ROVI: A VLM-LLM Re-Captioned Dataset for Open-Vocabulary Instance-Grounded Text-to-Image Generation

Supplementary Material

This supplementary material is structured into several sections that provide additional details, aiming to present more cases and analyses in image illustrations. Specifically, it will cover the following topics:

- Sec. A: Additional generation comparisons between our GLIGEN [4] model trained on ROVI and alternative methods [4, 7, 9, 10] trained on different datasets.
- Sec. B: Dataset snapshots and analysis demonstrating ROVI's superior quality and diversity advantages through representative examples across various visual domains.
- Sec. C: Detailed illustrations and explanations of our resampling stage methodology.
- Sec. D: Cases illustrating the limitations of ROVI's annotations.
- Sec. E: Analysis of failure cases in our GLIGEN model's generation process.
- Sec. F: Additional details regarding the user study presented in the main paper.

Please note that we have limited the display of bounding box annotations. Fig. 14 presents a fully labeled case; however, the number of bounding boxes exceeds practical limits for clear visualization. Therefore, we selectively annotate only the salient objects in subsequent figures to enhance clarity and readability.

WEB CAP. President Han-Chieh Chao took a photo after awarding the first honorary alumni award to Professor Simon
VLM DESC. A photograph of two men standing in an office, with one holding a framed certificate. The man on the left is wearing glasses and a dark blue shirt, while the man on the right is wearing a light-colored polo shirt and glasses. Behind them are two large framed Chinese calligraphy pieces on the wall, and a wooden cabinet with various items on top, including a plant and some framed pictures. The room has a neutral color scheme with white walls and a wooden floor.

Figure 1. Effect of pre-detection VLM-LLM re-captioning, our key contribution. *Web CAP.* refers to the original text source, while *VLM DESC.* denotes the generated VLM description in ROVI, specifically utilizing InternVL1.5 [1]. In this context, all detected entities are summarized by an LLM, Llama3 [3], from the exhaustive VLM description. The detection effectively describes the image, and the detailed open-vocabulary categories offer a clearer view of the visual content, significantly surpassing traditional detection results that rely on basic categories. Furthermore, instance grounding in image generation can benefit from the richness of categories and compositional elements present within the VLM description.

A. More Comparisons on Generation

Figure 2. The above presents a comparison of the generated results. Note that the GLIGEN model trained on ROVI (*Our GLIGEN*) differs from the official GLIGEN [4] trained on other datasets. *InstDiff* refers to InstanceDiffusion [9] All results are based on the test data mentioned in the paper's main text, where both bounding box annotations and prompts have undergone manual verification. We omit prompts for readability. Note the combinations of attributes formed by color, texture, and overall coherence of the images, as well as their aesthetic qualities.

Figure 3. In the results mentioned above, the prompt in the third row requests the generation of a *black-and-white* image; however, neither MIGC [10] nor InstanceDiffusion achieved this objective. Notably, there are some specific open-vocabulary categories, such as *whale sculpture* in the fourth row, *teardrop pendant* in the fifth row, and *medium-length hair* in the sixth row. If these open-vocabulary categories are not correctly understood, the generated content may deviate from the requirements, resulting in inaccuracies such as depicting a teardrop as a circular shape or rendering medium-length hair as long hair.

