DEIG: Detail-Enhanced Instance Generation with Fine-Grained Semantic Control

Shiyan Du[1], **Conghan Yue**[2], **Xinyu Cheng**[3], **Dongyu Zhang**[1,*]

[1]Sun Yat-sen University, Guangzhou, China
[2]Fudan University, Shanghai, China
[3]Yale University, New Haven, United States
dushy5@mail2.sysu.edu.cn, chyue25@m.fudan.edu.cn, xinyu.cheng@yale.edu, zhangdy27@mail.sysu.edu.cn

Abstract

Multi-Instance Generation has advanced significantly in spatial placement and attribute binding. However, existing approaches still face challenges in fine-grained semantic understanding, particularly when dealing with complex textual descriptions. To overcome these limitations, we propose DEIG, a novel framework for fine-grained and controllable multi-instance generation. DEIG integrates an Instance Detail Extractor (IDE) that transforms text encoder embeddings into compact, instance-aware representations, and a Detail Fusion Module (DFM) that applies instance-based masked attention to prevent attribute leakage across instances. These components enable DEIG to generate visually coherent multi-instance scenes that precisely match rich, localized textual descriptions. To support fine-grained supervision, we construct a high-quality dataset with detailed, compositional instance captions generated by VLMs. We also introduce DEIG-Bench, a new benchmark with region-level annotations and multi-attribute prompts for both humans and objects. Experiments demonstrate that DEIG consistently outperforms existing approaches across multiple benchmarks in spatial consistency, semantic accuracy, and compositional generalization. Moreover, DEIG functions as a plug-and-play module, making it easily integrable into standard diffusion-based pipelines.

Introduction

Multi-Instance Generation (Li et al. 2023; Zheng et al. 2023; Zhou et al. 2024; Gu et al. 2025) has emerged as a promising direction in controllable image generation, where the goal is to generate images containing multiple semantically distinct instances at user-specified spatial locations. Recent approaches often build upon diffusion models, integrating spatial priors such as bounding boxes (Zhou et al. 2024), masks (Kim et al. 2023; Bar-Tal et al. 2023) or scribbles (Wang et al. 2024a) to improve spatial alignment and identity consistency. These methods enable finer control over instance placement and visual composition, making them suitable for downstream applications such as fashion synthesis and artistic creation.

Nevertheless, current methods remain limited in their ability to handle rich and fine-grained region descriptions.

Figure 1: **Fine-Grained Generation.**Given bounding boxes and detailed descriptions, our method accurately generate multi-attribute instances, while existing methods fail to preserve fine-grained semantic details.

As illustrated in Fig. 1, although these methods can reliably generate instances with simple prompts, they often fail with complex, multi-attribute inputs involving multiple attributes such as multi-color designs and combinations of color, texture, and material. This restricts their applicability in scenarios requiring high-detail synthesis.

We attribute these limitations to two primary factors. First, current approaches predominantly focus on preventing semantic leakage, while neglecting the deeper semantic comprehension required for generating fine-grained visual details. (Zhou et al. 2024; Wang et al. 2024a; Gu et al. 2025) Second, the training data used in these methods are typically annotated with coarse-grained templates, lacking detailed instance-level descriptions. This restricts the model's ability to learn rich semantic-visual mappings from the data, thereby impeding the generation of visually and semantically coherent content.

In this work, we propose DEIG, a framework for generating Multi-Instance images with fine-grained attribute control. Inspired by recent advances in long-text alignment for text-to-image generation (Hu et al. 2024; Han et al. 2024), we extend global prompt-based generation toward instance-level detail generation. Specifically, we in-

*Corresponding author.

troduce the *Instance Detail Extractor* (IDE), which transforms high-dimensional embeddings from LLMs encoder into compact, instance-aware representations, enabling localized alignment between complex textual descriptions and visual regions. Leveraging a high-quality captioned dataset and a *Detail Fusion Module* (DFM), DEIG enables precise and attribute-consistent multi-instance generation.

To evaluate instance-level controllability under fine-grained prompts, we introduce DEIG-Bench, a challenging benchmark specifically designed to address key limitations of existing datasets—namely, the underrepresentation of human instances and the reliance on single-attribute prompts. DEIG-Bench provides multi-attribute, compositional descriptions along with tailored evaluation protocols for both human and object instances. For human-centric scenes, we focus on color compositionality across wearable regions; for object-centric scenes, we gradually increase attribute complexity, encompassing color, material, and texture. To enable robust and reliable assessment, we leverage two distinct VLMs in a question-answering setup to evaluate complex semantic consistency. To summarize, our main contributions are as follows:

- We propose DEIG, a novel framework that enhances instance-level detail representation and semantic understanding, to address the limitations of existing methods in handling rich, fine-grained region descriptions.
- We introduce DEIG-Bench, a comprehensive evaluation suite specifically designed for multi-attribute, multi-Instance generation, to fill the gap of lacking benchmarks for evaluating fine-grained semantic prompts.
- We conduct extensive experiments on multiple widely used benchmarks to demonstrate that DEIG significantly outperforms previous methods, particularly in generating detailed multi-attribute instances with improved semantic fidelity.

