PositionIC: Unified Position and Identity Consistency for Image Customization

Supplementary Material

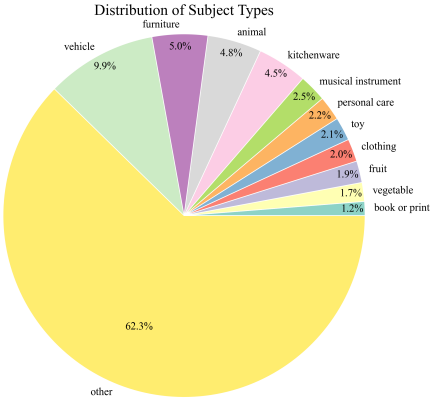

Figure 10. Category distribution of PIC-98K.

In Section 6, we give a detailed description of our data synthesis and filtering pipeline: Bidirectional Multi-dimensional Perception Data Synthesis (BMPDS). We further present a selection of example images. Section 7 introduces the PositionIC-Bench we use for position evaluation.

Section 8,9 and 10 provide a detailed discussion of specific experimental setups and additional ablation studies. Furthermore, more example demonstrations can be found in Figure 12.

6. Bidirectional Multi-dimensional Perception Data Synthesis

6.1. Detailed Instruction of GPT-4o

As shown in Figure 16, the message given to GPT-4o consists of **Instruction**, **Evaluation Metric** and **Response**. In the part of **Instruction**, we have defined the input format and evaluation metric dimensions, and required GPT-4o to select no fewer than three features for scoring based on the content of textual description. In the next part of **Evaluation Metric**, we detail the metric standard and provide GPT-4o examples to evaluate. There are 6 levels, ranging from 0 to 5, representing the similarity between two descriptions regarding the same feature. After that, we prompt GPT-4o to return a dictionary in JSON format containing the subject types and the scores for each feature. If there is no similarity between the two descriptions, the subject is set to "none", indicating that the final score is 0.

Due to the substantial differences in textual descriptions between subjects, it is not feasible to predetermine the feature categories for evaluation. Therefore, we allow the LLMs to select at least three features and assign individual scores to each. The final score is calculated as the average of all feature scores. For descriptions with significant discrepancies, the LLMs are permitted to assign a score of zero to the samples.

Figure 18 demonstrates the samples of Multi-dimensional Perception Data Filter. We have highlighted the correlated features in the description. In the first sample, the teddy bear share the same physical characteristics except for their posture, hence earning the highest appearance score and slightly lower posture score. In the second sample, the deer is missing antlers, which resulted in the lowest score on the "antler" feature.

6.2. Details of PIC-98K Dataset

We propose PIC-400K utilizing our Bidirectional Multi-dimensional Perception Data Synthesis, an automatic and effective high-consistency data synthesis pipeline. Samples of the filtered data PIC-98K are shown in the Figure 17. BMPDS can synthesize high-fidelity multi-subject images while maintaining high resolution. Against previous works, the position of subjects is controllable and it is randomly placed to train the position control capability of PositionIC.

There are over 9000 subject descriptions in PIC-98K, including multiple categories such as fruits, animals, and transportation vehicles, which basically cover common objects. The distribution of subjects is shown in Figure 10, vehicles, furniture, animal, and kitchenware constitute a significant proportion, with most difficult-to-classify subjects categorized as "other".

To ensure the dataset quality, **1)** we carefully design BMPDS as a multi-dimensional filter score by averaging four complementary metrics (CLIP, DINOv2, VLM and LLM). In particular, LLM evaluates descriptions extracted by an expert model. This multi-modal synergy mitigates biases inherent in single model. **2)** We conducted a human audit (correlation between LLM and human evaluators in Figure 7(b) reaches 89%). The audit yields FPR=9.37% and FNR=18.48%. 80% of samples have unanimous human agreement, on which the filter reaches 95% accuracy with FPR=4.35% and FNR=5.88%. The failures mainly come from inherently ambiguous pairs that are difficult even for humans (e.g., whether a hat-wearing dog matches a hatless reference), indicating a subjective gray area.

7. PositionIC-Bench

We manually select 252 single-subject samples and 296 multi-subject samples in the benchmark, where the object bounding boxes conform to standard proportions and in-

Table 4. VIEScore result.