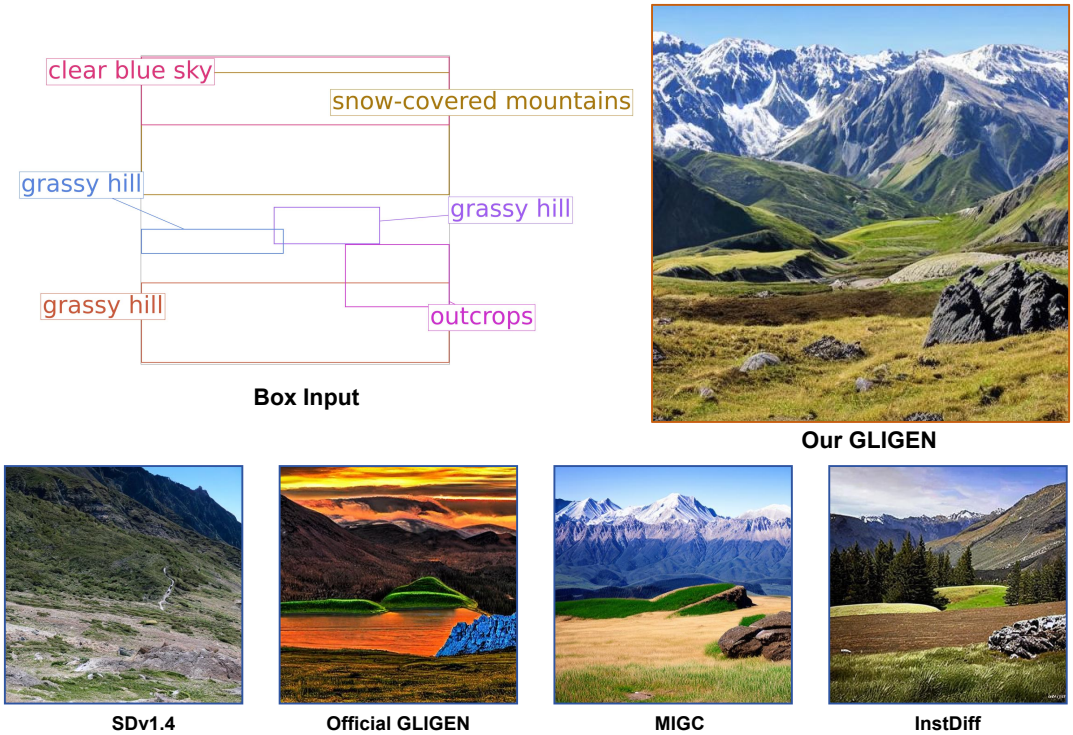


Figure 4. The superior visual quality and coherence demonstrated in our generated images can be attributed to the rigorous curation standards employed in ROVI. Specifically, ROVI's emphasis on high-resolution source images and stringent aesthetic quality criteria ensures that the training data better aligns with the demands of contemporary text-to-image synthesis tasks. This careful dataset construction directly translates to improved generation capabilities, as evidenced by the comparative results shown in the accompanying figures.

B. Dataset Snapshots & Advantages Overview

This section provides a comprehensive analysis of the key advantages of the ROVI dataset, followed by representative examples across diverse visual domains that demonstrate these capabilities in practice.

B.1. Dataset Advantages Overview

ROVI is specifically designed to meet the demanding requirements of advanced grounded text-to-image generation, where precise spatial control and rich semantic understanding are essential. Compared to existing detection-centric datasets, ROVI provides several key advantages:

Attribute-rich instance labels: Instance labels include detailed object attributes such as colors, textures, materials, and semantic properties, integrated with contextual descriptions rather than generic category names. Examples include "green algae-covered rocks" in Fig. 5 (middle-left) and "white leather office chair" in Fig. 7.

Complex scene understanding: Comprehensive coverage of visual hierarchies and relational structures enables detailed characterization of objects, environments, and their compositional interactions. This is best illustrated by the medalists in Fig. 10.

Diverse domain coverage: Broad applicability across varied content domains at large scale (1M images), including challenging visual scenarios with artistic styles and fine-grained categories.

Superior data quality and scale: Higher resolution, aesthetic quality, and category diversity with more instances per image and richer annotations aligned with natural image distributions.

Challenging category coverage: Successfully encompasses complex visual scenarios across diverse domains, including artistic styles, fine-grained object categories, and contextually-dependent semantic descriptions. Fig. 9 provides some examples with different artistic styles.

Global-local coherence: Annotations are linked to global image context through compositional elements, ensuring instance descriptions maintain contextual consistency for effective grounding. The text highlights in Fig. 10 illustrate how the generated prompt references detected instances.

B.2. Representative Examples

Figure 5. Landscapes constitute a significant proportion of our high-aesthetic dataset. Traditional detection approaches with restricted vocabularies typically produce monotonous summarization with minimal detail. Our dataset provides rich and precise annotations that effectively capture key features across various natural scenes. The middle-right example demonstrates Named Entity Recognition capabilities, where *cleveland peak* preserves specific geographical information from the original web caption that would be lost in generic detection approaches. Such detailed open-vocabulary categories offer substantially clearer characterization of landscape elements compared to basic detection outputs.

