# MoEdit: On Learning Quantity Perception for Multi-object Image Editing

## Supplementary Material

## A. Details of Network Architecture

This section provides a detailed explanation of the design of two key modules in MoEdit: the Feature Compensation (FeCom) module and the Quantity Attention (QTTN) module. It covers their structural composition, dimensional transformations, inputs, and outputs. The FeCom module is designed to minimize interlacing during the extraction of object attributes, while the QTTN module ensures consistent quantity preservation.

### A.1. Feature Compensation Module

#### A.1.1 Overall

This module processes an input image $I$ and a text prompt $c_q$, which specifies objects and their quantities, to generate an enhanced feature $I_g$, characterized by distinct and separable object attributes. The module is composed of three main components: an image encoder (from CLIP [38]), a text encoder (from SDXL [36]), and a feature attention mechanism. The feature attention mechanism aligns textual and visual representations of object and quantity information, enabling effective extraction of object attributes. Fig. 9 provides a detailed overview of the FeCom module, with each step annotated to indicate its function. The notation "$\rightarrow$ (...)" represents the output dimensions at each stage.

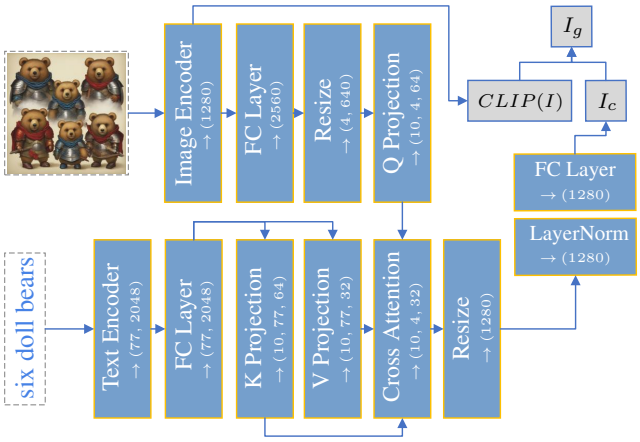


Figure 9. The detailed structure diagram of FeCom module. The image encoder is derived from CLIP [38], and the text encoder is sourced from SDXL [36]. The module takes two inputs—images and text prompts—and generates a single output: enhanced features $I_g$. In this process, $CLIP(I)$ denotes the inferior image features extracted by the image encoder of CLIP, while $I_c$ represents the compensation features used to enhance it. Each step is explicitly defined, with "$\rightarrow$ (...)" denoting the dimensionality of the corresponding output.

#### A.1.2 Discussion of LayerNorm

The FeCom module incorporates a LayerNorm layer, which is critical for balancing original image features, aligning textual and visual information, and enabling the QTTN module to effectively consider each object both individually and part of the whole image. Fig. 10 underscores its significance, with Reference representing the input image.

In Fig. 10 (a), "m/o LayerNorm" denotes the results obtained when the LayerNorm layer from Fig. 9 is repositioned to the final stage, after the last fully connected (FC) layer, during training. This adjustment leads to a marked decline in both quantity consistent perception and object attributes extraction. Conversely, "w/o LayerNorm" refers to the scenario where the LayerNorm layer is entirely removed during training. While this configuration preserves quantity consistency, it introduces aliasing of object attributes.

In Fig. 10 (b), to elucidate these degradations, we analyze the original image features extracted by the image encoder of CLIP and the compensation features generated by the FeCom module. "Compensation Features (m/o)" refers to compensation features obtained under the "m/o LayerNorm" condition, with a value range of $[3, -3]$, while the original image features span $[4, -6]$. This narrow range suggests insufficient feature intensity, which fails to adequately mitigate the in-between interlacing in inferior original image features, thereby compromising quantity consistent perception and object attributes extraction. Conversely, "Compensation Features (w/o)" are produced under the "w/o LayerNorm" condition, with a value range of $[60, -80]$. This excessively broad range indicates oversuppression of original image features, resulting in substantial information loss, particularly in object attributes. In comparison, the compensation features constructed by the FeCom module in MoEdit (denoted as "Compensation Features (Ours)") fall within a balanced range of $[15, -10]$. This range effectively preserves the original image information while minimizing in-between interlacing.