Related Work

Controllable Diffusion models

While text-to-image diffusion models generate high-quality outputs (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Podell et al. 2023; Betker et al. 2023), recent work focuses on enhancing controllability by introducing various conditioning mechanisms. Subject-driven tasks (Gal et al. 2022; Ruiz et al. 2023; Ma et al. 2024; Wang et al. 2024b; Chen et al. 2024), aim to preserve subject identity and enable personalized generation. Spatial control tasks (Zhang, Rao, and Agrawala 2023; Li et al. 2023; Mou et al. 2024; Mo et al. 2024) specify the spatial arrangement of content using layout or structural information. Text-conditioned control targets fine-grained alignment with complex or compositional prompts (Feng et al. 2022; Rassin et al. 2023; Hu et al. 2024; Han et al. 2024; Liu et al. 2025; Wu et al. 2025). Collectively, these tasks enhance the controllability of diffusion models and enable their application in different domains.

Multi-Instance Generation

With the widespread adoption of diffusion models, numerous studies have explored their potential for Multi-Instance Generation, focusing on synthesizing images with multiple semantically distinct instances arranged according to given layouts. Existing methods are typically categorized into training-free and training-based methods

Training-Free Methods Training-free methods (Kim et al. 2023; Bar-Tal et al. 2023) primarily operate during inference by manipulating attention mechanisms (Chen, Laina, and Vedaldi 2024; Xiao et al. 2024; Phung, Ge, and Huang 2024) or modifying loss functions (Xie et al. 2023; Wang et al. 2025). These methods often optimize the latent space to align generated instances with target spatial configurations. Although they offer considerable flexibility and require no retraining, they frequently compromise image fidelity, as the generation may deviate from the data distribution originally learned by the pretrained diffusion model.

Training-Based Methods In contrast, training-based methods incorporate instance-level cues, such as embeddings (Li et al. 2023; Zhou et al. 2024; Wang et al. 2024a; Gu et al. 2025) or semantic layouts (Jia et al. 2024; Wu et al. 2024), directly into the training pipeline. To mitigate attribute interference across instances, techniques like divide-and-conquer (Zhou et al. 2024) and hierarchical generation (Cheng et al. 2024) are employed. Recent DiT-based methods (Zhang et al. 2024, 2025; Zhou et al. 2025a) achieve notable improvements in scalability and image quality. However, their high computational overhead makes them impractical for deployment on consumer-grade GPUs.

Despite their progress, existing methods struggle with compositional prompts involving fine-grained attributes, and remain limited in human-centric scenarios. This hinders their applicability in domains requiring precise semantic alignment and spatial control.

Method

Overview

In this task, users provide instance-level conditions, including bounding boxes, fine-grained descriptions, and a global context $\mathcal{P}$. Formally, the input is $\mathcal{C} = \{\mathcal{P}, (b_1, p_1), \ldots, (b_n, p_n)\}$, where each (b_i, p_i) denotes the location and semantics of the i-th instance. The objective is to generate an image that aligns with both spatial and semantic constraints while preserving global coherence. To this end, we propose DEIG, a pipeline for multi-instance generation with fine-grained semantic control over the attributes of each instance, as illustrated in Fig. 2.

Instance-Level Semantic Enhancement

As discussed above, existing methods often neglect instance-level semantic understanding, limiting their ability to generate fine-grained content. To address this, we introduce a dedicated module that effectively extracts and represents instance-specific semantics.

Instance Detail Extractor Inspired by recent advances in long-text alignment with LLMs for text-to-image generation (Hu et al. 2024; Han et al. 2024; Liu et al. 2025; Wu et al. 2025), we replace traditional multi-modal encoders