Method	Single			Multi		
	$SS\uparrow$	$QS\uparrow$	$OS\uparrow$	$SS\uparrow$	$QS\uparrow$	$OS\uparrow$
DreamBooth	7.188	6.436	6.615		-	
ELITE	6.677	6.553	6.564		-	
RealCustom	7.812	7.540	7.634		-	
SSR-Encoder	7.028	5.336	5.988		-	
OminiControl	7.549	6.728	6.966		-	
MIP-Adapter		-		7.419	6.807	7.046
BLIP-Diffusion	6.927	5.439	6.062	5.373	6.185	5.628
MS-Diffusion	7.589	6.521	6.957	6.979	6.367	6.573
OminiGen	7.785	6.880	7.236	7.364	7.008	7.116
DreamO	8.159	8.099	8.101	<u>7.759</u>	<u>7.664</u>	<u>7.651</u>
UNO	<u>8.316</u>	<u>8.235</u>	<u>8.254</u>	7.572	7.128	7.281
Ours	**8.373**	**8.471**	**8.404**	**7.837**	**8.185**	**7.993**

Figure 11. Showcases of PositionIC-Bench.

clude challenging positional relationships. We show some samples of PositionIC-Bench in Figure 11. Our bench includes various subjects such as furniture, animals, plants, and portraits. Not limited to conventional object placement, PositionIC-Bench's bounding boxes have more complex spatial relationships where objects are placed on different planes. At the same time, the bounding boxes of smaller objects is appropriately enlarged to obtain more accurate evaluation scores.

8. Training Details and Result

8.1. Setting of Training

The value of the semantic density σ_i can be manually set or determined by Eq. (8). In Eq. (8), we set the semantic density to a geometric progression with a common ratio of $1 + 10\lambda$, which represents that σ_i increases incrementally from far to near to reflect the difference between foreground and background objects.

$$\sigma_i = \frac{10\lambda * (1 + 10\lambda)^i}{(1 + 10\lambda)^n - 1}, \tag{8}$$

where n is the number of reference images and $i = \{0, 1, 2, ..., n-1\}$ represents the object index ordered from far to near. λ is a hyperparameter used to control the variation in semantic density. A smaller value of λ results in less difference in semantic density between objects. During training, we set it to 5.

8.2. More Result

As shown in Figure 12, we downloaded several foreground images from the internet as additional visual demonstrations, covering four categories: Characters, Items, Animals, and Space. In the Characters and Items sections, PositionIC exhibits high fidelity and is capable of generating images with coherent spatial relationships. In the Animals section, we demonstrate the plausibility of object interactions under different prompts. In the Space section, PositionIC is able to achieve good results even for difficult examples (e.g., generating a planetary ring surrounding the Earth).

To evaluate the performance of PositionIC in scenarios involving multiple subjects and more complex semantic interactions, we selected several images from the MS-COCO[21] 2017 training set as reference subjects. Figure 12 (a) showcases the generation results with five reference images, where PositionIC consistently maintains the identity of each object while strictly adhering to the provided bounding boxes. Furthermore, Figure 12 (b) illustrates a scene where two individuals are shaking hands,