Figure 6. Human activities present significant challenges for traditional bounding box detection approaches. Users consistently require detailed object descriptions for refined generation outcomes, creating a gap between detection capabilities and generation needs. Our dataset demonstrates progress in addressing this challenge through open-vocabulary categories that facilitate detailed descriptions of human emotions, attire, and behavioral states. The examples show how our approach captures nuanced human expressions and contextual activities that standard detection vocabularies cannot adequately represent, enhancing the descriptive richness for generative applications.

Figure 7. Indoor scenes involve numerous objects and complex visual hierarchies that challenge traditional detection methods. The illustrated examples demonstrate that our VLM-LLM re-captioning workflow achieves comprehensive coverage even for intricate indoor scenarios. Open-vocabulary categories show notable advantages in capturing color variations, texture details, and combinatorial relationships between objects. The examples reveal how our approach maintains contextual understanding of object arrangements and spatial relationships that are essential for accurate scene representation in indoor environments.

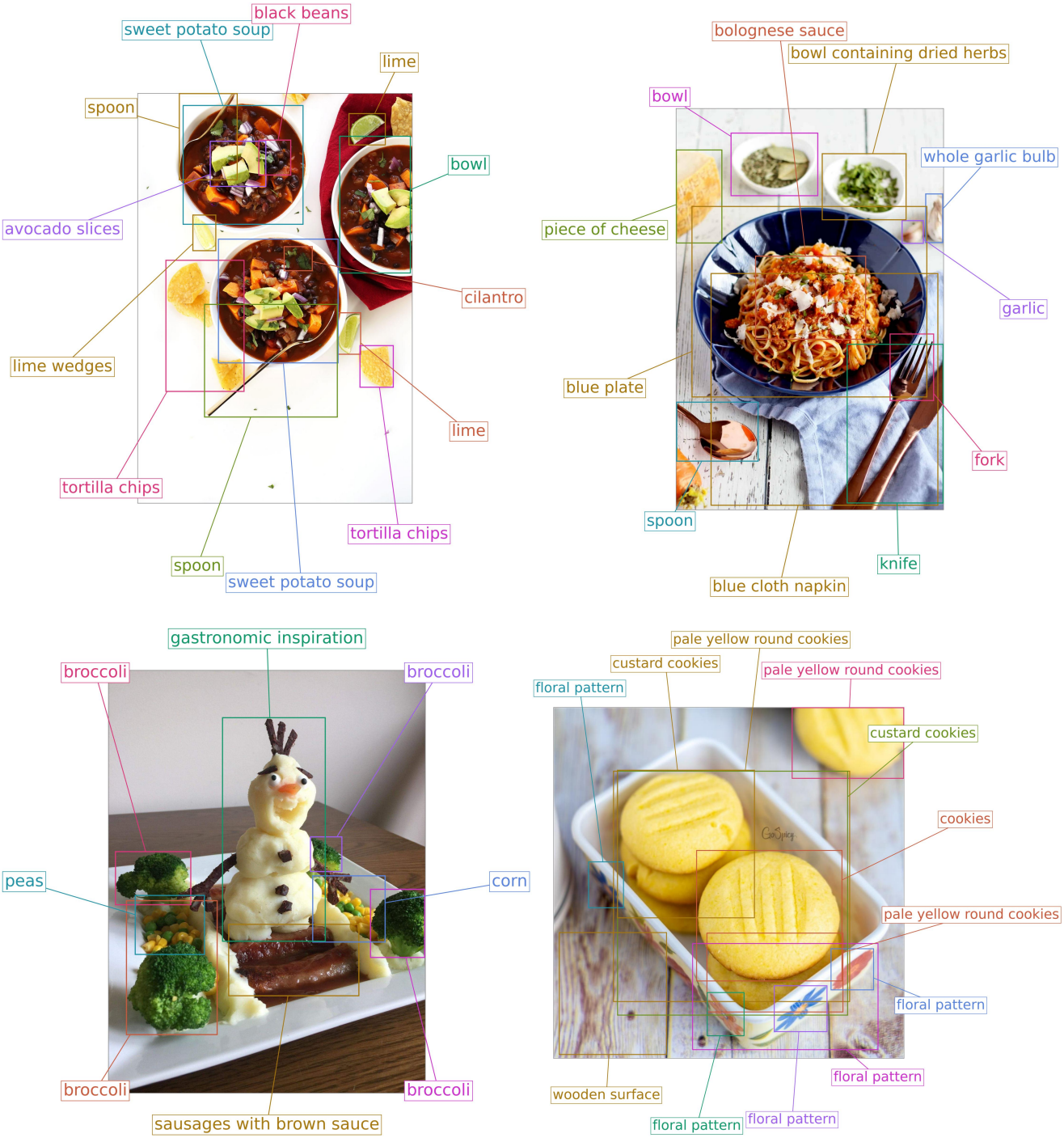
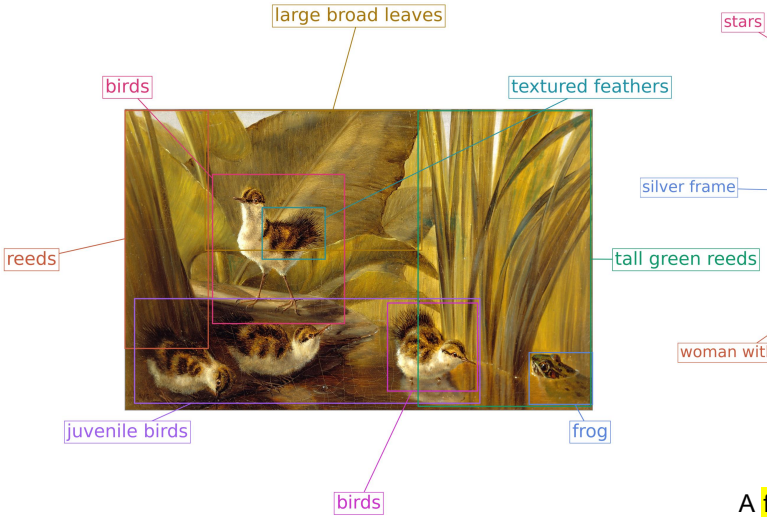
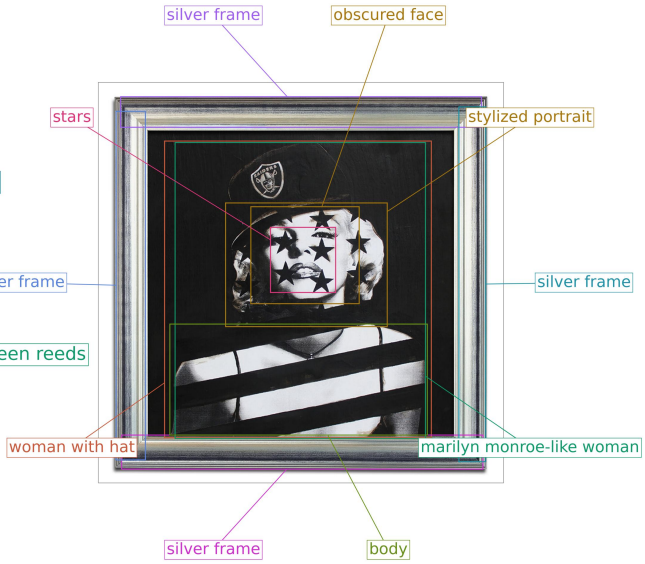


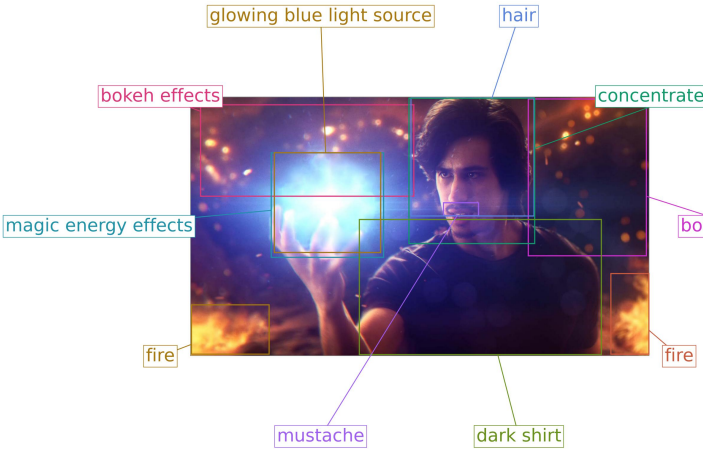
Figure 8. Food categories present particular complexity in object-centric detection and constitute a significant proportion of social media images. Traditional detection methods encounter challenges with food imagery, often producing outputs that are either excessively dense or too sparse for effective generative model training. Our open-vocabulary framework demonstrates benefits through specific categorization such as *whole garlic bulb* in the upper right and *sausage with brown sauce* in the lower left. These examples illustrate how detailed food categorization captures culinary specifics that generic detection approaches typically miss, providing more precise semantic understanding for food-related content.