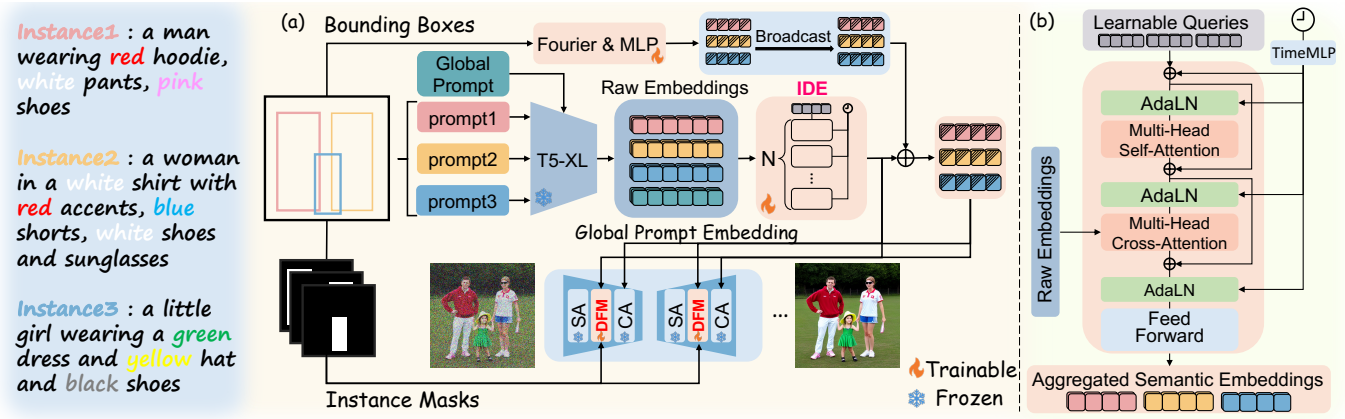

Figure 2: **Overview of the DEIG pipeline.** (a) DEIG enables the use of a frozen large text encoder to extract raw instance embeddings, which are refined by the IDE and fused into the UNet via the DFM for fine-grained control. (b) Structure of the IDE, which refines learnable queries via time-aware self- and cross-attention to produce compact instance embeddings.

with a frozen text encoder to better capture fine-grained instance semantics. As the resulting embeddings are high-dimensional and costly to use directly, we introduce the *Instance Detail Extractor* (IDE)—a lightweight module that distills rich text features into compact, instance-aware representations via learnable queries. While query-based distillation has proven effective in prior work (Li, Li, and Hoi 2023; Hu et al. 2024; Han et al. 2024; He et al. 2024; Zhou et al. 2025b), our IDE is specifically designed for multi-instance generation, supporting token-level alignment and temporal conditioning compatible with diffusion models.

Structurally, IDE integrates stacked self-attention and cross-attention layers to refine instance-specific semantics, as illustrated in Fig. 2(b). It operates on encoded text features $\mathbf{E}_\tau \in \mathbb{R}^{B\times N\times S_\tau\times C}$, where S_τ is the embedding sequence length. To avoid direct use of these high-dimensional features, IDE introduces learnable queries $\mathbf{Q} \in \mathbb{R}^{B\times N\times S\times C}$, with $S \ll S_\tau$. We refer to S as the **Aggregated Semantic Dimension**, which acts as a bottleneck to compress and organize instance-specific information efficiently.

Each IDE layer refines the queries via temporally-aware attention. The process begins with timestep conditioning through a lightweight TimeMLP, followed by adaptive layer normalization (AdaLN) (Perez et al. 2018), which adaptively modulates the query features using the same temporal embedding. A self-attention block captures intra-instance dependencies, and a subsequent cross-attention module aligns the queries with the high-dimensional text features from the frozen encoder. For the i-th layer, the key transformation can be expressed by:

$$\mathbf{H}^{\mathbf{i}}_{\mathbf{ca}} = \mathrm{CrossAttn}\left(\mathrm{AdaLN}\left(\mathbf{H}^{\mathbf{i}}_{\mathbf{sa}}, \mathbf{T}_{\mathbf{emb}}\right), \left[\mathbf{H}^{\mathbf{i}}_{\mathbf{sa}}, \mathbf{E}_\tau\right]\right), \quad (1)$$

where $\mathbf{T}_{\mathrm{emb}}$ is the timestep embedding vector that modulates the visual features via AdaLN. $\mathbf{H}^{\mathbf{i}}_{\mathbf{sa}} \in \mathbb{R}^{B\times N\times S\times C}$ denotes the self-attended instance queries at the i-th layer, the output $\mathbf{H}^{\mathbf{i}}_{\mathbf{ca}} \in \mathbb{R}^{B\times N\times S\times C}$ captures cross-modal interactions between textual and visual embeddings, thereby completing the feature transformation at layer i.

The output is subsequently passed through a residual feed-forward network, completing one refinement cycle. After stacking N such layers, IDE produces a set of compact and expressive instance-level embeddings, which we refer to as **Aggregated Semantic Embeddings**. We extract them from the second layer of the encoder in the UNet model at an intermediate diffusion timestep and visualize them across different semantic dimensions S. As illustrated in Fig. 3(b), each dimension attends to specific fine-grained attributes described in the input. Collectively, these dimensions constitute a comprehensive representation of each instance, confirming that our approach achieves more precise semantic alignment than previous methods.

Detail Fusion Module

To integrate the Aggregated Semantic Embeddings into generation, we propose the *Detail Fusion Module* (DFM), where Grounding Embeddings Broadcast aligns spatial cues with the semantic dimension, and Instance-based Masked Attention applies masking strategies to avoid attribute leakage.

Grounding Embeddings Broadcast To align aggregated semantics with spatial generation, we adopt a broadcast-based spatial fusion approach across all S dimensions. Each instance's spatial coordinate b_i is transformed via Fourier encoding and broadcast as:

$$\mathbf{f_i} = \mathcal{B}(\mathcal{F}(b_i), S), \quad \mathbf{e}_i = \mathcal{B}(\mathbf{e}_i, S) \in \mathbb{R}^{(B,N,S,C)} \quad (2)$$

$$\mathbf{G}_{\mathrm{ase},i} = \mathrm{MLP}([m\cdot\mathbf{f_i} + (1-m)\cdot\mathbf{e}_i, \mathbf{E}_{\mathrm{ase},i}]) \quad (3)$$

For i-th instance, $\mathbf{f}_i$ denotes the broadcasted spatial embedding, and $\mathbf{e}_i$ is a learnable null embedding used when spatial information is absent. $\mathcal{F}(b_i)$ computes the Fourier encoding of coordinates, and $\mathcal{B}(\cdot, S)$ denotes broadcasting along dimension S. The binary mask $m \in \{0, 1\}$ selects between the two. The output $\mathbf{G}_{\mathrm{ase},i}$ is a fused embedding, combining spatial and semantic cues, and is later used in instance-based masked attention.

