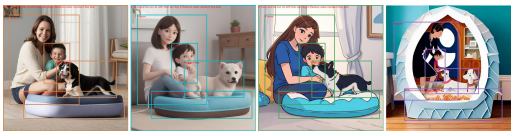
HiCo: Hierarchical Controllable Diffusion Model for Layout-to-image Generation

Bo Cheng **Yuhang Ma** **Liebucha Wu** **Shanyuan Liu**
Ao Ma **Xiaoyu Wu** **Dawei Leng** † **Yuhui Yin**

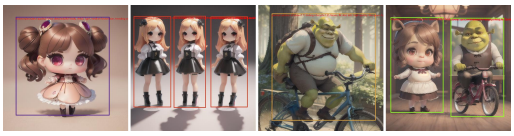
360 AI Research
{chengbo1, mayuhang, wuliebucha, liushanyuan}@360.cn
{maao, wuxiaoyu1, lengdawei, yinyuhui}@360.cn

(a) Layout grounded generation with closed-set short descriptions

(b) Layout grounded generation with open-ended fine-grained descriptions



(c) HiCo is compatible with different SD variants



(d) Multi-concept generation by HiCo with multi LoRAs

(e) Fast inference with HiCo-LCM / HiCo-Lightning

Figure 1: HiCo model serves to enhance layout controllability for text-to-image generation, by integrating bounding box condition of different objects hierarchically. The proposed unique conditioning branch structure can produce more harmonious and holistic image with complex layout.

Abstract

The task of layout-to-image generation involves synthesizing images based on the captions of objects and their spatial positions. Existing methods still struggle in complex layout generation, where common bad cases include object missing, inconsistent lighting, conflicting view angles, etc. To effectively address these issues, we propose a **Hi**erarchical **Co**ntrollable (HiCo) diffusion model for layout-to-image generation, featuring object seperable conditioning branch structure. Our key insight is to achieve spatial disentanglement through hierarchical modeling of layouts. We use a multi branch structure to represent hierarchy and aggregate them in fusion module. To evaluate the performance of multi-objective controllable layout generation in natural scenes, we introduce the

†Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4094

HiCo-7K benchmark, derived from the GRIT-20M dataset and manually cleaned. https://github.com/360CVGroup/HiCo_T2I.

1 Introduction

Text-to-image (T2I)[1–4] diffusion models like Stable Diffusion, GLIDE[3], have rapidly developed for their exceptional quality and diverse generative capabilities. However, the T2I models lack fine-grained control over visual composition and spatial layout via text prompts alone.

Layout-to-image generation[1, 4, 5], which aims to produce high-quality and realistic images from layout conditions. This article mainly studies the generation of layout images based on object text description and positional coordinates. Existing methods can be mainly divided into two categories: training-free methods[6–9] and training-based methods[4, 10–12]. Training-free methods usually use cross-attention to get the ability to control position. Training-based methods typically utilize new network structures or specific attention. As shown in the Figure 2, in complex scenarios, the training-free method represented by CAG[6] has a serious problem of object missing. The training-based methods represented by GLIGEN[4] can alleviate the phenomenon of object missing, but the generated images often exhibit distortion.

Figure 2: The generation of CAG[6], GLIGEN[4] and HiCo in complex layouts.

To address the issues, we propose the Hierarchical Controllable(HiCo) diffusion model. The approach disentangle the spatial layouts by multiple branch networks. Specifically, the branch design of HiCo is inspired by external condition introduction methods similar to ControlNet[13] and IP-Adapter[14]. Hierarchical layout features are separately extracted by branches of weight sharing, and refinedly aggregated with Fuse Net. The overall architecture of our approach is shown in Figure 3. Our method demonstrates superior performance in terms of object omissions and image quality as shown in Figure 2. This is attributed to our hierarchical modeling approach, which has particular advantages in generating complex layouts.

To further validate the effectiveness of our method, we conducted experiments on both the closed-set COCO dataset and the open-ended GRIT dataset, and achieved excellent performance on both. Furthermore, our method exhibits a flexible scalability, including switching checkpoints and integrating plugins like LoRA, LCM. Refering to Figure 1 for details.

Due to the lack of objective benchmark for multi-objective controllable layout in natural scenes, we introduce the HiCo-7K benchmark. HiCo-7K is uniformly sampled from GRIT-20M[15] dataset, revalidated for object regions using Grounding-DINO[16], and filtered based on semantic relevance using CLIP[17]. Furthermore, we conduct deep manual cleaning on it.

Our primary contributions are shown as following:

1. We propose the HiCo model, which achieves spatial disentanglement through hierarchical modeling of layouts. Our method can generate more desirable images in complex scenarios, and exhibit a flexible scalability.

2. We propose a benchmark HiCo-7K, which has been revalidated and cleaned by algorithms and professionals. It can objectively evaluate the task of layout image generation in natural scenes.

3. Our method achieves state-of-the-art performance on both the open-ended HiCo-7K dataset and the closed-set COCO-3K[18] dataset.

https://doi.org/10.52202/079017-4094

2 Related Work

2.1 Diffusion Models

Diffusion models are generative models that synthesize images from random noise by iterative image denoising. They have showed its excellent potential to generate high semantic relevance and aesthetic quality images than GAN-related models[19]. DDGAN[20], DiffusionVAE[21] study the combination of diffusion model and other generative methods. Compared to denoising diffusion probabilistic models (DDPM)[22], DDIM[23] and PLMS[24] mitigate the lengthy sampling procedure by reducing number of sampling steps. The latent diffusion model(LDM)[1] leverages VAE to encode images to latent codes with smaller resolution, saving efforts to train super-resolution models for generating high-resolution images. ControlNet[13] and IP-Adapter[14] enabled additional image-guided conditions into frozen T2I diffusion models (e.g., sketch, segmentation masks, canny edge).

2.2 Layout-to-Image Generation

Layout-to-image technology seeks to generate realistic images with corresponding spatial layouts based on graphical or textual inputs that convey layout information. Early layout-to-image techniques primarily used GAN-related technologies[25–27]. Although these works achieved encouraging progress, their generation effects and application scenarios were extremely limited.