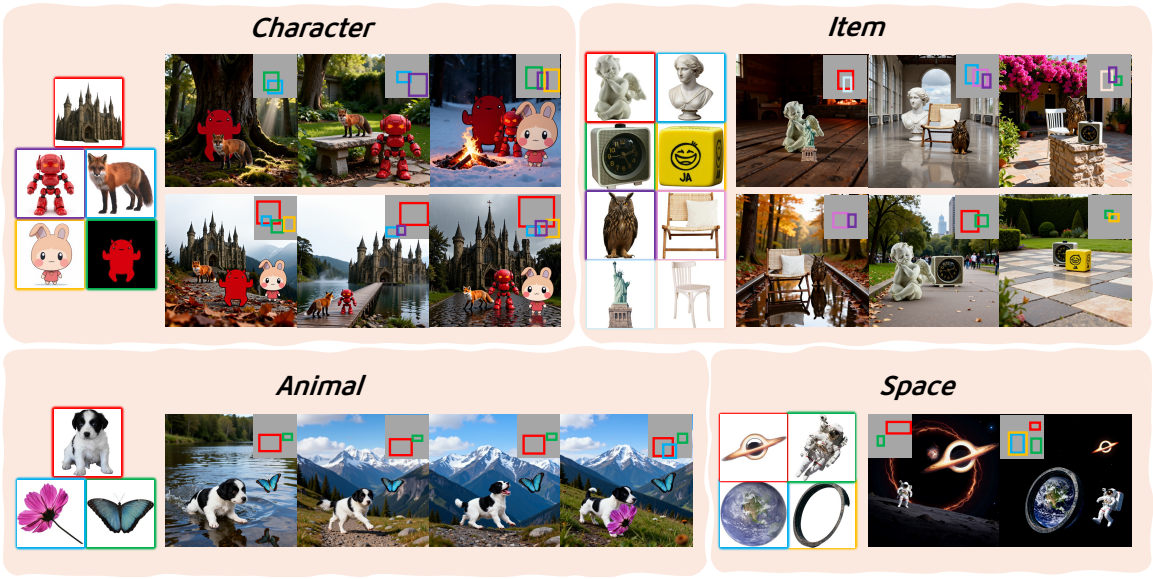

Figure 12. More results about PositionIC.

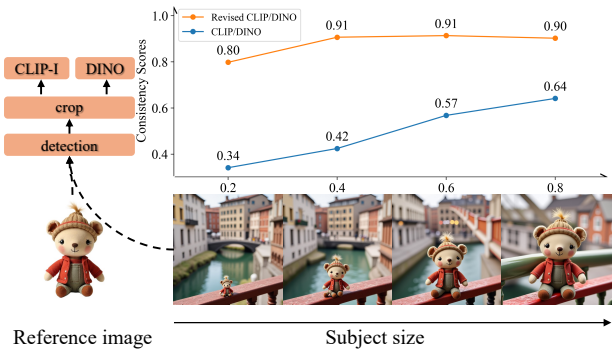

Figure 13. Inaccuracy of directly using CLIP-I and DINO. The revised score is less affected by the size of the subject which can reflect the subject consistency more authentically.

Figure 14. Qualitative results of ablation study. Visibility-Aware Attention(VAA) and our filter pipeline are capable of effective position control and consistent customization.

demonstrating PositionIC's capability to handle complex inter-object interactions.

8.3. VLM-based evaluation metrics

To provide a more comprehensive assessment, we utilized VLM-based metrics, specifically evaluating all models via VIEScore[14] across three dimensions: Semantics Score (SS), Quality Score (QS), and Overall Score (OS). GPT-4o was employed as the underlying VLM backbone for this evaluation. As illustrated in Table 4, PositionIC achieves the superior performance across all evaluated metrics.

9. Revised Evaluation Metric

In this section, we elaborate on the calculation methods for our evaluation metrics. As shown in Figure 13, directly using the entire image to compute the CLIP or DINO metrics is unreasonable, as both CLIP and DINO compute the similarity of global image features. To avoid the sensitivity, we crop the subject from original image as source images for evaluation.

10. More Ablation Study

In this section, we conduct more detailed ablation studies, including ablations on the data filter and Visibility-Aware Attention(VAA).

10.1. Impact of Visibility-Aware Attention

As shown in Figure 5, CLIP-I and DINO scores significantly drop when training without VAA. We infer that incorporating a attention mask allows the model to focus more effectively on features in a smaller region rather than globally, which accelerates the convergence and improves the consistency.

Figure 15. Results from COCO2017 and failure cases.

10.2. Qualitative Results

As shown in Figure 14, we visualize the effect of VAA and data filtering. It can be seen that without VAA, the model loses the ability to control position. When using unfiltered data, although the data volume is larger, the generated quality is poorer.

Method	CLIP-I↑	CLIP-T↑	DINO↑
w/o VAA	0.784	0.269	0.686
w/ VAA	0.846	0.269	0.823

Table 5. Ablation study of VAA. Our model performs better subject fidelity on DreamBench after restricting the attention area of reference images.

11. Limitation

While PositionIC has made significant advancements in layout control and subject consistency, certain limitations persist. Since the training dataset contains a maximum of four reference images per sample, the generation success rate tends to decline when five or more subjects are provided during inference. As illustrated in Figure 15(c), the appearance of the fifth subject (pizza) is distorted. Furthermore, extreme occlusion cases can lead to generation failures; for instance, the flowers are omitted in the generated image shown in Figure 15(d). Future work will focus on exploring solutions for generating a larger number of objects and managing extreme occlusion cases.