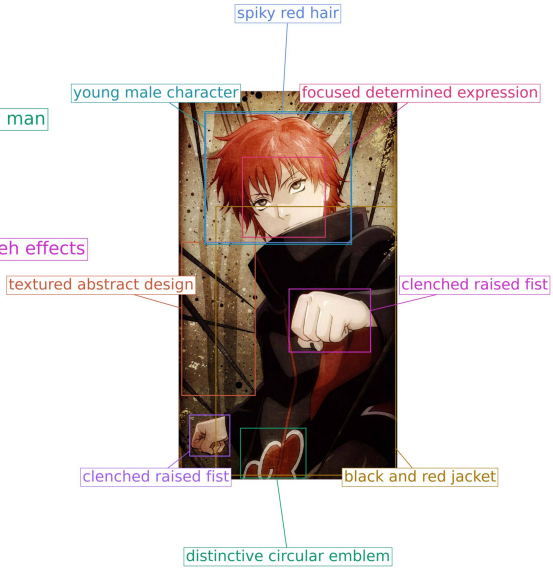
A realistic oil painting depicting a natural scene ...



A framed artwork featuring a stylized black and white portrait of a woman with blonde hair and a hat, reminiscent of Marilyn Monroe ...



A 3D render of a man with dark hair and a mustache ...



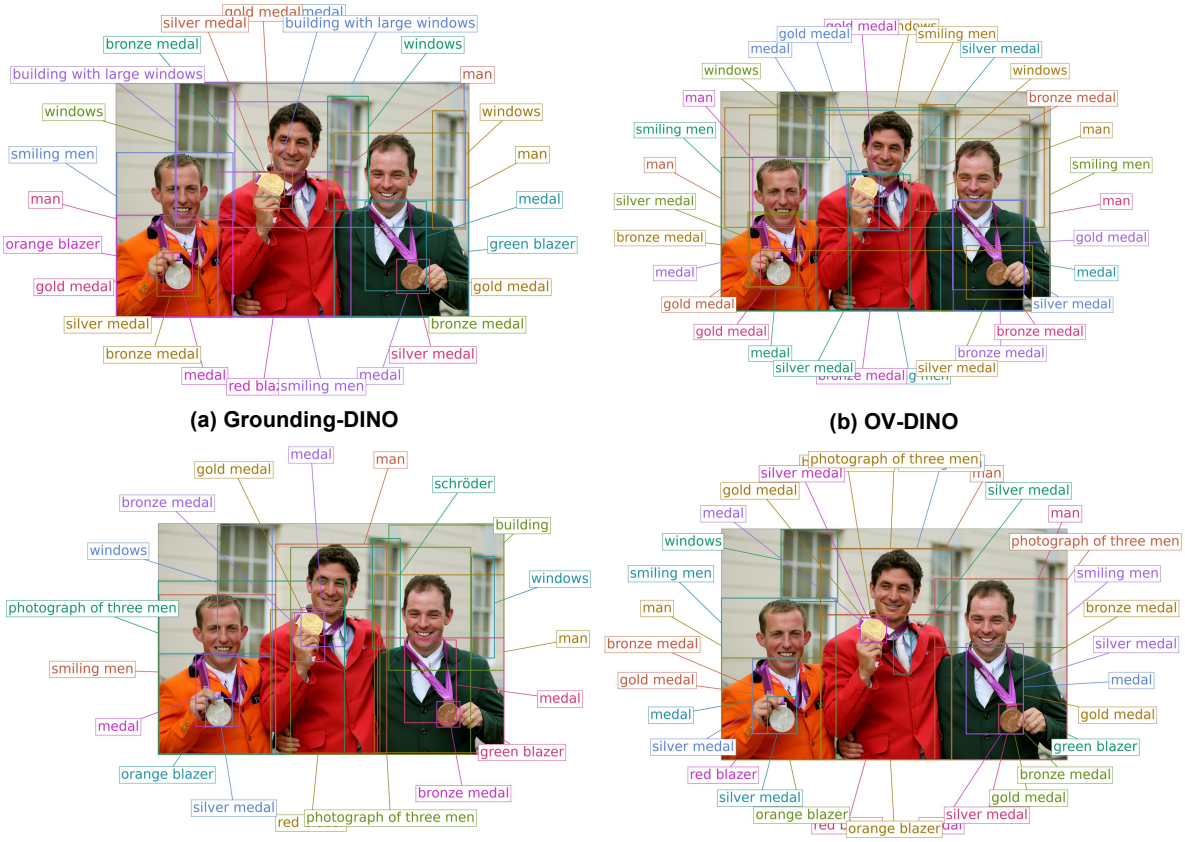
A stylized illustration of a young male character with spiky red hair ...

