, image, and coordinates obtained from the original pictures shown in row (a). In row (d), we generate the image features when the object-level image is not provided in the prompt, using the feature model. The results indicate that the feature model is capable of generating reasonable object-level image features based on object-level text and coordinates. This enables our model to handle image generation issues when the image modality is missing.

We then perform the ablation study on the coordinate model, and present the results in Fig.4. Rows (c-d) utilize the same global and object-level text prompts as in rows (a-b). However, we assume that coordinates have not been

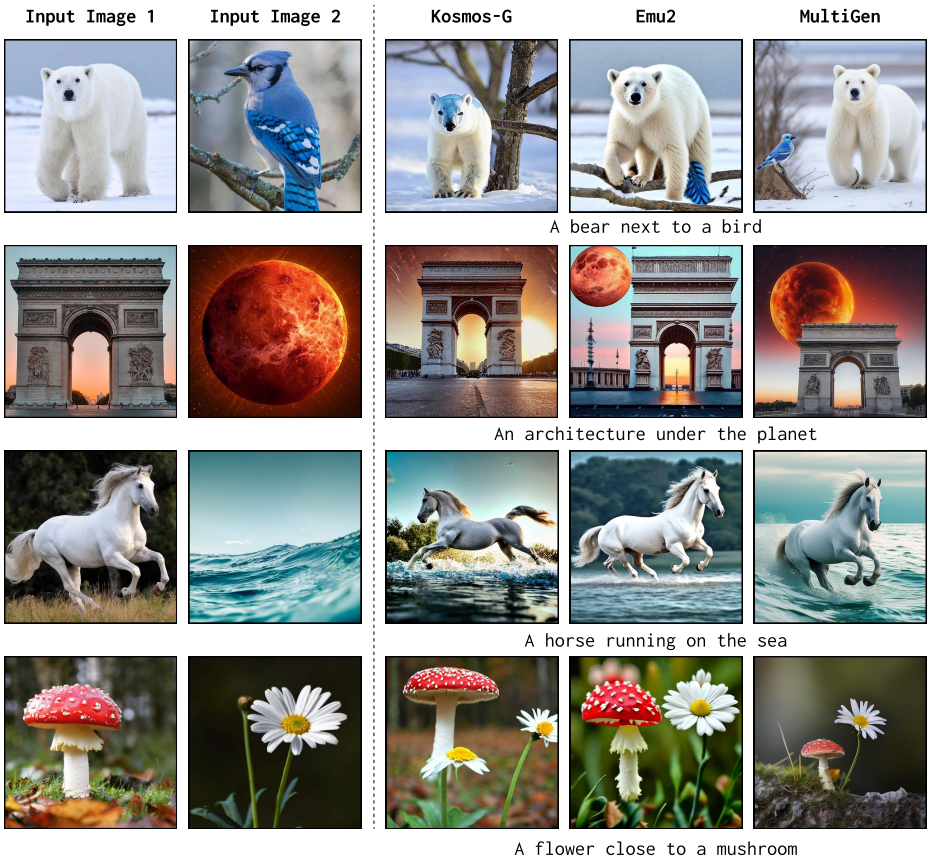

Fig. 2: Full qualitative comparison results of zero-shot image generation results from *multi-object* multi-modal prompts. We compare MultiGen (fifth column) with Kosmos-G [3] (third column) and Emu2 [4] (fourth column). When combining images of multiple objects, it is evident that MultiGen does not result in missing objects. Additionally, it excels in seamlessly fusing images with backgrounds.

provided, so the model needs to independently generate reasonable coordinates, based on which it generates image features and, ultimately, the image itself. Row (c) displays the final generated results, and row (d) shows the underlying coordinates. The results demonstrate that our coordinate model can effectively generate plausible layouts, positioning objects in reasonable locations. This allows our model to still generate images even when the coordinate modality is not provided. The coordinate model also ensures that MultiGen rarely misses objects when generating images based on prompts involving multiple objects, as we have displayed in Sec.1 and Fig. 2.