Different guiding conditions[28–31] endow diffusion models with diverse capabilities. Researchers have proposed various methods to generate layout-controllable images using object descriptions and spatial positions. They mainly design a brand new network or special attention, such as layout attention or attention redistribution. Our approach build on basic pre-trained model by introducing weight-shared multi-branch structures for enhanced local controllability.

Researches[32, 33] on the combination of large language model(LLM) and diffusion model, can enhance strong performance in instruction-following and controllable image generation and editing tasks.Unlike HiCo which requires layout and image specification directly from user input, LMD[34] and SLD[35] resort to LLM for image scene description and layout arrangement via text automatically. It's worth pointing out that HiCo can be integrated with LMD and SLD, by treating HiCo as the replacement of their train-free layout controllable image generation module.

3 Method

3.1 Preliminary

The SD model is a diffusion model that operates within the latent space, which consists of three main components: The VAE-encoder[36] encodes images into the latent space, while the VAE-decoder reconstructs the latent representation into realistic images. The CLIP[17, 37] text encoder projects a sequence of tokenized texts into the sequence embedding. The UNet model is trained to predict the added Gaussian noise using LDM loss.

$$L_{LDM} = E_{\epsilon(x),t,\epsilon\sim N(0,1)}\left[||\epsilon - \epsilon_\theta(z_t, t)||_2^2\right] \tag{1}$$

where t is uniformly sampled from time steps $\{1, \ldots, T\}$, z_t is t-step latent of the input. ϵ_θ is noise prediction model.

The conditional guided generation in SD involves incorporating text conditions, reference images, masks, and other conditions into the SD model through various techniques. Controlnet is a common method of introducing external control conditions through collateral structures, and its training goal is to predict noise at different stages with a learnable branch network, denoted as ϵ_θ. Given the latent z_0, the model adds noise to reach z_t, where t represents the number of noise-adding steps. Here, c_t indicates the textual control condition, and c_f stands for a specific control condition. The objective function can be expressed as a function $L_{condition}$.

$$L_{condition} = E_{z_0,t,c_t,c_f,\epsilon\sim N(0,1)}\left[||\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)||_2^2\right] \tag{2}$$

 https://doi.org/10.52202/079017-4094

3.2 HiCo Diffusion Model

We adopt a common external condition introduction method similar to ControlNet and IP-Adapter, and explore its innovative application in the design of controllable layout networks. We proposed a multi-branch HiCo Net, which independently models the background and multiple foregrounds, and hierarchically expresses the local semantics and spatial layout relationship of the image in a fine-grained manner. On the fusion of branches, we experimented with a variety of fusion methods and proposed a non-parametric Fuse Net, decouples branches through masks and achieves excellent performance. The overall network structure is shown in Figure 3.

HiCo Net.The multi-branch HiCo Net is introduced to generate the global background and foreground instances for different layout regions. The branch of HiCo Net adopts the structure of ControlNet, independently decoupling layout conditions to hierarchically model foreground objects.

$$L_{HiCo} = E_{z_0,t,c_t^k,c_b^k,\epsilon\sim N(0,1),k\sim[1,K]}\left[||\epsilon - \epsilon_\theta(z_t, t, \mathcal{F}(c_t^k, c_b^k)||_2^2\right] \tag{3}$$

Here, c_t^k represents the textual control condition of k-th sub area, and c_b^k represents the bounding box control condition of k-th sub area. F represents the Fuse Net.

We define the instruction to a layout-to-image model as a composition of sub-caption and sub-boundingbox.

$$\begin{aligned} \text{Instruction }: & y = (c_i, b_i), i \in [1, K], with \\ \text{caption }: & c = [c_g, c_1, ..., c_i, ..., c_K] \\ \text{boundingbox }: & b = [b_g, b_1, ..., b_i, ..., b_K] \end{aligned} \tag{4}$$

where K represents the number of regions, with c_g and b_g representing the textual descriptions and positions of the background image, c_i and b_i corresponding to the descriptions and positions of the individual regions. The HiCo Net processes the different regional conditions to generate intermediate features that correspond to the textual descriptions within the predefined regions.

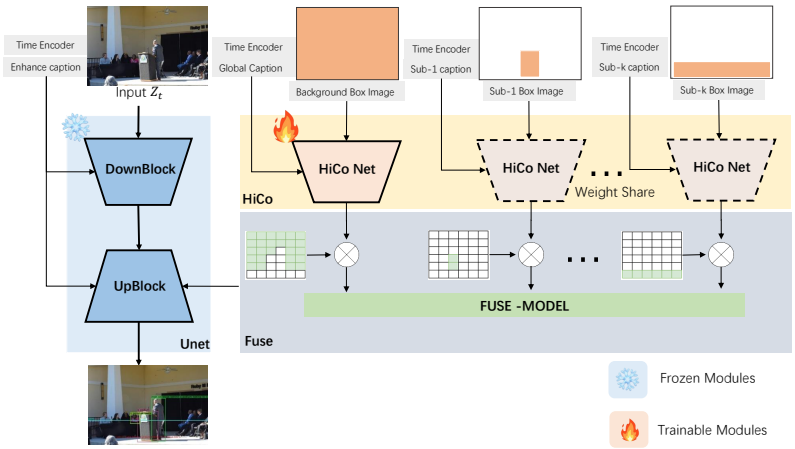


Figure 3: The overall architecture of our approach.

Fuse Net.The module fuses intermediate features from sub-regions and acts as the external features of the UNet model. It have different forms, including summation, averaging, mask, learnable-weights, and other methods. Based on different task types, various fusion structures can be selected. Our approach mainly decouples the content of different foreground and background regions through the mask fusion form. The detailed fusion process is defined as:

$$\mathcal{F}(c_t, c_b) = M_g \cdot \epsilon_{hico}(z_t, t, c_t^g, c_b^g) + \sum_{k=1}^{K} M_k \cdot \epsilon_{hico}(z_t, t, c_t^k, c_b^k), M_g = 1 - \sum_{k=1}^{K} M_k \tag{5}$$

Here, c_t^k, c_b^k represent the text and position control conditions of *k-th* sub area, and c_t^g, c_b^g represent the textual description and position of the background image. M_k represents the *k-th* object area mask information, and M_g represents the background area mask information.