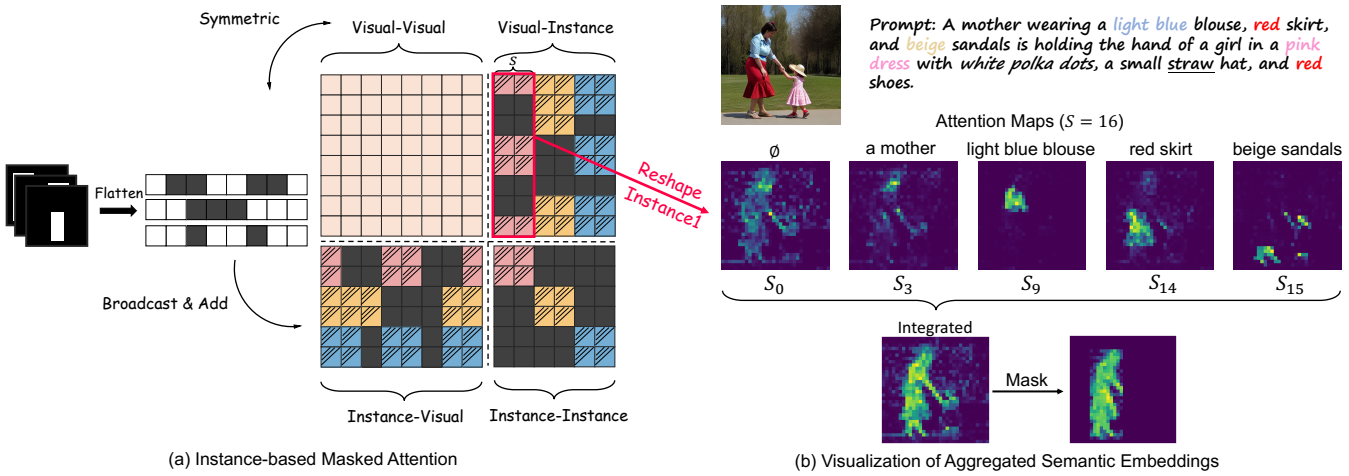

Figure 3: **Workflow Visualization of Fine-Grained Instance Generation.** (a) Instance-based masked attention mechanism divides the attention map into four sub-regions, applying masks to restrict cross-instance interactions and prevent semantic leakage. (b) Visualization of aggregated semantic embeddings across different semantic dimensions.

Instance-based Masked Attention Following prior embedding-guided fusion frameworks, we freeze the self- and cross-attention layers of UNet and insert a gated self-attention module between them to enable instance level semantic fusion. To mitigate attribute leakage across instances, which often occurs with standard self-attention, we introduce a masking mechanism based on instance partitioning.

As shown in Fig. 3(a), the gated self-attention operates on the concatenation of visual and instance embeddings, generating an attention map naturally divided into four interpretable subregions. To regulate these interactions, we define a binary mask $\mathbf{M} \in \{0, -\infty\}^{L \times L}$, where $L = N_{\text{visual}} + N \times S$. For any $v \in \mathbf{V}_{\text{visual}}$, $g \in \mathbf{G}_{\text{ase}}$, the masking rules are as follows:

(a) Visual-Visual Attention. We observe that masking visual embeddings can noticeably degrade image fidelity. Therefore, we allow all visual embeddings to attend to each other without masking:

$$\mathbf{M}_{v_i, v_j} = 0 \quad , \forall i, j \in 1, \ldots, N_{\text{visual}} \tag{4}$$

(b) Symmetric Instance–Visual Attention. Each instance embedding is allowed to attend only to visual embeddings from the same instance, and vice versa. For any two different instances, interactions between them are masked by setting the corresponding attention scores a value of negative infinity. This operation can be formulated as:

$$\mathbf{M}_{v_i, g_j} = \mathbf{M}_{g_i, v_j} = \begin{cases} 0, & \text{if Instance}(v_i) = \text{Instance}(g_j) \\ -\infty, & \text{otherwise} \end{cases} \tag{5}$$

(c) Instance-Instance Attention. Instance embeddings attend only to others within the same semantic group and all cross-group interactions are masked with negative infinity, following the same rule as above.

$$\mathbf{M}_{g_i, g_j} = \begin{cases} 0, & \text{if Group}(g_i) = \text{Group}(g_j) \\ -\infty, & \text{otherwise} \end{cases} \tag{6}$$