The quantitative comparisons are shown in Table 4. Notably, the "m/o LayerNorm" have a significant impact on the overall performance of MoEdit, primarily due to their limited compensatory capacity. In contrast, under the "w/o LayerNorm" condition, despite the excessive strength of the compensation features, image aesthetics, quality, and numerical accuracy remain largely unaffected. This stability can be attributed to the fact that these compensation fea-

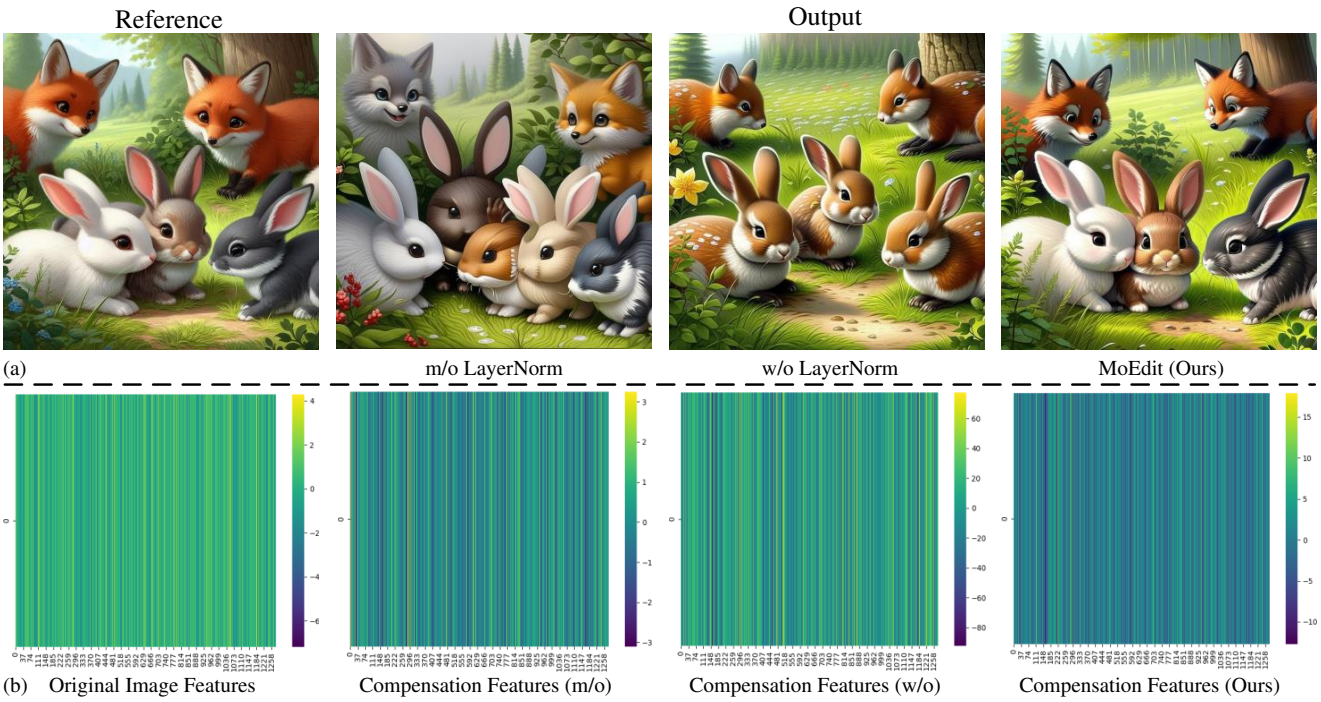


Figure 10. The importance of the LayerNorm layer. Please refer to Sec. A.1.2 for a comprehensive evaluation and detailed discussions regarding these configurations.

tures are derived from textual prompts containing object and quantity information. However, the loss of object attributes leads to a noticeable decline in textual-visual alignment and the similarity between the original and resulting images.

As demonstrated in Fig. 10 (a), MoEdit achieves superior performance visually in both preserving quantity consistency and extracting object attributes.