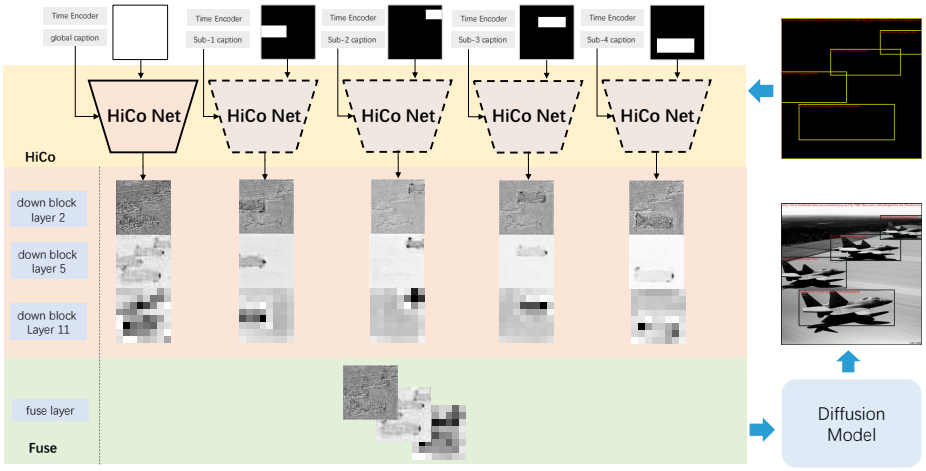


Figure 4: The visualization of the features on different layers of the HiCo branch and Fuse Net.

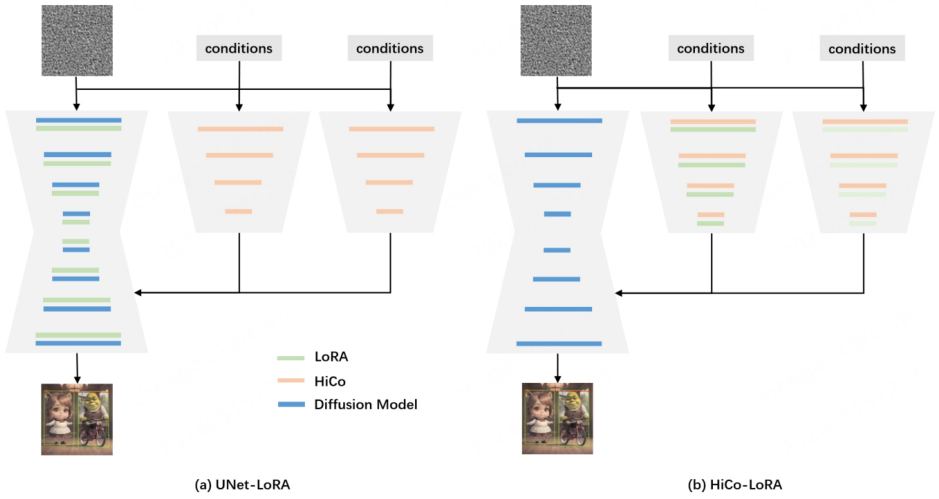


Figure 5: The model fine-tuning technique based on various positions of LoRA. (a) Adding LoRA parameters on UNet. (b) Adding LoRA parameters on HiCo.

3.3 Hierarchical Controllable Analysis

The introduction of external control conditions through a side branch structure is a common condition guidance approach in diffusion models. Our HiCo model employs an innovative weight-sharing mechanism across its branch structures, which adeptly generates distinctive features for both foreground objects and the background image, tailored to the specific conditions of the caption and bounding box. These features are strategically integrated during the upsampling phase, a critical step in the diffusion model's generative workflow.

Figure 4 depicts the generative process of the HiCo model for four foreground objects, showcasing the visualization of features of layer-2, layer-5, and layer-11 at the 50th denoising step during the downsampling stage. The visualization of shallow features reveal that various branches exhibit a more pronounced response to their respective layout areas. The intermediate features indicate further refinement of object positions, contours, and semantics generated by different branches. Furthermore, the deep features underscore the model's adept handling of regional information, essential for the controlled layout of the image. The fusion process of the HiCo branches employs a mask technique.

https://doi.org/10.52202/079017-4094

However, it is crucial to elucidate how this fusion contributes to the overall semantic coherence of the spatial layout.

3.4 Expansion Capability

Low-Rank Adaptation (LoRA)[38] stands out as an efficient fine-tuning technique. Our HiCo model exhibits excellent compatibility with rapid generation plugins, whether it's utilizing LCM-LoRA[39]for quickly generating 512×512 resolution images on HiCo-SD1.5 or Lightning-LoRA/Lightning-UNet[40] for quickly producing 1024×1024 resolution images on HiCo-SDXL.

Multi-concept[41–43] controllable layout generation effectively blend different elements such as characters, style, and objects into a cohesive image. Integrating multi-LoRA models of the same type into the UNet can easily lead to confusion of different conceptual features. This will make it difficult to naturally generate different conceptual elements in the same image. Specifically, training LoRA on the HiCo branch, as shown in Figure 5, has been shown to significantly enhance performance in scenarios requiring the injection of multiple concepts. For more results, please refer to Appendix A.5.

4 Evaluation

4.1 Dataset

The HiCo model can use various types of grounding data to achieve controllable layout generation across different scenarios.

Open-ended Grounded Text2Image Generation. For training datasets, the fine-grained detailed description data, comprises 1.2 million image-text pairs with regions and descriptions sourced from GRIT-20M[15]. We performed algorithms and manual cleaning on the raw data, resulting in a dataset with an average of 4.3 objects per image.

Closed-set Grounded Text2Image Generation. For training datasets, the coarse-grained categorical description data, we selecte a subset of approximately images from COCO Stuff[18] based on criteria such as region size, labeled as COCO-75K. This dataset includes 171 categories and an average of 5.5 objects per image.

Evaluation Dataset. The evaluation datasets include two types of data: the coarse-grained COCO-3K and the fine-grained HiCo-7K. We introduce the HiCo-7K benchmark for evaluation of multi-objective controllable layout in natural scenes. The HiCo-7K dataset is derived from GRIT-20M and has undergone iterative cleaning through both algorithmic and manual processes. It consists of 7,000 images, with an average of 3.78 objects per image. We have detailed the construction pipeline of the custom dataset HiCo-7K in Appendix A.2.