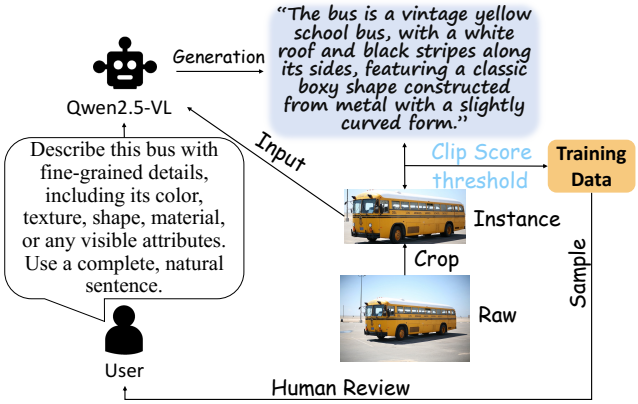

Figure 4: **Detail-Enriched caption generation pipeline.** Instances are described by a VLM from cropped images and filtered using CLIP scores and human review.

The final masked attention output is computed via a standard self-attention:

$$\hat{\mathbf{A}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V} \tag{7}$$

As shown in Fig. 3(b), the mask effectively suppresses attention that would otherwise leak into other instances or the background. Finally, the visual embeddings are then updated through gated residual mechanism:

$$\mathbf{V}_{visual} = \mathbf{V}_{visual} + \eta \cdot \tanh\gamma \cdot \mathcal{ES}(\hat{\mathbf{A}}), \tag{8}$$

where $\mathcal{ES}(\cdot)$ denotes *embedding slices* operation, which extract a subset of the visual embedding outputs. η and γ are learnable scalars that control the update strength.

Detail-Enriched Instance Captions

High-quality datasets are vital for accurate instance-level generation. We curate a dataset from MS-COCO (Lin et al.

Method	Qwen2.5-VL									InternVL3									mIoU ↑
	MAA_{human} ↑				MAA_{obj} ↑					MAA_{human} ↑				MAA_{obj} ↑					
	C1	C2	C3	AVG	L1	L2	L3	L4	AVG	C1	C2	C3	AVG	L1	L2	L3	L4	AVG	
GLIGEN	0.23	0.05	0.02	0.10	0.20	0.08	0.08	0.03	0.10	0.28	0.12	0.05	0.15	0.32	0.17	0.17	0.11	0.19	0.71
MIGC	0.51	0.11	0.03	0.22	0.63	0.29	0.29	0.20	0.36	0.60	0.23	0.11	0.31	0.75	0.54	0.47	0.36	0.54	0.72
InstanceDiffusion	0.54	0.17	0.05	0.25	0.53	0.24	0.32	0.20	0.33	0.61	0.23	0.12	0.32	0.62	0.49	0.46	0.36	0.49	0.75
ROICtrl	0.61	0.23	0.09	0.31	0.56	0.32	0.27	0.16	0.33	0.68	0.33	0.16	0.39	0.67	0.48	0.42	0.30	0.47	0.71
DEIG (Ours)	**0.82**	**0.74**	**0.69**	**0.75**	**0.67**	**0.41**	**0.38**	**0.27**	**0.44**	**0.86**	**0.82**	**0.81**	**0.83**	**0.79**	**0.58**	**0.50**	**0.44**	**0.58**	**0.79**

Table 1: **Quantitative Results on DEIG-Bench.** We report *MAA* for human (C1–C3) and object (L1–L4) instances under two VLMs, Qwen2.5-VL (Bai et al. 2025) and InternVL3 (Zhu et al. 2025), as well as spatial alignment via *mIoU*.

Method	Instance Success Rate (%) ↑						mIoU ↑					
Level	L2	L3	L4	L5	L6	AVG	L2	L3	L4	L5	L6	AVG
GLIGEN	41.56	32.29	28.13	25.38	29.79	29.91	36.70	29.10	24.92	23.37	27.22	27.03
MIGC	75.00	65.83	66.88	62.63	64.79	65.84	64.29	55.94	56.68	53.63	56.25	56.44
InstanceDiffusion	68.44	58.96	59.84	55.75	56.77	58.63	61.91	54.63	53.71	50.17	51.30	53.06
ROICtrl	72.19	65.63	64.69	59.88	60.94	63.25	61.75	57.32	56.28	52.68	53.56	55.27
DEIG (Ours)	**75.23**	**72.00**	**70.02**	**71.00**	**73.13**	**72.25**	**60.43**	**66.50**	**62.01**	**61.65**	**62.68**	**62.64**

Table 2: **Quantitative results on MIG-Bench.** MIG-Bench focuses on color-centric attribute evaluation via *Instance Success Rate*, which checks color match within a predefined gamut, and assesses spatial alignment using *mIoU* with Grounding-DINO.