## A.2. Quantity Attention Module

### A.2.1 Overall

The module takes as input the image features $I_g$, enhanced by the FeCom module, and the noise $z_t^4$ from the fourth block $B_4$ of the U-Net. It outputs perception information of quantity consistensy that can be interpreted by the U-Net to regulate the entire editing process. The module comprises three key components: the extraction module $E_t$, attention interaction, and U-Net injection. $E_t$ is designed to disentangle each object attribute from the whole $I_g$ while simultaneously capturing the global information of $I_g$. The attention interaction component translates the information extracted by $E_t$ into a format that is interpretable by the U-Net. Finally, the U-Net injection component leverages this transformed information to control the image editing process, preserving quantity consistency. Fig. 11 presents the detailed structure of the QTTN module, with each step annotated to indicate its function. The notation "$\rightarrow$ (...)" represents the output dimensions at each stage.

| Method | NIQE ↓ | HyperIQA ↑ | CLIP Score | |
|---|---|---|---|---|
| | | | Whole ↑ | Edit ↑ |
| m/o LayerNorm | 2.8726 | 73.28 | 0.3005 | 0.2689 |
| w/o LayerNorm | 2.7081 | 75.99 | 0.3078 | 0.2778 |
| w/o Extraction | 2.7232 | 75.47 | 0.3105 | 0.2782 |
| MoEdit (Ours) | **2.6501** | **77.87** | **0.3274** | **0.2790** |
| **Method** | **LPIPS ↓** | **Numerical ↑** | **Q-Align** | |
| | | | **Quality ↑** | **Aesthetic ↑** |
| m/o LayerNorm | 0.3101 | 32.65 | 4.7511 | 4.4371 |
| w/o LayerNorm | 0.2943 | 84.77 | 4.8933 | 4.7552 |
| w/o Extraction | 0.2975 | 78.73 | 4.8344 | 4.6211 |
| MoEdit (Ours) | **0.2731** | **86.79** | **4.9219** | **4.8047** |

Table 4. Quantitative comparisons. "m/o LayerNorm" denotes the results obtained when the LayerNorm layer from Fig. 9 is repositioned to the final stage, after the last FC layer, during training. "w/o LayerNorm" refers to the scenario where the LayerNorm layer is entirely removed during training. "w/o Extraction" illustrates the absence of Extraction module in Fig. 11. For more discussion, please refer to Secs. A.1.2 and A.2.2.

### A.2.2 Discussion of Extraction

In the QTTN module, $E_t$ is a critical component. Although implemented by a single FC layer, $E_t$ extracts clear and structured information about each object both individually and part of the whole $I_g$. This functionality cannot be directly incorporated into the attention interaction process. Fig. 12 underscores the importance of $E_t$, with the "w/o

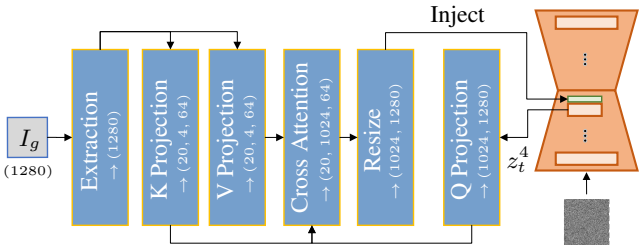


Figure 11. The detailed structure diagram of QTTN module. Each step is annotated with its specific purpose, with "→ (...)" denoting the dimensionality of the corresponding output. This module receives two inputs: $I_g$, originating from the FeCom module, and input noise $z_t^4$, from the fourth block $B_4$ of the U-Net. Its primary purpose is to generate a quantity consistent perception that can be effectively interpreted by the U-Net. The core of the Extraction module $E_t$ is a FC layer, which is designed to consider each object both individually and part of the whole $I_g$. The functionality of $E_t$ is detailed in Fig. 12.

Reference w/o Extraction MoEdit (Ours)

Figure 12. The importance of the $E_t$. Please refer to Sec. A.2.2 for a comprehensive evaluation and detailed discussions regarding these configurations.

Extraction" condition illustrating its absence.

Without $E_t$, the distinctive and separable object attributes in $I_g$ can still ensure quantity consistent perception. However, compared to MoEdit, the absence of $E_t$ leads to suboptimal object attributes extraction, reduced image details, and poor anatomical representations. This limitation arises because the Q, K, and V Projections within the attention interaction cannot effectively handle both the extraction of each object information and the transformation of information patterns simultaneously. Therefore, MoEdit incorporates an independent FC layer to ensure the high quality information of each object required for effective attention interactions. This unique functionality underscores the rationale for designating this FC layer as Extraction.

The quantitative comparisons are summarized in Table 4. Notably, "w/o Extraction" exhibits a decline across multiple metrics, including image aesthetics, quality, text-image alignment, similarity between original and resulting images, and numerical accuracy.