4.2 Experimental Setup

We can apply the HiCo architecture to various network structures such as SD1.5, SD2.1, or SDXL[44] to achieve controllable generation. Specifically, for SD1.5, We utilize the AdamW optimizer with a fixed learning rate of 1e-5 and train the model for 50,000 iterations with a batch size of 256. We train HiCo with 8 A100 GPUs for 3 days. Training HiCo-SDXL requires more iterations and fine-tuning on a smaller set of high-quality data. After training, our HiCo model also supports rapid generation plugins with LoRA, LCM, SDXL-Lightning.

4.3 Quantitative Results

4.3.1 Coarse-Grained Closed-set Text2Img Generation.

We train and evaluate HiCo model on COCO-stuff datasets to develop its layout-to-image capabilities. We use Frechet Inception Distance(FID)[45] to evaluate the perceptual quality of the generated images. We use YOLO score[5] to evaluate the recognizability of the object in the generated images. YOLO score uses a pretrained YOLOv4[46] model to detect the objects in the generated images.

The HiCo model achieves the best results in image fidelity and spatial semantic dimensions, and the generated images have a more beautiful and reasonable visual effect as shown in Figure 6. Meanwhile,

it can generate images with a resolution of 512 and not just the categories of COCO-3K, which is an ability that other models do not possess. The quantitative comparison is presented in Table 1.

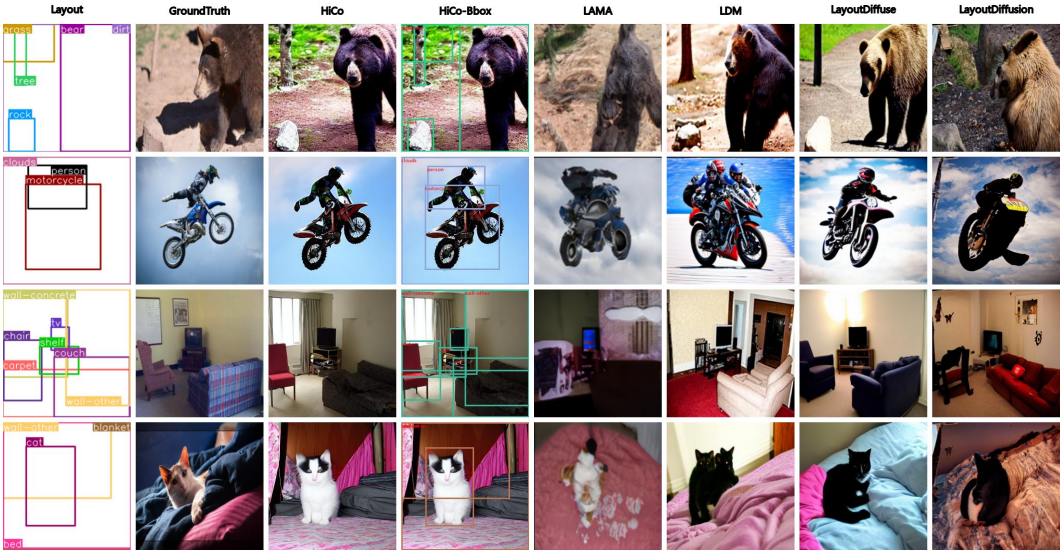


Figure 6: Qualitative comparison of HiCo and other methods on COCO-3K. Compared with other methods, HiCo can generate high-quality images under complex layout conditions. The cases generated by LAMA, LDM and LayoutDiffuse is referenced from LayoutDiffuse.

Table 1: Quantitative comparison of HiCo and other methods on COCO-3K. All generated images are evaluated under 256×256 resolution. † indicates that the experimental value is referenced from LayoutDiffuse.

Methods	FID↓	AP↑	AP50↑	AP75↑	AR↑
LAMA†[5]	31.12	20.61	28.54	22.69	26.48
LDM†[1]	24.60	32.13	54.23	34.34	39.86
LayoutDiffuse†[10]	20.27	36.58	**59.59**	38.06	46.09
LayoutDiffusion[11]	48.77	14.20	24.2	14.3	16.2
HiCo(Ours)	**20.02**	**36.60**	58.10	**39.4**	**46.60**

4.3.2 Fine-Grained Open-ended Text2Img Generation.

We train and evaluate HiCo using a high quality data of natural scenes. We use FID, Inception Score (IS)[47], Learned Perceptual Image Patch Similarity (LPIPS)[48] to evaluate the perceptual quality of the images generated with layout information. The results are presented in Table 2. Our model can generate high-quality images with rich objects even in complex scenarios, as shown in Figure 7.

Furthermore, we use Grounding-DINO[16] to detect the instance caption and calculate the maximum IoU between the detection boxes and the ground truth box. If the maximum IoU is higher than the threshod 0.5 and the Local CLIP Score[49] of them is higher than 0.2, we mark it as position correctly generated. We use AR, AP, AP50 and AP75 to evaluate the spatial performance. The results are presented in Table 3.

Table 2: Quantitative comparison of perception dimension with other models on HiCo-7K. All generated images are evaluated under 512×512 resolution. The results indicate that HiCo has better fidelity and perception.

Methods	FID↓	IS↑	LPIPS↓
MtDM[9]	24.83	25.27±1.54	0.76±0.05
GLIGEN[4]	19.65	26.59±1.45	**0.73±0.07**
CAG[6]	19.37	26.24±1.42	0.76±0.06
MIGC[12]	26.92	22.17±1.16	0.77±0.07
InstanceDiff[50]	16.99	26.19±1.09	**0.73±0.06**
HiCo(Ours)	**14.24**	**28.31±0.79**	**0.73±0.07**

https://doi.org/10.52202/079017-4094

Table 3: Quantitative comparison of spatial location dimensions on HiCo-7K. Experiments show that HiCo has better positional control and image text consistency.