Method	$Acc_{c.}$	$CLIP_{c.}$	$Acc_{t.}$	$CLIP_{t.}$	AP	AP_{50}
GLIGEN	25.0	0.217	15.8	0.205	0.20	0.35
MIGC	52.3	0.252	23.2	0.221	0.22	0.40
InstanceDiffusion	53.4	0.252	25.9	0.227	**0.40**	0.57
ROICtrl	56.9	0.255	23.7	0.223	0.26	0.51
DEIG (Ours)	**58.8**	**0.258**	**26.1**	**0.228**	0.34	**0.57**

Table 3: **Quantitative results on InstDiff-Bench.** Attribute alignment is assessed using *Accuracy* and *CLIP scores* for color ($Acc_{c.}$, $CLIP_{c.}$) and texture ($Acc_{t.}$, $CLIP_{t.}$), while spatial precision is measured by *AP* and AP_{50} using YOLOv8.

2014) using Qwen2.5-VL (Bai et al. 2025) to generate detailed, context-aware captions averaging 20–30 words per instance. Unlike template-based methods, our approach produces natural and semantically rich descriptions. To ensure quality and minimize hallucinations, we remove grayscale and low-fidelity images via VLM assessment and apply a two-stage verification process. This process involves (1) computing the CLIP (Radford et al. 2021) score for each image–caption pair and keeping those above a predefined threshold, and (2) performing human verification on a random subset of 500 pairs to confirm overall consistency and reliability. The data construction pipeline is shown in Fig. 4.

Experiment

To thoroughly evaluate the effectiveness of DEIG, we conduct comprehensive experiments, including comparisons with previous state-of-the-art baselines and detailed ablation studies. We adopt the encoder of Flan-T5-XL (Chung et al. 2024) as the text encoder in all experiments. Additional experimental settings and more qualitative results are provided in the Appendix.

DEIG-Bench

Existing multi-instance generation benchmarks often lack fine-grained supervision and realistic human-centric compositional prompts, limiting evaluation of real-world appearance complexity.

To address these gaps, we introduce **DEIG-Bench**, a benchmark for evaluating fine-grained, multi-attribute generation of both human and object instances. It features compositional prompts and structured complexity levels that reflect real-world attribute entanglement. DEIG-Bench is constructed from 400 filtered images in the MS-COCO validation set, ensuring each image contains 3–10 visible instances for reliable attribute recognition.

DEIG-Bench evaluates fine-grained compositionality in both human and object instances. For humans, we define three difficulty levels: C1, C2 and C3, based on color combinations across wearable regions. For objects, we define four attribute levels from L1 to L4, where L1 includes only color, L2 adds material on top of color, L3 adds texture on top of color, and L4 combines all three for the highest semantic and visual complexity. Evaluation combines *mIoU*, computed using Grounding-DINO (Liu et al. 2024), for spatial alignment, and two VLMs for semantic validation. We also introduce *Multi-Attribute Accuracy* (MAA) to measure how well models bind multiple attributes to each instance, which

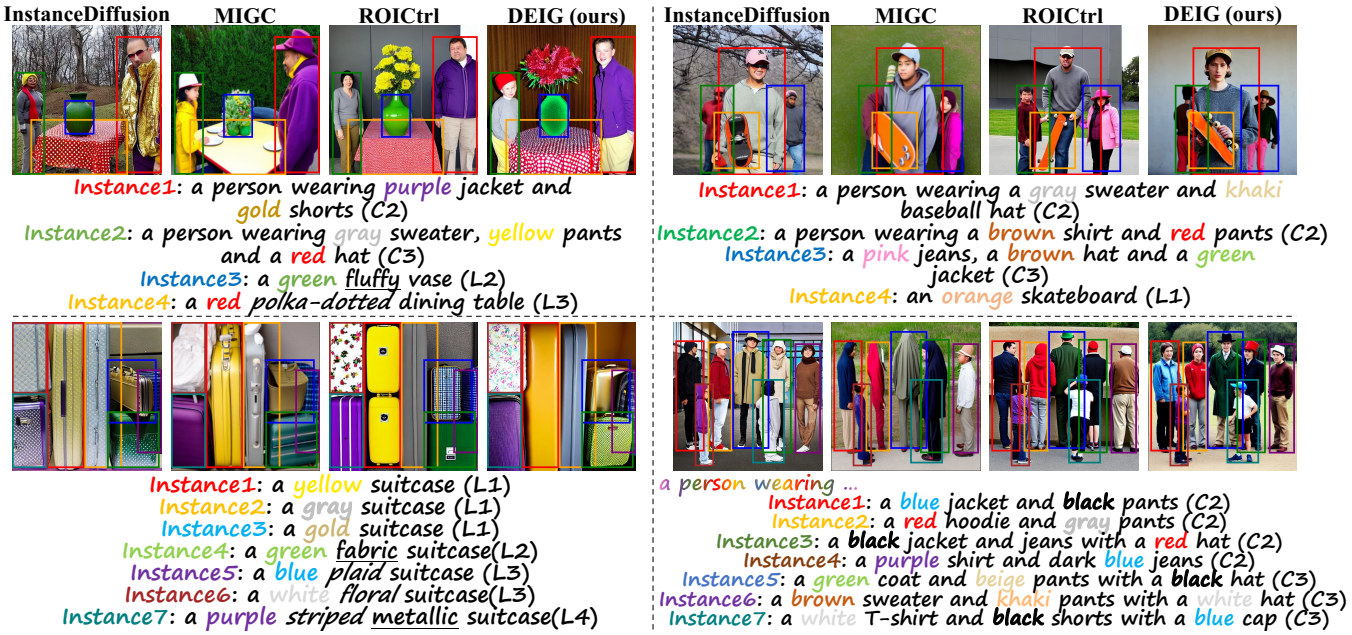

Figure 5: **Qualitative comparison on DEIG-Bench.** DEIG exhibits accurate generation of fine-grained, multi-attribute instances across varying levels of complexity, demonstrating superior compositional control and semantic alignment.