# B. Supplementary Experiments

## B.1. Full Qualitative Comparisons

In our main paper, MoEdit was comprehensively evaluated using six objective metrics, HyperIQA [42], AesBench [13], Q-Align [54], NIQE [32], CLIP Score [11], and LPIPS [62], and two subjective metrics, MOS and Numerical Accuracy. The evaluation compared MoEdit against seven methods: SSR-Encoder [63], $\lambda$-Eclipse [35], IP-Adapter [61], Blip-diffusion [24], MS-diffusion [49], Emu2 [43], and TurboEdit [56]. This section provides additional details on these comparison methods and summarizes the code sources for all methods and evaluation metrics in Table 5, facilitating future research replication.

**SSR-Encoder.** This method is tailored for subject-driven generative tasks, encoding selective subject representations. By focusing on extracting and representing subject-specific features, it enhances the ability to maintain subject consistency in generated outputs.

**$\lambda$-Eclipse.** A multi-concept text-to-image generation model that leverages the latent space of CLIP to achieve efficient multimodal alignment. This method allows the flexible integration of multiple concepts while preserving semantic consistency between textual and visual inputs.

**IP-Adapter.** A text-compatible image-prompt adapter designed for text-to-image diffusion models. The introduction of efficient adapter modules enables the model to utilize image prompts to enhance generation quality while maintaining compatibility with textual input.

**Blip-diffusion.** This method combines pre-trained subject representations with text-to-image generation and editing. By integrating precise subject modeling, it offers more controllable image generation and editing while improving semantic alignment between text and images.

**MS-diffusion.** A method for multi-subject image personalization that integrates layout guidance for structured scene control. This method generates multi-subject, layout-accurate images in zero-shot scenarios, enhancing the diversity and consistency of outputs.

**Emu2.** This study investigates the contextual learning capabilities of generative multimodal models, showcasing their ability to perform diverse tasks in unsupervised settings via input prompts. The findings highlight the potential of these models for cross-task generalization and implicit learning, underscoring their flexibility and adaptability.

**TurboEdit.** An efficient text-driven image editing method that rapidly generates modifications aligned with textual descriptions. By optimizing the editing workflow, it provides a real-time, interactive image editing experience.

Due to space constraints in the main paper, qualitative comparisons were limited to conditions 3, 7, and 9+. In this section, results for the remaining conditions are presented in Figs. 13–17. Furthermore, Fig. 18 offers additional visual

| Method | | URL |
|---|---|---|
| SSR-Encoder | [63] | https://github.com/Xiaojiu-z/SSR_Encoder |
| $\lambda$-Eclipse | [35] | https://github.com/eclipse-t2i/lambda-eclipse-inference |
| IP-Adapter | [61] | https://github.com/tencent-ailab/IP-Adapter |
| Blip-diffusion | [24] | https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion |
| MS-diffusion | [49] | https://github.com/MS-Diffusion/MS-Diffusion |
| Emu2 | [43] | https://github.com/baaivision/Emu |
| TurboEdit | [56] | https://betterze.github.io/TurboEdit/ |
| **Metric** | | **URL** |
| HyperIQA | [42] | https://github.com/SSL92/hyperIQA |
| Q-Align | [54] | https://github.com/Q-Future/Q-Align |
| AesBench | [13] | https://github.com/yipoh/AesBench |
| NIQE | [32] | https://github.com/csjunxu/Bovik_NIQE_SPL2013 |
| CLIP Score | [11] | https://github.com/Taited/clip-score |
| LPIPS | [62] | https://github.com/richzhang/PerceptualSimilarity |

Table 5. Code sources of seven comparison methods and six objective metrics

demonstrations of MoEdit output.

### B.2. Visualization of Ablation Results

**Scale variation.** Due to the limited space in the main paper, we presented only the results for $\lambda = 1$ with varying $\beta$ and $\beta = 1$ with varying $\lambda$. However, these results alone are insufficient to fully capture the mechanisms of the two modules. To address this, more comprehensive results are provided in Fig. 19, offering a clearer demonstration of the interaction between the modules.

When $\lambda$ is fixed, increasing $\beta$ enhances the ability of the QTTN module to disentangle each object attribute from the whole image and to extract global information, resulting in higher-quality outputs. This improvement, however, relies heavily on the input features $I_g$ of the QTTN module, which must exhibit sufficient distinction and separability of object attributes. Conversely, when $\beta$ is fixed, reducing $\lambda$ increases the degree of aliasing of object attributes. Consequently, the QTTN module becomes less effective at extracting information of each object both individual and part of the whole, leading to a deterioration in the output quality.