Instruction

You will receive two paragraphs of text, which are detailed descriptions of two different images. The input is in the following format:
describe_1:detailed describe <end>.describe_2:detailed describe <end>.
There will be a common subject in these two images.
For example, both paragraphs describe a dog. The first paragraph is a dog swimming, and the second paragraph is a dog running.
The description given to you will include a description of the subject. You need to find the common subject in the two images based on these descriptions and determine whether the two subjects are the same. Note that you need to distinguish the same at the instance level. For example, the first dog is a normal Shiba Inu, and the second dog is a Shiba Inu with different patterns. Then the two subjects are not the same. You will score the similarity of the subject from the following dimensions:
1. The similarity of key features. Such as the dog's body shape, body proportions, species, etc.
2. Distinguish between permanent features and temporary features. For example, patterns and colors are permanent features, while wearing a hat and being dirty are temporary features. Permanent features are more reliable than temporary features.
You need to decide on at least 3 features to score, and using as many feature dimensions as possible to judge.

Evaluation criteria

The scoring criteria are:
0 points: completely different objects, such as a dog and a car
1 point: completely different, but similar, such as a dog and a cat
2 points: the same object, but not guaranteed to be the same instance, such as two dogs
3 points: the same object, and the same type, such as two corgis
4 points: almost identical objects, such as two dogs with the same pattern
5 points: completely identical objects, with almost the same text description

Response

Note that you need to judge the credibility of the feature for identifying the subject. The higher the credibility, the greater the weight of its similarity. Finally, you need to output your score in the form of a python dictionary, in the following format:
{{
"subject":"", "<feature1>":5, "<feature2>":3, ...
}}
You need to fill in the value corresponding to the subject with the name of the subject you identified, such as dog. f you think there is no common subject in these two text descriptions, fill in "none".\n
At the same time, replace <feature1>, <feature2>, etc. with the feature dimensions you decided.

Figure 16. Prompt template of MLLMs in Multi-dimensional Perception Data Filter.