Methods	LocalCLIP Score↑	LocalIoU Score↑	AR↑	AP↑	AP50↑	AP75↑
GroundT	25.07	66.40	49.45	30.13	42.77	31.75
MtDM	24.81	25.02	5.21	0.59	2.45	0.13
GLIGEN	24.96	48.54	28.33	12.67	24.15	12.14
CAG	24.54	24.43	4.62	0.47	2.01	0.21
MIGC	24.94	48.03	25.8	10.18	23.42	7.56
InstanceDiff	25.09	51.58	33.92	16.76	27.45	17.14
HiCo(Ours)	**25.17**	**59.98**	**41.21**	**21.53**	**33.22**	**22.66**

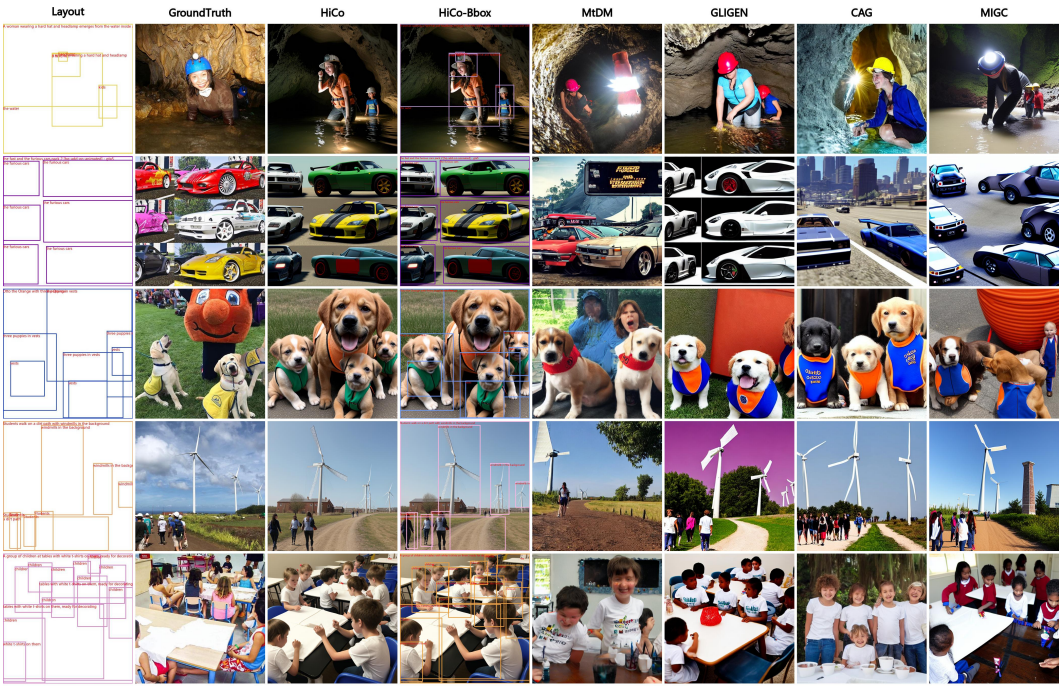


Figure 7: Qualitative comparison with other models on HiCo-7K. Compared with other methods, HiCo can generate high-quality images for both simple and complex layout information.

Zero-shot Evaluation. We further evaluate the zero-shot performance of HiCo trained in natural scenes on COCO-3K, detailed results are shown in table 4. The preferences of HiCo controllability is not the best on COCO-3K. The reason for this problem is that our model was trained on a 1.2M fine-grained long caption, which belongs to out of distribution data for COCO data.

Table 4: Quantitative comparison of spatial location dimensions of zero-shot capability on COCO-3K.

Methods	LocalCLIP Score↑	LocalIoU Score↑	AR↑	AP↑	AP50↑	AP75↑
GLIGEN	25.83	73.17	53.39	36.87	65.35	38.09
MIGC	25.94	75.06	57.55	40.82	70.57	42.21
InstanceDiff	**26.60**	**85.91**	**78.10**	**65.28**	**80.56**	**71.93**
HiCo-Real(Ours)	26.27	79.57	68.40	52.47	78.61	57.05

4.4 Human Evaluation

We use a multi-round and multi-participant cross-evaluation approach to assess human preferences, focusing on aspects such as object quantity, spacial location, and global image quality. Details on the experimental setup can be found in supplementary material of Appendix A.4.

Spatial Location & Semantics.Target quantity dimension assesses whether the number of objects in the generated image aligns with the preset value. The semantics and position dimension examine the relevance of the objects to the textual description and their regional placement within the image.

https://doi.org/10.52202/079017-4094

Global Image Quality. The image global quality dimension includes five sub-dimensions: relativity, clarity, rationality, aesthetics and risk.

Table 5: The human evaluation results encompass two dimensions: spatial semantic location and global image quality. The numerical range is from 0 to 100, with a higher score indicating better performance. It should be noted that the risk dimension is the proportion of generated risk images, the lower the better.

Method	Spetial Location& Semantics			Global Image Quality					Overall
	quantity	semantics	position	relativity	clarity	rationality	aesthete	risk↓	
GroundT	100	100	100	93.36	90.69	97.34	92.56	0.00	95.62
SD-Real	80.56	-	-	67.41	83.58	60.62	63.67	2.00	-
MtDM	54.96	33.26	22.04	48.08	59.04	40.49	40.91	9.00	39.41
GLIGEN	80.44	73.56	72.78	56.78	61.67	53.67	51.89	4.44	67.76
CAG	50.89	32.11	37.89	48.67	55.11	49.00	44.89	7.33	43.94
MIGC	69.89	62.44	62.67	49.22	57.44	47.00	46.67	6.11	59.03
InstanceDiff	88.33	80.00	66.78	61.56	62.33	56.00	55.33	2.50	70.54
HiCo	**89.44**	**86.11**	**82.22**	**63.67**	**82.33**	**58.33**	**60.89**	**2.00**	**78.08**

Table 5 demonstrates the human evaluation results conducted on 300 controllable layout images. The results indicate that, in terms of spatial position and semantic dimension, our approach outperforms other models. Moreover, it achieves near-par performance to the RealisticVisionV51 model(SD-Real) in the fine-grained dimension of global image quality, suggesting that despite the enhanced controllability, the generative capabilities of our model remain robust and effective.

4.5 Ablation Studies

The ablation focuses on UNetGlobalCaption(**UGC**), GlobalBackgroundBranch(**GBB**) and FuseNet(**FN**). Furthermore, FuseNet is a non-parametric network that includes the following types, such as (1)Summation. (2)Average. (3)Mask. Experiments are performed on HiCo-7K using the HiCo-base model with the same training amount. The results are presented in Table 6.