### B.3. Limitations

When the spatial relationship between objects and their surrounding 3D environment changes, MoEdit often struggles to deliver satisfactory editing results. This limitation stems from the baseline model, SDXL [36], which lacks the capability to effectively construct accurate 3D relationships in multi-object scenes. As illustrated in Fig. 20, prompts like *"on the table"* require consideration of not only the interaction between existing objects and their environment but also the 3D spatial alignment between the newly introduced table and multiple objects. This challenge results in MoEdit producing artifacts such as misaligned object segments, truncation, and loss of surrounding elements like clothing. Meanwhile, other methods exhibit even poorer performance, either failing to achieve meaningful edits or simply modifying the ground to a wooden texture, thereby creating a illusion of objects being placed on a table. To address these challenges, future work could focus on integrating 3D environmental data, transcending the limitations of 2D information, to achieve better alignment and coherence between multi-object and their scenes.

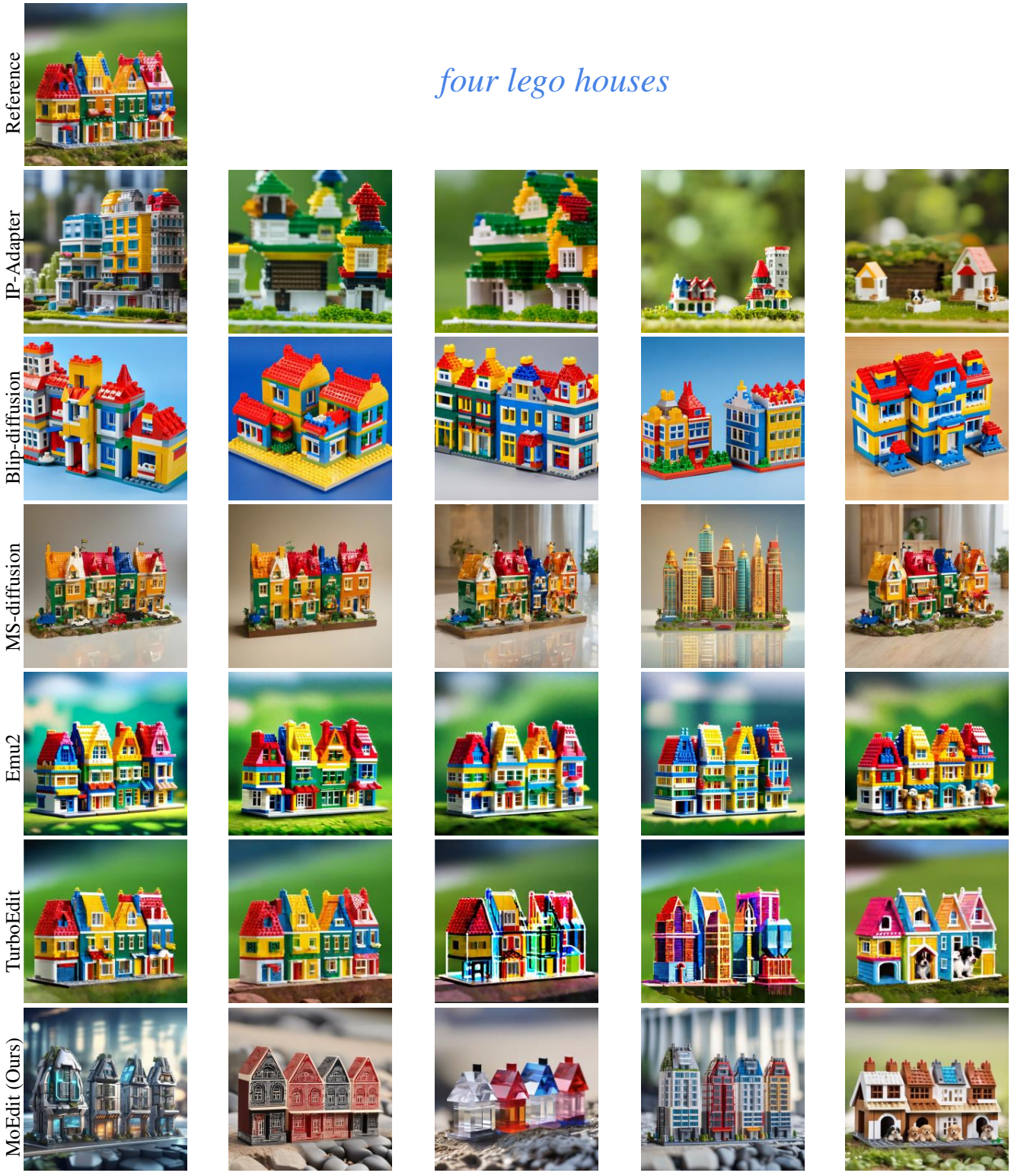


Figure 13. Qualitative comparisons among a set of four objects.