IDE	DFM	Cap.	mIoU ↑	MAA_{human} ↑	MAA_{obj} ↑
	✓	✓	0.73	0.51	0.35
✓		✓	0.75	0.70	0.41
✓	✓		0.70	0.31	0.29
✓	✓	✓	**0.79**	**0.75**	**0.44**

Table 4: **Ablation study of DEIG components.** We evaluate the individual contributions of IDE, DFM, and Captions supervision by measuring MAA with Qwen2.5-VL and mIoU.

is calculated as the ratio of instances correctly identified by the VLM to the total number of instances.

For broader validation, we additionally report results on MIG-Bench (Zhou et al. 2024) and InstDiff-Bench (Wang et al. 2024a).

Comparison

We compare our method with previous SOTA approaches for Multi-Instance Generation. Specifically, we select GLIGEN, MIGC, InstanceDiffusion, and ROICtrl as baselines for comparison.

Quantitative Results Tab. 1 presents the quantitative results on DEIG-Bench. On human-centric tasks, DEIG significantly outperforms all baselines across attribute complexities, particularly under multi-color combinations, indicating stronger compositional generalization enabled by our detail-aware semantic extraction and fusion modules. For object-centric generation, DEIG also achieves consistently higher scores, with larger gains on color attributes than on material and texture, likely because color correlates more directly

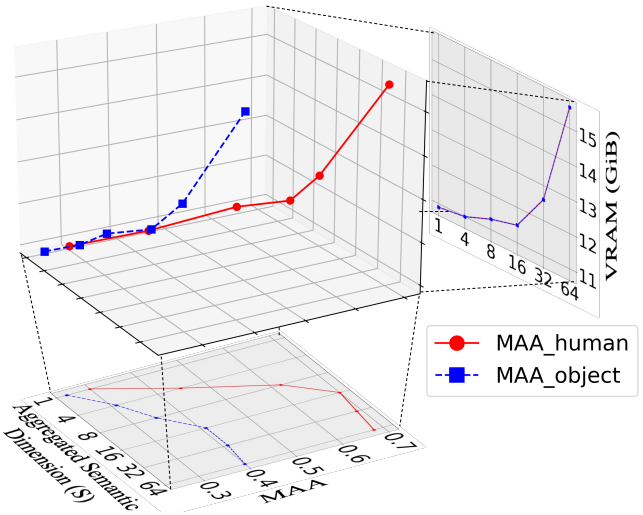

Figure 6: **Trade-off Between Semantic Precision and Computational Cost.** 3D visualization of how the aggregated semantic dimension S affects MAA and VRAM usage under FP16 precision.

with the RGB space and is easier for diffusion models to learn, whereas material and texture require more abstract semantic understanding beyond surface-level appearance.

Tab. 2 and Tab. 3 present the results on MIG-Bench and InstDiff-Bench. DEIG shows strong controllability under dense layouts and diverse attribute prompts. While spatial alignment on InstDiff-Bench is slightly lower due to instance-masked attention limiting interactions in crowded

Figure 7: **Plug-and-play adaptation to community diffusion models.** DEIG preserves fine-grained semantic control when integrated into community diffusion models, demonstrating strong compatibility and generalization.

regions, it consistently surpasses baselines in accuracy and CLIP alignment, especially for color attributes.

Qualitative Comparison DEIG effectively generates images from multi-attribute prompts with distinct spatial separation and strong attribute fidelity, as illustrated in Fig. 5. While other methods struggle to represent multiple attributes details per instance, our approach preserves semantic integrity across instances, with each accurately reflecting its described color, material, and texture combination.

Fig. 7 shows DEIG adapted to a community diffusion backbone without retraining. It preserves spatial layout and fine-grained generation quality, highlighting the plug-and-play nature and compatibility with common diffusion pipelines.

Ablation Study

We conduct an ablation study to evaluate the contributions of three components in DEIG: IDE, DFM, and detail-enriched instance captions. As shown in Tab. 4, removing the captioning module leads to the largest drop in MAA, highlighting the importance of fine-grained semantic supervision. Excluding IDE reduces semantic alignment, confirming its role in detailed instance representation, while removing DFM slightly degrades accuracy due to increased semantic leakage between instances. Overall, performance drops more for human instances than for object instances, indicating higher sensitivity to fine-grained control.

