VGDiffZero: Text-to-image Diffusion Models Can Be Zero-shot Visual Grounders

Xuyang Liu[1,2*] Siteng Huang[2*], Yachen Kang[2], Honggang Chen[1], Donglin Wang[2†]

[1] College of Electronics and Information Engineering, Sichuan University; [2] School of Engineering, Westlake University
liuxuyang@stu.scu.edu.cn, {huangsiteng, kangyachen, wangdonglin}@westlake.edu.cn, honggang_chen@scu.edu.cn

Motivation: From Generative to Discriminative

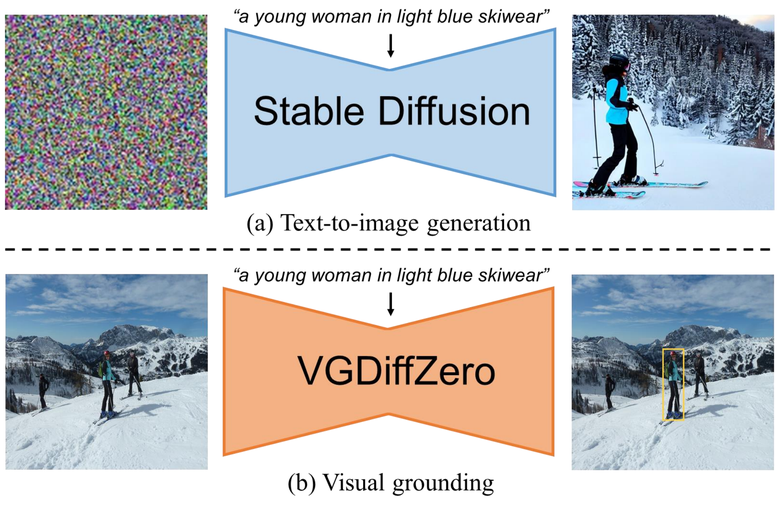

Figure 1: Illustration of two types of vision-language tasks.

Directly adapt the pre-trained text-to-image diffusion models to visual grounding.

Recent progresses [1,2] in leveraging the pre-trained diffusion models for discriminative tasks have shown **two key advantages** of text-to-image diffusion models: (1) Strong abilities of vision-language alignment. (2) Sufficient spatial relation knowledge and fine-grained disentangled concepts.

Motivated by these two advantages, we seek to directly leverage the power of pre-trained **generative** diffusion models, particularly Stable Diffusion [3], for a **discriminative** vision-language task of visual grounding.

Contributions

Our main contributions can be summarized as threefold:

- We propose VGDiffZero, a novel diffusion-based framework for zero-shot visual grounding. To the best of our knowledge, this is the **first attempt** to address the **discriminative task** of visual grounding using a **generative model** under the zero-shot setting.
- We propose a comprehensive region-scoring method that incorporates **global and local contexts** of input images to enable accurate proposal selection.
- Extensive experiments on three widely-used visual grounding benchmarks of RefCOCO, RefCOCO+ and RefCOCOg demonstrate the effectiveness of our proposed VGDiffZero for zero-shot visual grounding.