Table 6: The results of ablation studies on HiCo-7K of UGC, GBB and FN.

UGC	GBB	FN	FID↓	AR↑	AP↑	LocalCLIP Score↑	LocalIoU Score↑
×	✓	sum	20.81	33.12	16.54	25.18	52.25
×	✓	avg	18.78	36.3	18.58	25.02	56.67
×	✓	mask	**15.26**	**39.51**	**20.02**	25.22	**59.07**
✓	✓	mask	16.85	38.19	19.05	25.23	57.03
✓	×	mask	21.82	30.07	12.07	**25.24**	49.64

4.6 Inference Performance

For inference run time and memory usage, we conducted two additional comparisons. The first comparison is horizontal, among 6 different models including: GLIGEN, InstanceDiff, MIGC, CAG, MtDM as well as our HiCo. Specifically, we evaluated the inference time and GPU memory usage for directly generating 512×512 resolution images on the HiCo-7K using a 24GB VRAM 3090 GPU, results in Figure 8-(a) show that HiCo has the shortest inference time and the 2nd lowest GPU memory footprint.

The multi-branch of HiCo has two inference modes: "parallel mode" and "serial mode". In order to verify the performance advantages of HiCo when the number of objects increases, the second comparison is vertical: we assessed the inference time and GPU memory usage for generating 512×512 resolution images on the HiCo-7K with different number of objects. Since each object is processed by a separate branch in HiCo, the inference can be accelerated by inferring all the branches in one batch, in "parallel mode", which as shown in Figure 8-(b), is much faster than the "serial mode", inferring all the branches one by one in serial.

 https://doi.org/10.52202/079017-4094

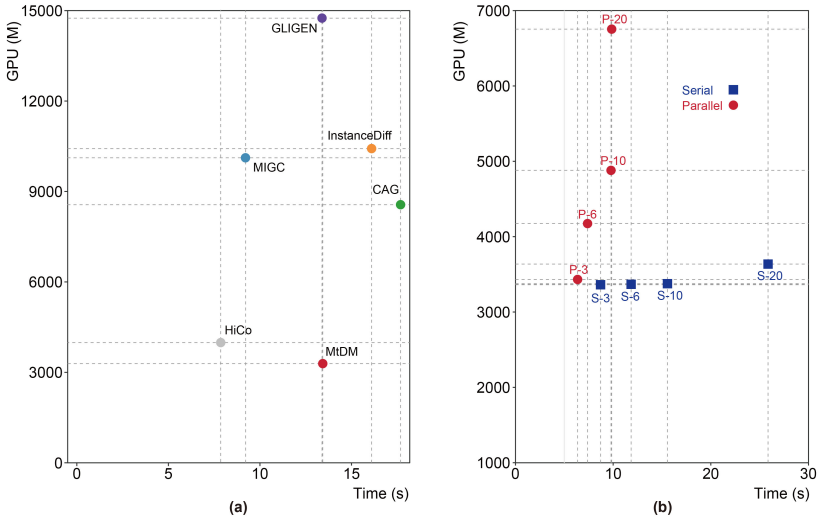


Figure 8: Quantitative comparison of inference performance. (a) Comparison between different methods on HiCo-7K. (b) Comparison between different object quantities on HiCo.

5 Conclusion

HiCo is a controllable layout generation model based on diffusion model, guided by multiple branch structures. This approach allows users to specify the location and detailed textual descriptions of target regions while maintaining the rationality and controllability of the generated content. Through training and testing on data with varying degrees of granularity in natural scenarios, as well as conducting algorithm metric evaluation and subjective human assessment, the superiority of this method is demonstrated. However, there is still potential for further improvement, especially in the areas of image content editing and integrating multiple style concepts. By combining current controllable generation capabilities can boost the overall playability of AI-generated artwork.

Limitation. The complex interactions and occlusion order of overlapping areas are significant challenges for HiCo model and even the field of layout-to-image generation. HiCo achieves hierarchical generation by decoupling each object's position and appearance information into different branch, meanwhile controlling their overall interactions through a background branch with global prompt and the Fuse Net. HiCo is capable of handling complex interactions in overlapping regions by FuseNet. The occlusion order of overlapping objects is also specified via the global prompt by text description. But since there lacks corresponding occlusion order train data, the success rate is far from optimal. For current version of HiCo, there indeed lacks of more explicit mechanism for occlusion order controlling. For more results, please refer to Appendix A.3. We also found that the HiCo model still does not handle the generation of complex layouts for multiple concepts of LoRA very well. We intend to conduct further research to explore solutions to these issues in the future.

Social impact aspect. Our model is designed to assist users in generating creative images with controllable layouts. However, there is a risk of misuse of our method to generate inappropriate or sensitive content. Therefore, we believe that regulating the application of such models and developing risk detection tools are crucial. This will facilitate the progress of AI technology for the benefit of humanity.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[3] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[4] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[5] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13819–13828, 2021.

[6] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024.

[7] Peiang Zhao, Han Li, Ruiyang Jin, and S Kevin Zhou. Loco: Locally constrained training-free layout-to-image synthesis. *arXiv preprint arXiv:2311.12342*, 2023.

[8] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023.

[9] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

[10] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023.

[11] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023.

[12] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. *arXiv preprint arXiv:2402.05408*, 2024.

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[14] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[15] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[18] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[19] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

 https://doi.org/10.52202/079017-4094

[20] Yu Chen, Weida Zhan, Yichun Jiang, Depeng Zhu, Renzhong Guo, and Xiaoyu Xu. Ddgan: Dense residual module and dual-stream attention-guided generative adversarial network for colorizing near-infrared images. *Infrared Physics & Technology*, 133:104822, 2023.

[21] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[24] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[26] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5070–5087, 2021.

[27] Bo Wang, Tao Wu, Minfeng Zhu, and Peng Du. Interactive image synthesis with panoptic layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7783–7792, 2022.

[28] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[29] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[30] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[31] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

[32] Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, et al. Llms meet multimodal generation and editing: A survey. *arXiv preprint arXiv:2405.19334*, 2024.

[33] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

[34] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

[35] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336, 2024.