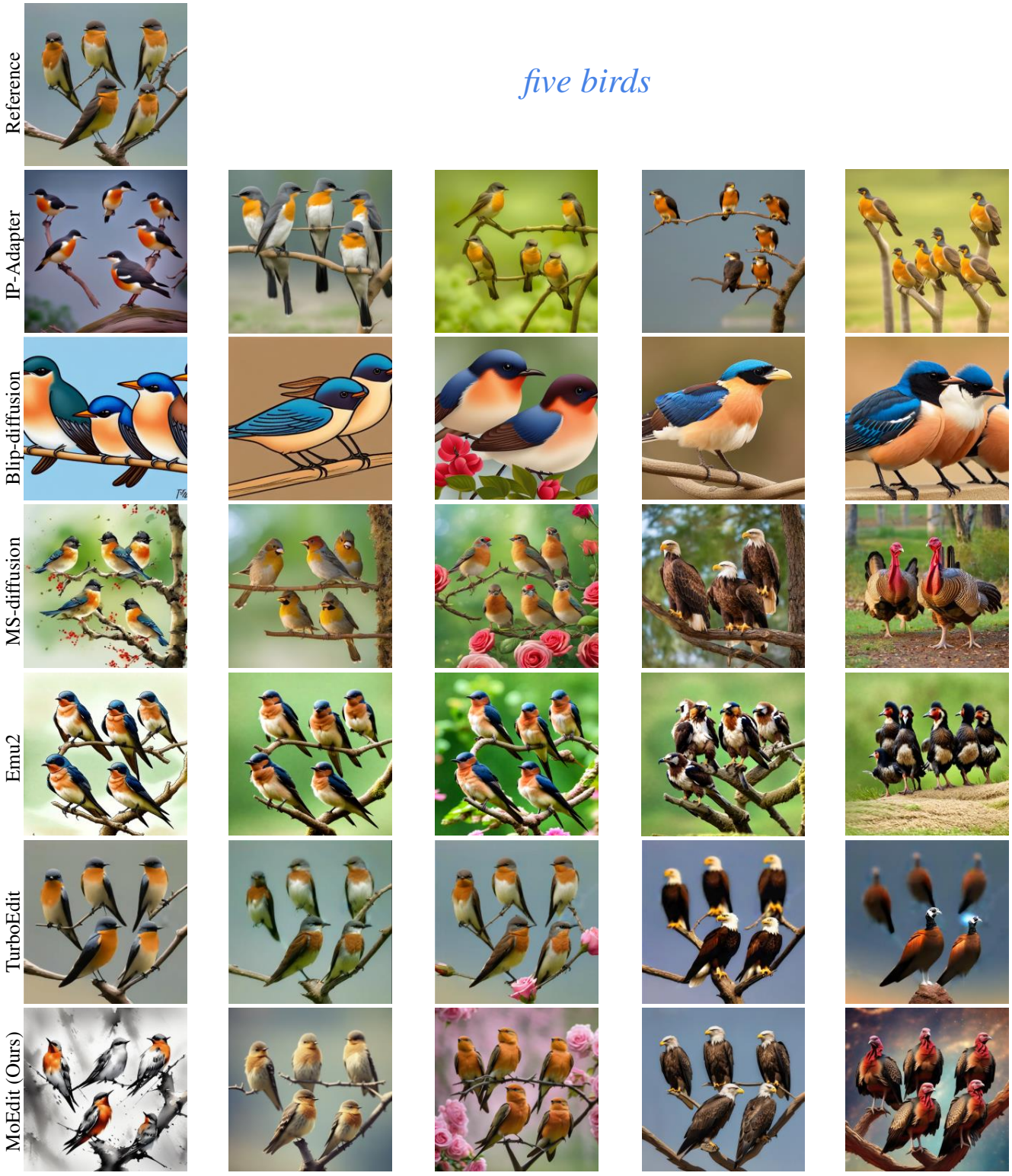


Figure 14. Qualitative comparisons among a set of five objects.


Figure 15. Qualitative comparisons among a set of six objects.