Figure 9. Our dataset encompasses stylized images from diverse internet sources, where detailed VLM descriptions demonstrate particular advantages. The examples show successful detection across various artistic styles, including oil paintings, black-and-white comics, 3D renderings, and anime. Each case captures primary objects while identifying focal points within different aesthetic contexts. The VLM's ability to understand and summarize artistic styles proves valuable for training generative models across diverse visual aesthetics, maintaining annotation quality regardless of stylistic variations from photorealistic content.

C. Aims of Resampling

This section discusses the resampling stage mentioned in the paper's main text. Referring to Fig. 10, we first list the results of VLM description and LLM summarization for the categories to be detected; relevant words or phrases caught by OVDs [2, 5, 6, 8] are highlighted in the text. Things evident are:

- Boxes from OVDs are pretty dense and exhibit significant overlap, which may introduce considerable potential bias. Furthermore, if each box is individually checked by the VLM, it would significantly increase the computational overhead.
- Instances of errors are also apparent, such as inaccuracies regarding the materials of gold, silver, and bronze medals, as well as the specific color of blazers.
- Notably, some detectors have incomplete category responses; for example, OV-DINO [8] failed to detect any categories related to blazers in this case.
- Additionally, there are isolated detection results, such as OWLv2 [6] responding to *schröder*, a category derived from a web caption that falls under NER (Named Entity Recognition). Given the actual capabilities of OVDs, this category is not exceptionally reliable when three individuals are present.



WEB CAP. London OS 120807 Gerco Schröder silvermedal, Steve Guerdat goldmedal and Cian O'Connor bronzmedal. Photo: Roland Thunholm Code: 718 35 . OS i London 2012
VLM DESC. A photograph of three men standing side by side, each holding a medal. The man in the center is wearing a red blazer and holding a gold medal, while the man on the left is in an orange blazer with a silver medal, and the man on the right is in a green blazer with a bronze medal. They are all smiling and appear to be at an outdoor event, with a building featuring large windows in the background.

Figure 10. Effect of OVDs after the pre-detection VLM-LLM re-captioning.