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[37] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[38] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023.

[39] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

[40] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

[41] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[42] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*, 2024.

[43] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023.

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[45] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[46] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[47] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[49] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023.

[50] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024.

 https://doi.org/10.52202/079017-4094

A Appendix / supplemental material

Overview

In this supplemental material, we provide the following items:

- Training and Inference Strategies.
- HiCo-7K Benchmark.
- Limitaion Discussion.
- Manual Evaluation Criteria.
- More results on HiCo, including different layout quantities, different base models, and different resolutions.

A.1 Training and Inference Strategies

During the training stage, we only optimize the parameters of the HiCo Net, while keeping the SD base model parameters fixed. The Fuse Net can use either non-parametric methods, like summation, averaging, mask or a parametric method, like simple MLP structure.

During the inference stage, the structure of the Fuse Net can be reasonably selected according to the size and importance of the foreground and background regions to achieve controlled generation effects in different scenarios. Additionally, LoRA network parameters can be added to the HiCo Net to fine-tune and learn new tasks, such as personalization, multi-language controllable generation and other tasks.

A.2 HiCo-7K Benchmark

We have detailed the construction pipeline of the custom dataset HiCo-7K in Figure 9. We found that GRIT-20M has some issues, such as a low labeling rate for objects with the same description and target descriptions being derived solely from the original captions. Compared to GRIT-20M, the pipeline of the HiCo-7K is as follows.

1. Extracting noun phrase. We use spaCy to extract nouns from captions and the LLM VQA model to remove abstract noun phrases. Meanwhile, we use the GroundingDINO model to extract richer phrase expressions.
2. Grounding noun phrase. We use the GroundingDINO model to obtain the bboxes. After that, we use NMS and CLIP algorithms to clean the bboxes.
3. Artificial correction. To address the issue of algorithmic missed detections for multiple objects with the same description in an image, artificial correction is employed to further enhance the labeling rate of similar objects.
4. Multi-captions with bounding box. We expand the basic text from the original captions and use GPT-4 to re-caption the target regions. The dataset of HiCo-7K contains 7000 expression-bounding-box pairs with referring-expressions and GPT-generated-expressions.

A.3 Limitaion Discussion

HiCo achieves hierarchical generation by decoupling each object's position and appearance information into different branch, meanwhile controlling their overall interactions through a background branch with global prompt and the Fuse Net. The Fuse Net combines features from foreground and background regions, as well as intermediate features from side branches, then integrates them during the UNet upsampling stage. As illustrated in Figure 10-(a),Figure 10-(b), HiCo is capable of handling complex interactions in overlapping regions without any difficult.The occlusion order of overlapping objects is also specified via the global prompt by text description, for example "bowl in front of vase", as illustrated in Figure10-(c),Figure10-(d). But since there lacks corresponding occlusion order train data, the success rate is far from optimal. For current version of HiCo, there indeed lacks of more explicit mechanism for occlusion order controlling.

We recognize this problem as our future work. Actually we're already working on the occlusion order data curation, which is a quite challenging task as it requires reliable depth estimation besides the object detection bounding boxes.

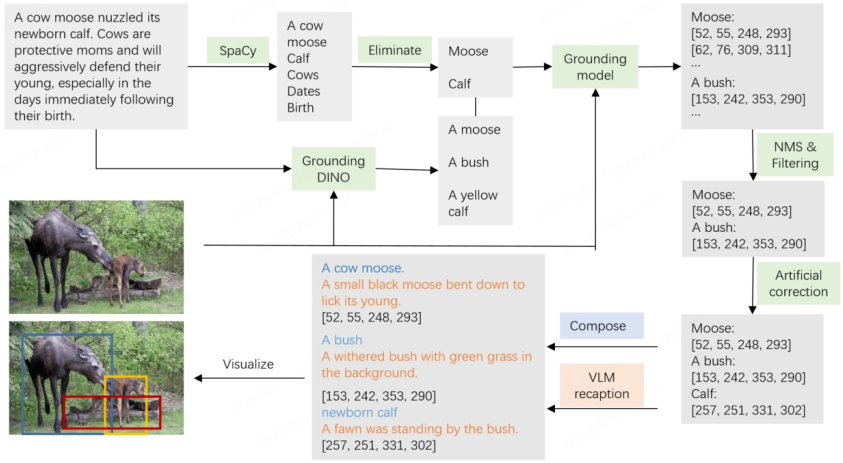


Figure 9: The pipeline of constructing HiCo-7K with grounded image caption pairs.

Figure 10: The image generated in overlapping and interactive scenarios.The generated caption is displayed in the image.

https://doi.org/10.52202/079017-4094

A.4 Manual Evaluation Criteria

Evaluation Dimensions

We designed the evaluation dimensions by reviewing existing literature and soliciting opinions from professional designers and a large number of users. Specifically, we divided the evaluation dimensions into two categories: local and global. Local dimensions include spatial location and semantics, such as the quantities, semantics, and positions of the object bounding boxes. Global dimensions include global image quality, such as relativity, clarity, rationality, aesthetics, risk, and overall score.

Dimension Definitions

Spatial Location& Semantics. Local dimensions include three sub-dimensions: quantity indicates the number of generated objects; semantics measures the consistency between the objects within the bounding box and the textual description of that region; position measures the deviation between generated objects and ground truth by calculating the IoU of bounding boxes.

Global Image Quality. The image global quality dimension includes five sub-dimensions: relativity is primarily used to evaluate the understanding and representation of text by image content; clarity is a commonly used metric for assessing image quality; rationality is used to describe the distortion, deformation, and disarray of image content; aesthetics encompasses the overall assessment of the visual appeal of the generated image, incorporating factors such as detailed depiction, color usage, creativity, and other relatively subjective judgment elements; risk evaluates elements related to nudity, violence, terror, and other sensitive content within the image.

For risk we assess the presence of such elements and represent it using binary values of 0 or 1. For the other dimensions, we categorize them into four levels ranging from 0 to 3 and then normalize them to 0-100. We use overall score to represent the comprehensive assessment of the image.

Evaluation Execution

The evaluation team consists of professional evaluators. They have rich professional knowledge and evaluation experience, allowing them to accurately execute the evaluation tasks based on the given dimensions.