Method: Zero-shot Visual Grounding via Text-to-image Diffusion Models

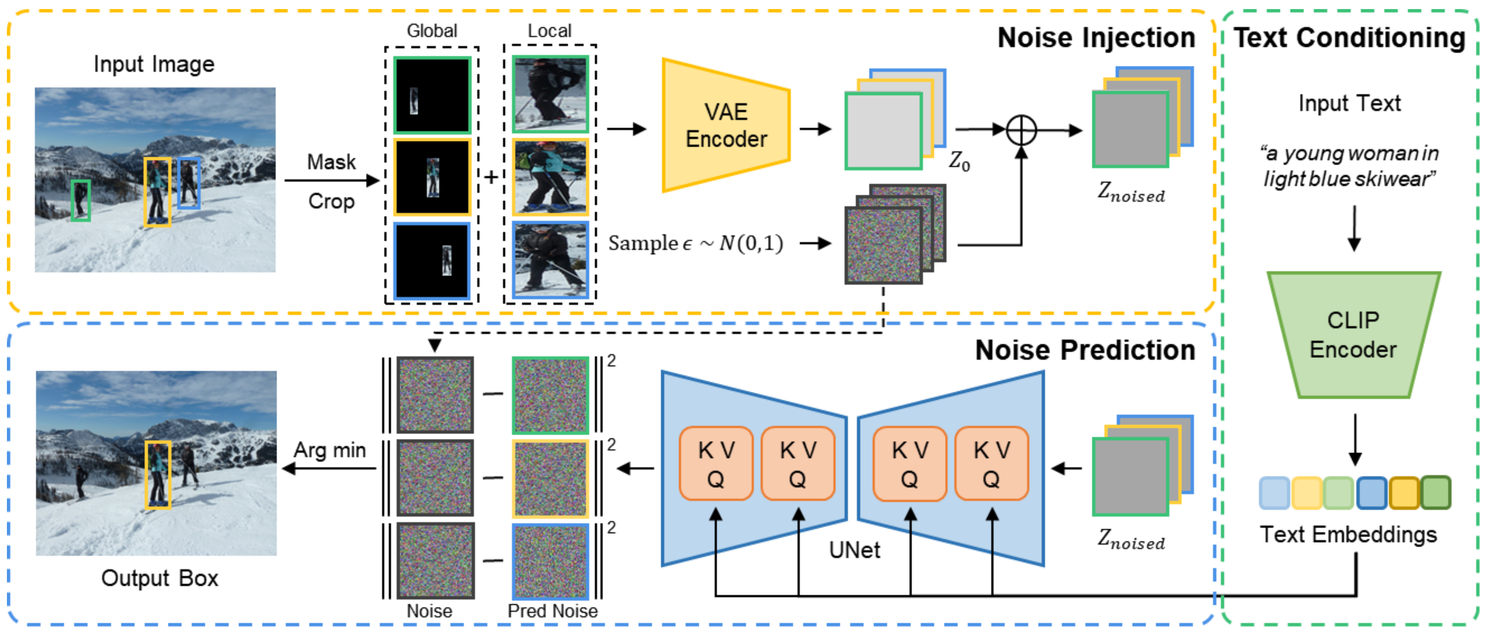

Figure 2: Overview of our VGDiffZero. Given an input image, isolated proposals are generated via cropping and masking, and then encoded individually into latent vectors Z_0. Gaussian noise $\epsilon \sim N(0,1)$ is injected into each latent vector to obtain noised latent representations Z_{noised}. Subsequently, each noised latent together with the text embeddings is fed into the UNet to select the best matching proposal as the final prediction.

Noise Injection

Multiple object proposals in the input image are first generated by a pre-trained object detector. To preserve the **global** and **local** spatial contexts along with the visual details of the proposals, we **mask** and **crop** the input image to isolate them, followed by encoding these isolated regions into latent representations Z_0 using a VAE encoder. Then Gaussian Noise $\epsilon \sim N(0,1)$ is added to these latent representations.

Noise Prediction

The input text is encoded into text embeddings by a pre-trained CLIP model. The denoising UNet then predicts the noise for the noised latent vectors Z_{noised}. Prediction errors assess the match between predicted and actual noise, with lower values suggesting improved image-text semantic alignment. The proposal with the least error is selected for prediction. Thus, VGDiffZero considers the global and local contexts of isolated proposals for comprehensive scoring.

Experiment

VGDiffZero Setup

We use the detected object proposals from a pre-trained Faster R-CNN , and each isolated proposal is resized to 512×512. VGDiffZero is built on the pre-trained Stable Diffusion 2.1-base [3] with DDPM Sampler and 1,000 timesteps. We use the text encoder initialized from CLIP-ViT-H/14.

Main Results

We evaluate VGDiffZero on three widely-used visual grounding benchmarks: RefCOCO, RefCOCO+ and RefCOCOg.

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	test A	test B	val	test A	test B	val	test
Random	15.61	13.47	18.23	16.30	13.29	19.98	18.79	18.35
CPT-Blk	26.90	27.50	27.40	25.40	25.00	27.00	32.10	32.30
Cropping	26.04	26.34	28.95	26.34	26.28	29.41	32.64	32.37
Masking	27.17	29.47	26.21	27.64	29.62	27.29	32.66	32.56
VGDiffZero w/Single IPM	26.78	29.56	27.28	27.41	29.55	27.21	32.82	32.39
VGDiffZero	**27.95**	**30.34**	**29.11**	**28.39**	**30.79**	**29.79**	**33.53**	**33.24**

Table 1: Comparison of accuracy (%) on RefCOCO, RefCOCO+ and RefCOCOg datasets under the zero-shot setting.

Analysis

- Table 2 shows that using the full expression outperforms the core phrase extraction on datasets with simple expressions, while the latter is more effective for complex sentences with multiple objects, allowing for clearer object identification.
- Table 3 shows that larger pre-training datasets improve benchmark accuracy, highlighting the value of extensive pre-training for better vision-language alignment leading to more accurate visual grounding performance.

Methods	RefCOCO	RefCOCO+	RefCOCOg
core-exp	26.86	27.13	**34.32**
full-exp	**27.95**	**28.39**	33.53

Table 2: Effect of different expression processing methods.

SD Version	RefCOCO	RefCOCO+	RefCOCOg
SD 1-2	27.11	26.73	32.34
SD 1-4	27.64	27.61	32.73
SD 1-5	27.86	27.97	32.81
SD 2-1	**27.95**	**28.39**	**33.53**

Table 3: Effect of different pre-trained diffusion models.

Conclusion

In this paper, we propose VGDiffZero, a novel zero-shot visual grounding framework that leverages pre-trained text-to-image diffusion models' vision-language alignment abilities. Through the designed comprehensive region-scoring method, our VGDiffZero can consider both the global and local contexts of each isolated object proposal. Extensive experimental results demonstrate that VGDiffZero achieves satisfactory performance on three general visual grounding benchmarks.

Misc

References

[1]. Alexander C Li, Mihir Prabhudesai, Shivam Duggal, et al., "Your diffusion model is secretly a zero-shot classifier," in *ICCV*, 2023.

[2]. Wenliang Zhao, Yongming Rao, Zuyan Liu, et al., "Unleashing text-to-image diffusion models for visual perception," in *ICCV*, 2023

[3]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al., "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

ICASSP 2024 KOREA