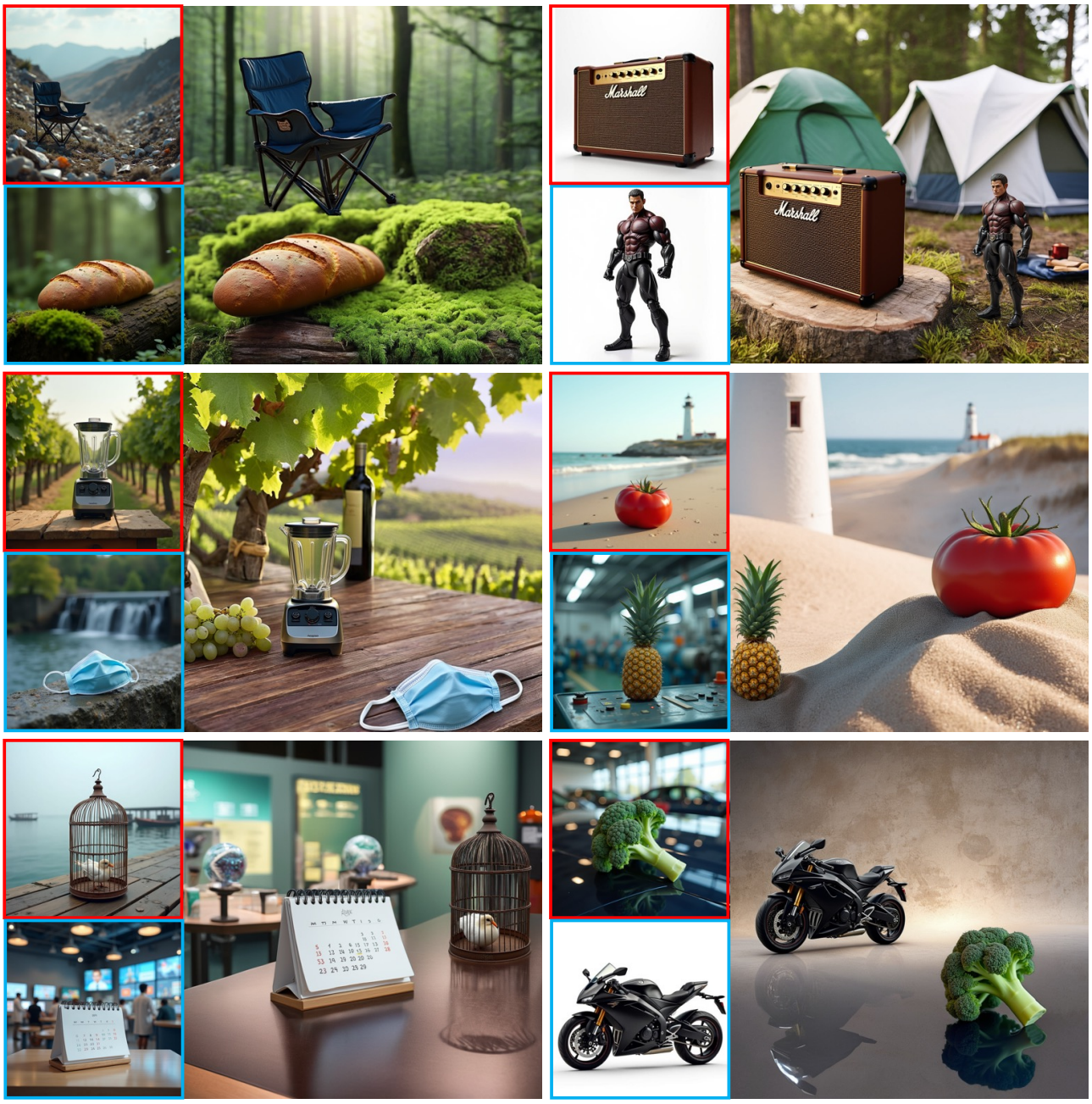

Figure 17. Showcases of PIC-98K Dataset.

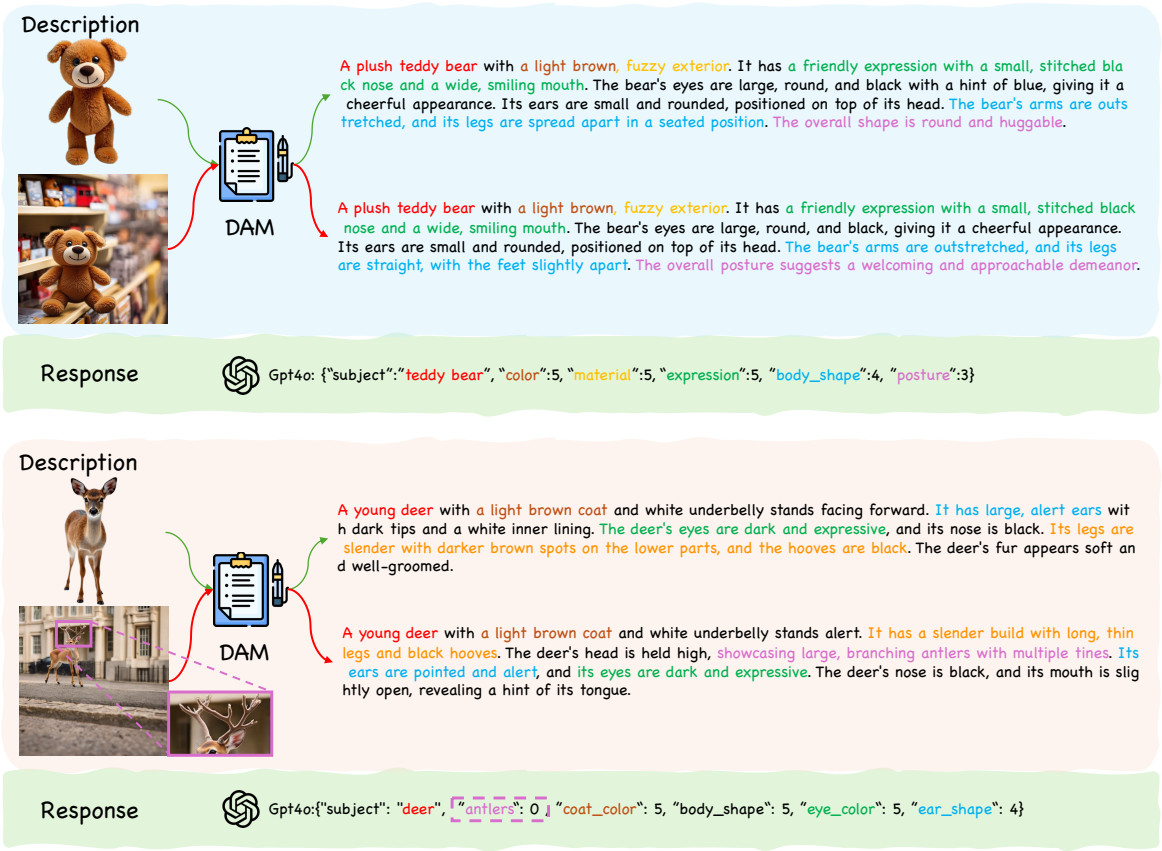

Figure 18. Examples of Multi-dimensional Perception Data Filter.