Specifically, our evaluation score is computed with the following steps:

1. According to the rules provided, evaluators rate each image on each dimension. Risk is scored as either 0 or 1, while the other dimensions range from 0 to 3.
2. Calculating the overall rate for each image: We calculate the overall rate by summing the weighted scores of each dimension. For the sub-dimensions of the local dimension, the weight is 0.2, while for the sub-dimensions of the global dimension, the weight is 0.1. The final score for overall rate is calculated as the weighted sum of the seven dimensions multiplied by the score of risk (which is either 0 or 1).
3. We calculate the mean of each dimension across the entire evaluation set, excluding the risk dimension, and mapped the scores to a scale of 0 to 100. For risk, we calculate the coverage rate of this dimension.

A.5 More Generation Results

Figure 11 shows more results generated by HiCo using fast generated plug-in LoRA. The four columns from left to right represent the generated results of 50-steps, 4-steps, 6-steps, and 8-steps, respectively.

Figure 12-(a) shows more results generated by HiCo-SD1.5 in HiCo-7K. Figure 12-(b) shows more results generated by HiCo-SDXL in HiCo-7K. The number of object layouts ranges from 3 to over 10. Despite complex layouts and rich descriptions, HiCo reliably ensures that each object is generated in the correct position with the right description.

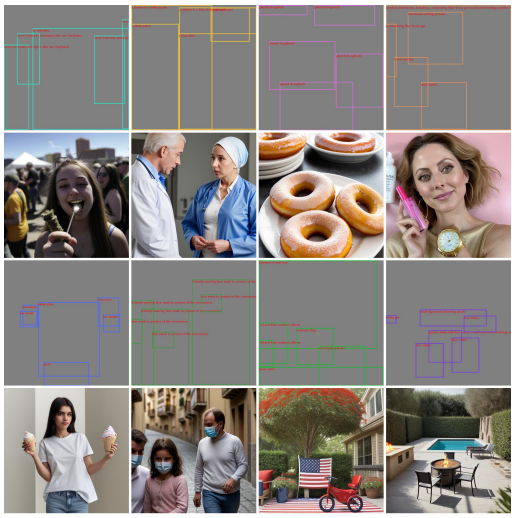
Figure 12-(c) shows more results generated by HiCo using different checkpoint from open source community. The results from the second to fourth lines are generated by the following three models, namely disneyPixarCartoon, flat2DAnimerge and dreamshaper.

Figure 12-(d) show the generation of different layout information with LoRA. Rows 1-3 demonstrate that HiCo can effectively generate complex layouts for a single concept with single LoRA. Row 4 shows the generation of complex layouts for multiple concepts with multi LoRAs using the HiCo model.

Through further experiments, we found that the HiCo model still does not handle the generation of complex layouts for multiple concepts very well. More effective methods need to be explored to address this issue.

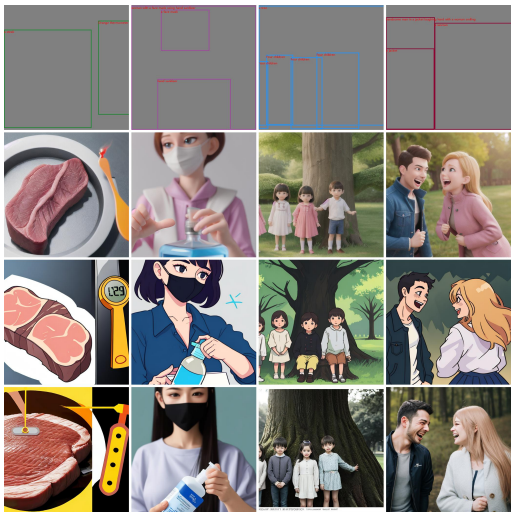
Figure 11: Qualitative experiments. Fast generation of complex layout information with HiCo-LCM / HiCo-Lightning.

https://doi.org/10.52202/079017-4094

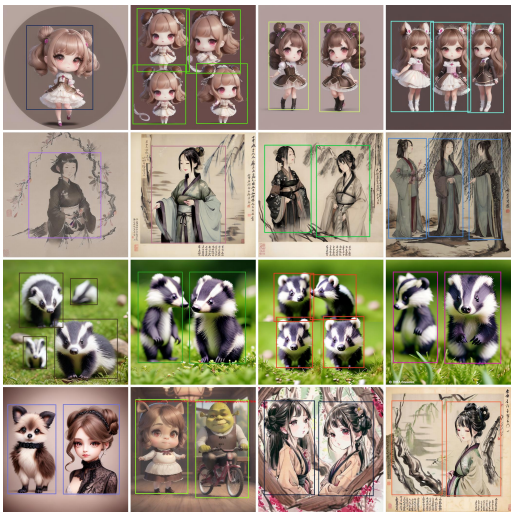


(a) Generation of complex layout information of HiCo-SD1.5.

(b) Generation of complex layout information of HiCo-SDXL.



(c) Generation of complex layout information of HiCo with different backbones.



(d) Generation of concepts by HiCo with multi LoRA.

Figure 12: Qualitative experiments generated by HiCo model to show the layout controllability.

https://doi.org/10.52202/079017-4094

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect our paper's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of the work in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-4094

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided detailed descriptions of the training data, network architecture, and experimental details to ensure the reproducibility of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper currently does not provide open-source code. However, we are actively working on it and plan to release it as soon as possible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided detailed experimental details and test configurations in Section 4.2, including datasets, hyperparameters, and so on.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the relatively high cost associated with training and evaluating generative models, it is challenging for us to report the **Experiment Statistical Significance**.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

https://doi.org/10.52202/079017-4094

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:In the experiment, we utilized the GPUs from our internal cluster, specifically training with 8 A100 cards for 3 days. If we include the resources for the exploratory phase, the total amount of resources consumed would be even greater.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper complies with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader effects of our research are examined in Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: While no specific actions have been taken to tackle this matter at present, we are dedicated to reducing the risk before introducing our generative model to the public.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We utilize open-source assets and acknowledge them accordingly in our references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

 https://doi.org/10.52202/079017-4094

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are presented in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The research conducted for this paper does not utilize crowdsourcing or engage with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The research conducted for this paper does not utilize crowdsourcing or engage with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/079017-4094