In summary, our experiments validate the effectiveness of the feature model and the coordinate model. This renders MultiGen a robust and versatile model

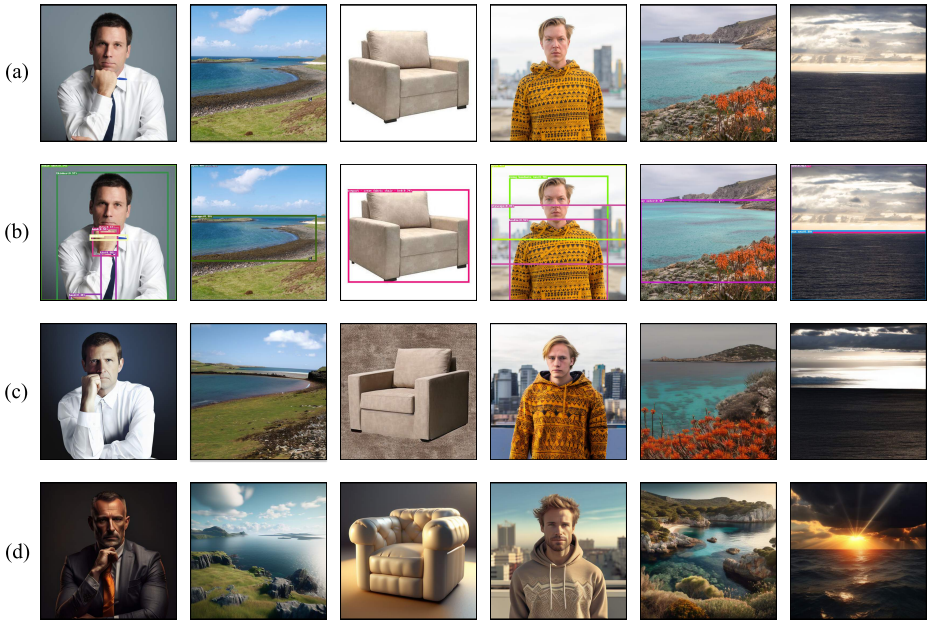

Fig. 3: Ablation study on feature model. (a) Randomly sampled original images. (b) Open-set detection results to obtain object-level images along with their corresponding text and coordinates. (c) MultiGen generation results. Images are generated based on the provided object-level images, texts, and coordinates, where image features are extracted using SigLIP-SO400M-384. (d) MultiGen generation results where images features are generated by the feature generation model.

capable of supporting prompts derived from any combination of object-level modalities, including images, text, and coordinates. Even when certain modalities are missing from the prompt, MultiGen can still effectively generate images.

3 Visualization of Generation Process

MultiGen employs the coordinate model and image feature model based on diffusion process to generate missing modalities when they are not available. In Fig. 5, we present the complete image generation process when only a text prompt is provided. Initially, the coordinate model generates coordinates corresponding to the objects based on the object-level text. We present the results of coordinate generation in the first row, and it can be observed that the coordinate model is capable of generating reasonable coordinates based on the text prompt and object-level text. Subsequently, using the text and coordinates, we generate object-level image features through the feature model. These elements are then integrated into augmented tokens to facilitate image generation. Remarkably, even with only a text prompt, MultiGen successfully generates high-quality images. This flexibility enables our method to effectively adapt to diverse scenarios.

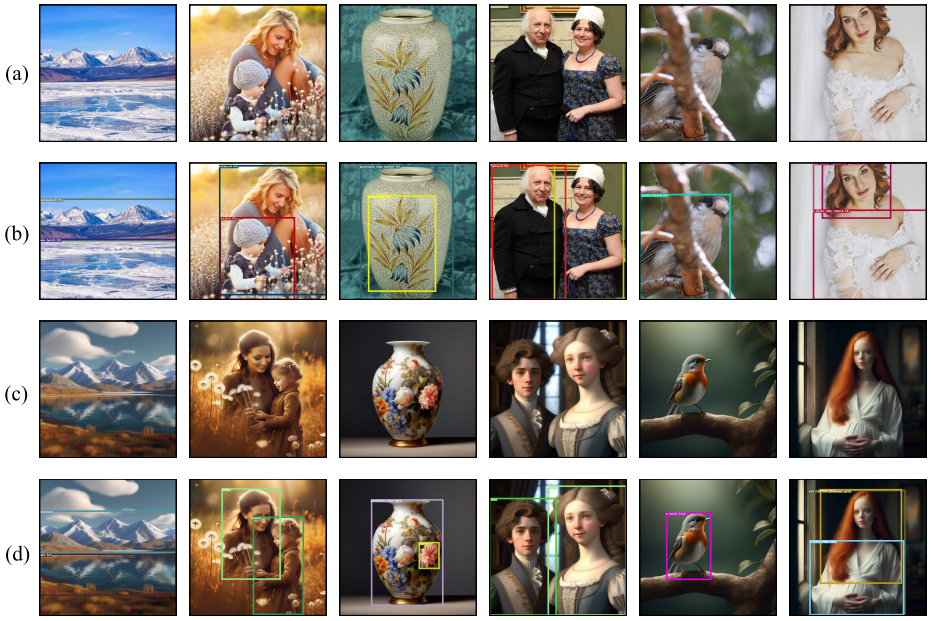

Fig. 4: Ablation study on coordinate model. (a) Randomly sampled original images. (b) Open-set detection results to obtain object-level images along with their corresponding text and coordinates. (c-d) MultiGen generation results. We use the same object-level text as in (b) and leverage the coordinates provided by the coordinate generation model. Subsequently, we generate the object-level image feature by feature model. (c) displays the final results, and (d) reveals the underlying coordinates.