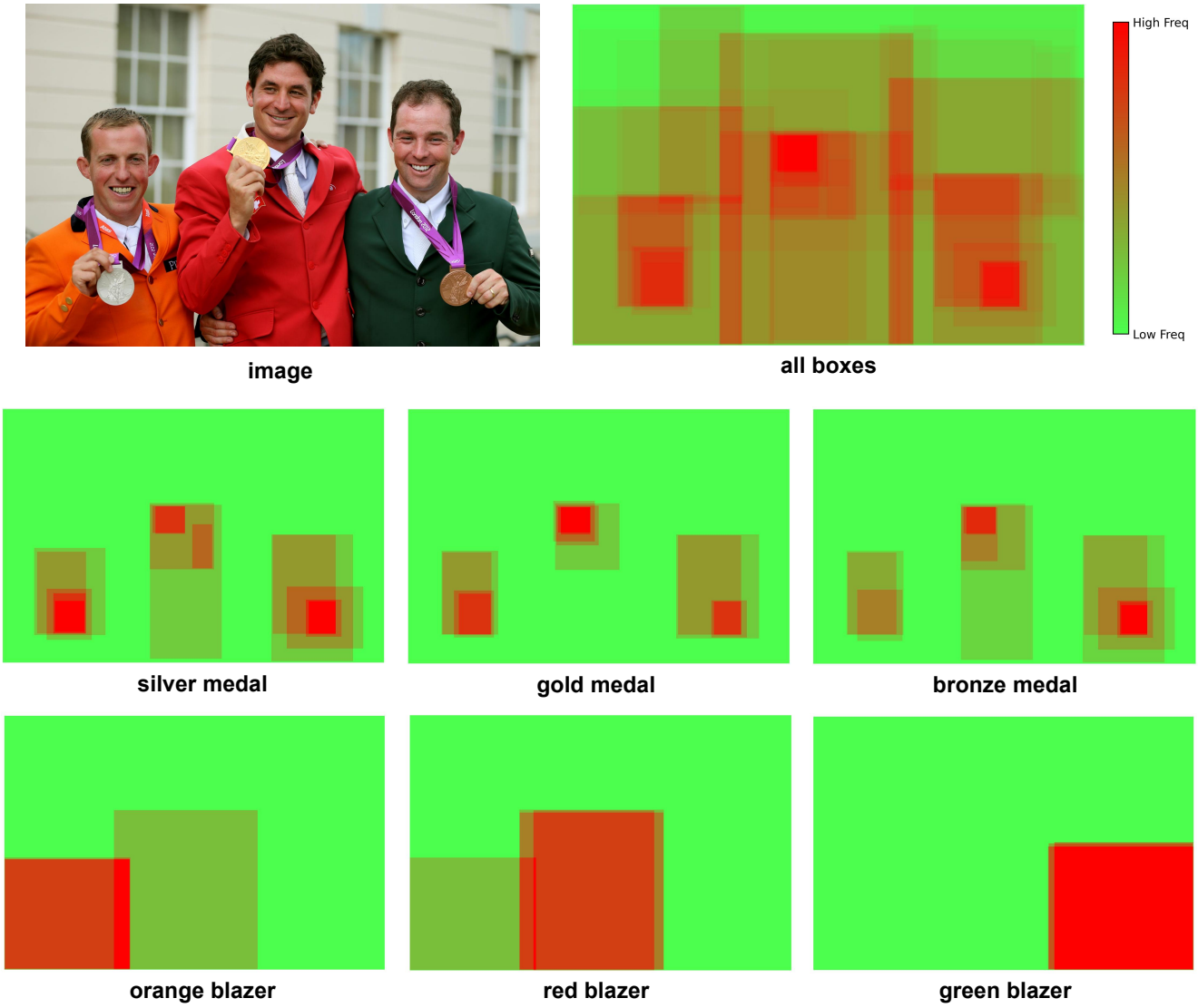


Figure 11. Heatmaps on categories of concern.

Here, we discuss how to integrate the detection results of OVDs from a fundamental perspective, specifically by examining the overlap relationships between bounding boxes. According to the frequency relationships revealed in Fig. 12, the area containing the three medals exhibits the highest frequency. This observation also explains why we retain basic categories such as *medal* and *blazer* along with their open-vocabulary category combinations, as both reflect the visual observation capabilities of OVDs.

We focus on the category combinations related to *medal* and *blazer*, as these are the most prone to errors. The heatmap indicates that we can determine priority based on overlap relationships by synthesizing the detection results from multiple OVDs. This results in the correct regions for these six groups of open-vocabulary categories achieving the highest frequency.

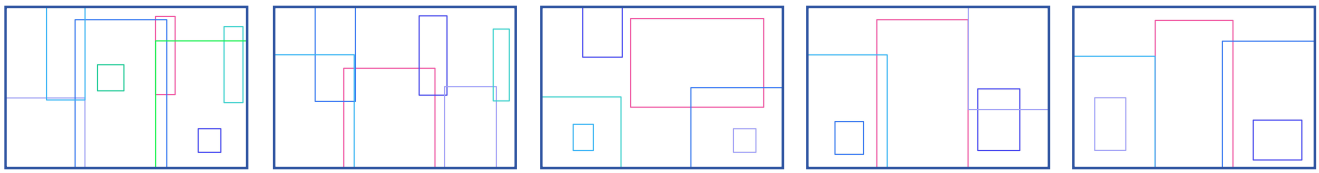


Figure 12. Box candidates in layers of resampling in relatively low overlap.