Figure 16. Qualitative comparisons among a set of eight objects.

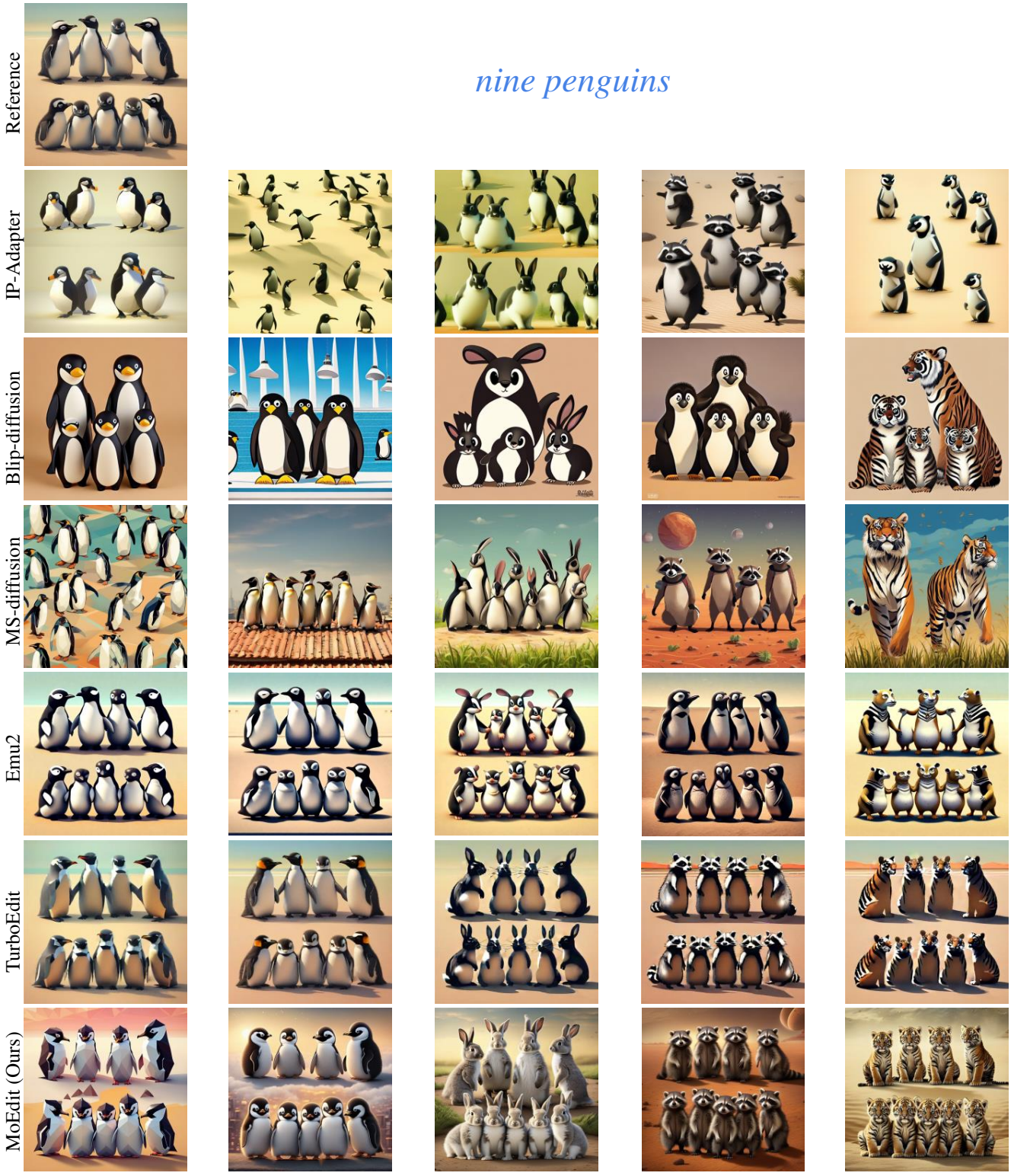


Figure 17. Qualitative comparisons among a set of nine objects.