We further analyze the effect of the aggregated semantic dimension S on performance and computational cost. As shown in Fig. 6, increasing S improves MAA for both human and object instances, with gains saturating around $S = 16$. Beyond this point, performance plateaus or slightly declines, while GPU memory usage increases steadily, revealing a trade-off between accuracy and efficiency. A setting of $S = 16 \sim 32$ provides a good balance.

Conclusion

In this paper, we propose DEIG, a detail-enhanced framework for fine-grained multi-instance generation. DEIG integrates instance-aware semantic extraction and masked attention fusion to improve attribute alignment and reduce leakage. Built as a plug-and-play module, it adapts to existing diffusion pipelines with minimal overhead. Extensive experiments on multiple benchmarks show that DEIG achieves strong spatial consistency and semantic fidelity, advancing controllable generation in complex multi-instance scenarios.

Acknowledgments

This research is supported by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515011741. The authors express their sincere gratitude to all collaborators and colleagues who provided valuable insights, constructive suggestions throughout the course of this work.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *International Conference on Machine Learning*, 1737–1752. PMLR.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 5343–5353.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6593–6602.

Cheng, B.; Ma, Y.; Wu, L.; Liu, S.; Ma, A.; Wu, X.; Leng, D.; and Yin, Y. 2024. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *arXiv preprint arXiv:2410.14324*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gu, Y.; Zhou, Y.; Ye, Y.; Nie, Y.; Yu, L.; Ma, P.; Lin, K. Q.; and Shou, M. Z. 2025. Roictrl: Boosting instance control for visual generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23658–23667.

Han, Y.; Wang, R.; Zhang, C.; Hu, J.; Cheng, P.; Fu, B.; and Zhang, H. 2024. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts. *arXiv preprint arXiv:2406.09162*.

He, X.; Liu, Q.; Qian, S.; Wang, X.; Hu, T.; Cao, K.; Yan, K.; and Zhang, J. 2024. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, X.; Wang, R.; Fang, Y.; Fu, B.; Cheng, P.; and Yu, G. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Jia, C.; Luo, M.; Dang, Z.; Dai, G.; Chang, X.; Wang, M.; and Wang, J. 2024. Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2480–2488.

Kim, Y.; Lee, J.; Kim, J.-H.; Ha, J.-W.; and Zhu, J.-Y. 2023. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7701–7711.

Li, D.; Li, J.; and Hoi, S. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36: 30146–30166.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22511–22521.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, M.; Ma, Y.; Yang, Z.; Dan, J.; Yu, Y.; Zhao, Z.; Hu, Z.; Liu, B.; and Fan, C. 2025. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5523–5531.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2024. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Mo, S.; Mu, F.; Lin, K. H.; Liu, Y.; Guan, B.; Li, Y.; and Zhou, B. 2024. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7465–7475.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Phung, Q.; Ge, S.; and Huang, J.-B. 2024. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7932–7942.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rassin, R.; Hirsch, E.; Glickman, D.; Ravfogel, S.; Goldberg, Y.; and Chechik, G. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36: 3536–3559.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Wang, J.; Yan, C.; Zhang, W.; Lin, H.; Wang, M.; Dai, G.; Gong, T.; Sun, H.; and Wang, J. 2025. SpotActor: Training-Free Layout-Controlled Consistent Image Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7718–7726.

Wang, X.; Darrell, T.; Rambhatla, S. S.; Girdhar, R.; and Misra, I. 2024a. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6232–6242.

Wang, X.; Fu, S.; Huang, Q.; He, W.; and Jiang, H. 2024b. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*.

Wu, W.; Li, Z.; He, Y.; Shou, M. Z.; Shen, C.; Cheng, L.; Li, Y.; Gao, T.; and Zhang, D. 2025. Paragraph-to-image generation with information-enriched diffusion model. *International Journal of Computer Vision*, 1–22.

Wu, Y.; Zhou, X.; Ma, B.; Su, X.; Ma, K.; and Wang, X. 2024. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*.

Xiao, J.; Lv, H.; Li, L.; Wang, S.; and Huang, Q. 2024. R&B: Region and Boundary Aware Zero-shot Grounded Text-to-image Generation. In *ICLR*.

Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Zhang, H.; Duan, Z.; Wang, X.; Chen, Y.; and Zhang, Y. 2025. Eligen: Entity-level controlled image generation with regional attention. *arXiv preprint arXiv:2501.01097*.

Zhang, H.; Hong, D.; Wang, Y.; Shao, J.; Wu, X.; Wu, Z.; and Jiang, Y.-G. 2024. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. *arXiv preprint arXiv:2412.03859*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22490–22499.

Zhou, D.; Li, Y.; Ma, F.; Zhang, X.; and Yang, Y. 2024. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6818–6828.

Zhou, D.; Xie, J.; Yang, Z.; and Yang, Y. 2025a. 3dis-flux: simple and efficient multi-instance generation with dit rendering. *arXiv preprint arXiv:2501.05131*.

Zhou, J.; Li, J.; Xu, Z.; Li, H.; Cheng, Y.; Hong, F.-T.; Lin, Q.; Lu, Q.; and Liang, X. 2025b. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13093–13103.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.