4 Limitations

Here, we discuss the limitations of MultiGen. Like all zero-shot image generation methods that handle multiple modalities, MultiGen finds it easier to generate common objects, while accurately reproducing rare objects can be more challenging. In other words, achieving the same level of object detail preservation as tuning-based methods poses a challenge. We observe that as the training steps increase, the similarity between generated objects and original objects improves. Additionally, it is possible to consider introducing more auxiliary information to maintain the shape of objects. These ideas will be our next direction for improvement. Besides, when coordinates are not given, there are occasional instances where the coordinates generated by the coordinate model are not reasonable. This may be attributed to the quality of the training data and the relatively fewer training steps of the model. We leave this as direction for future research. Despite these limitations, MultiGen remains an important step in generating images from multi-modal prompts due to its ability to effectively integrate multiple modalities and generate images based on prompts involving multiple objects.

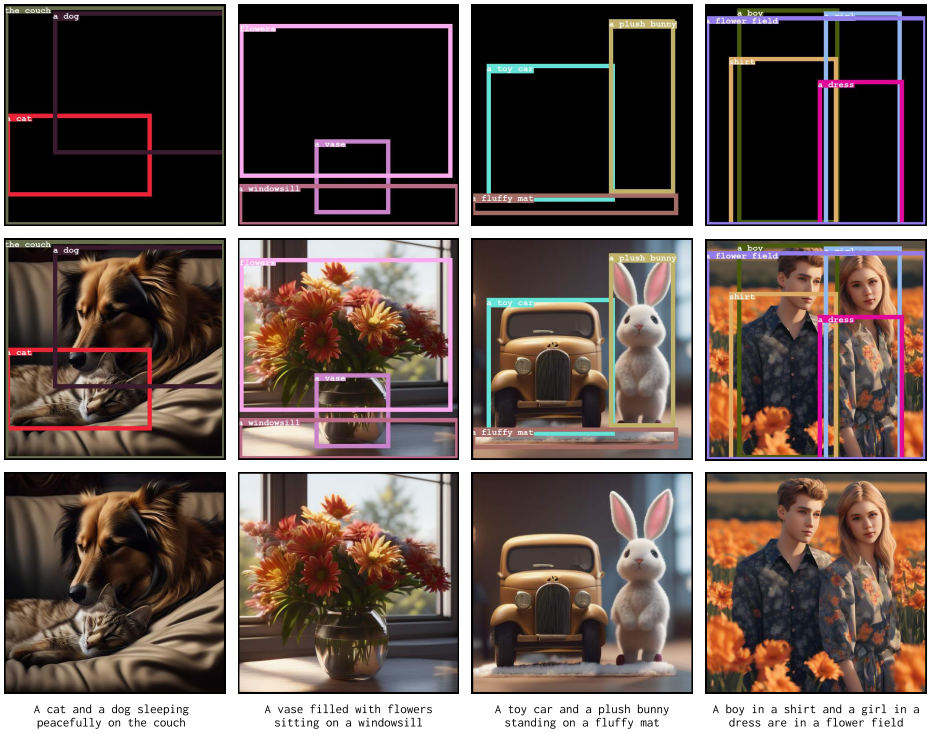

Fig. 5: Visualization of generation process. MultiGen has the ability to generate missing modalities. We demonstrate this by presenting the coordinates generated by MultiGen when provided with only a text prompt (first row), the corresponding images generated based on these coordinates (second row), and the final results (third row). Remarkably, even without explicitly providing coordinates, MultiGen is capable of generating reasonable ones. Furthermore, MultiGen can generate realistic image features and produces the desired image.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. arXiv:2211.09800 (2022)
2. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv:2305.14720 (2023)
3. Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., Wei, F.: Kosmos-G: Generating images in context with multimodal large language models. arXiv:2310.02992 (2023)
4. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. arXiv:2312.13286 (2023)