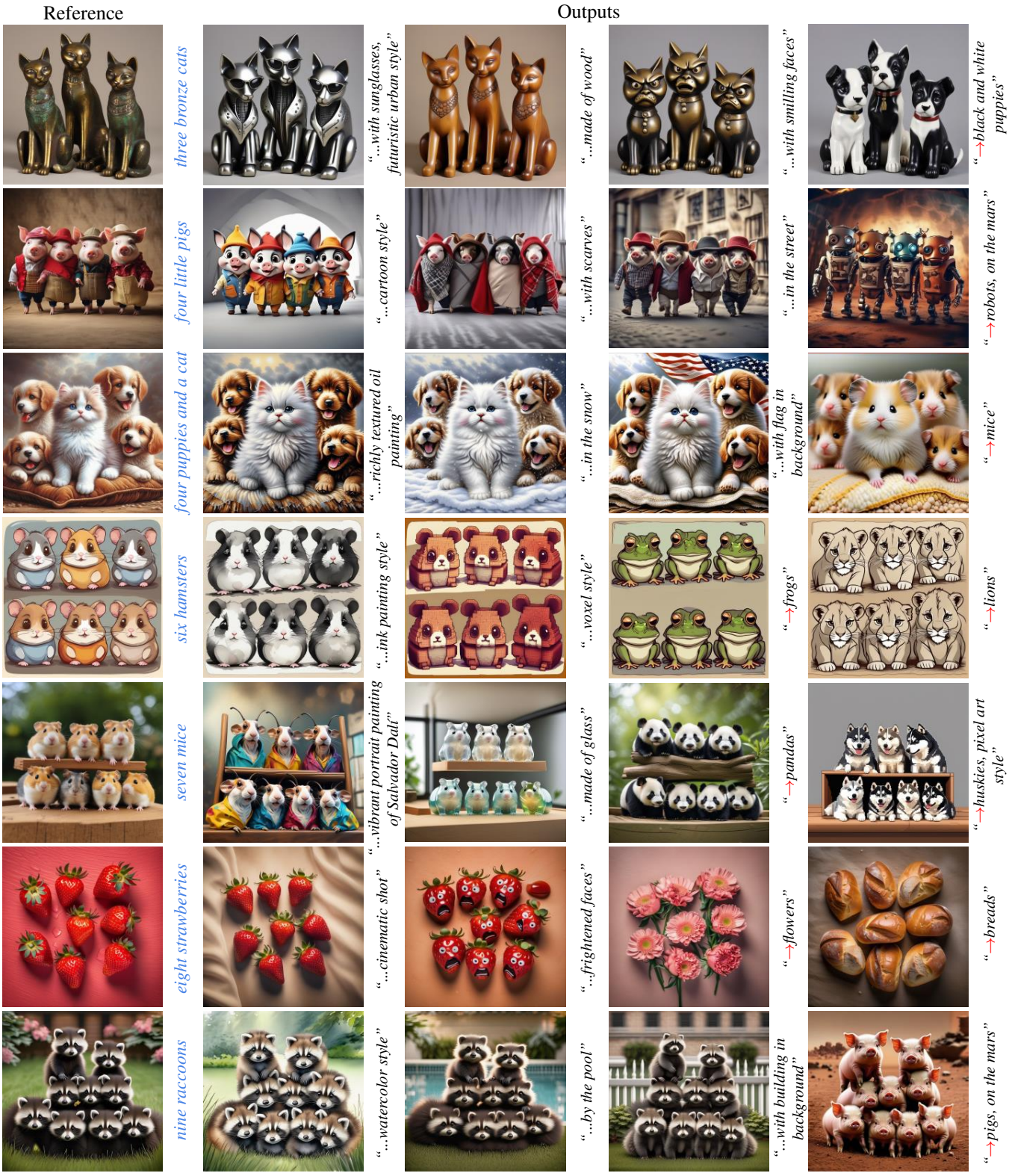


Figure 18. Additional qualitative results.


Figure 19. Scale variation. The reference image of all results are refered to Fig. 10. The top-left image illustrates the outcome when ($\lambda = 0, \beta = 0$). Each row corresponds to results obtained with a fixed $\beta$ and varying $\lambda$. From left to right, $\lambda$ increases incrementally, as indicated by $\uparrow x$, where $x$ denotes the increment from the baseline $\lambda = 0$. Similarly, each column corresponds to results generated with a fixed $\lambda$ and varying $\beta$. From top to bottom, $\beta$ increases incrementally, marked by $\uparrow x$, where $x$ represents the increment from the baseline $\beta = 0$. Detailed discussions of these results are provided in Sec. B.2.


Figure 20. The illustration of limitations. The reference refers to the input image. All output images are edited based on the text prompts *"on the table"*. For more discussion, please refer to Sec. B.3.