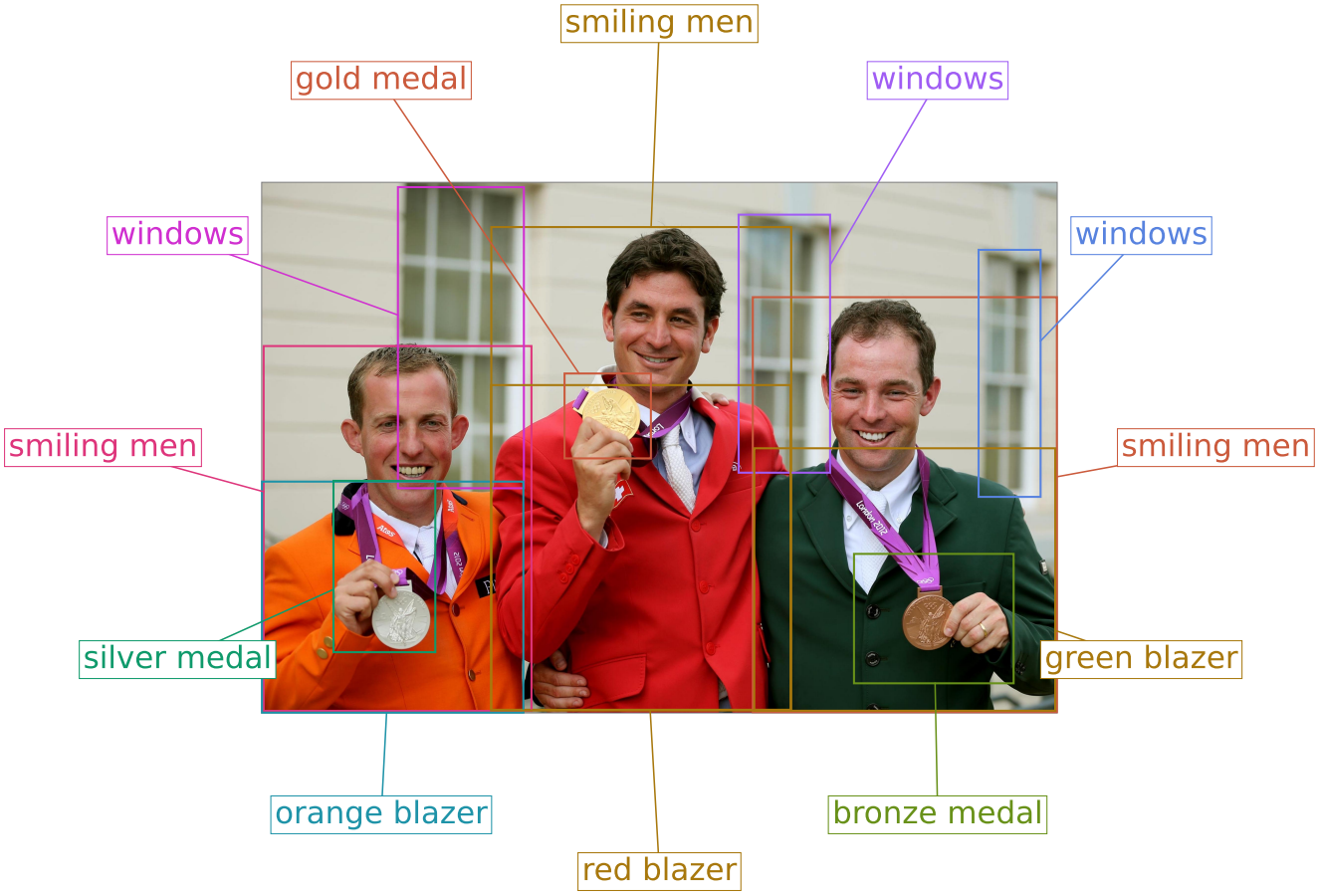


Figure 13. Effect of a limited number of resampled and cross-checked boxes.

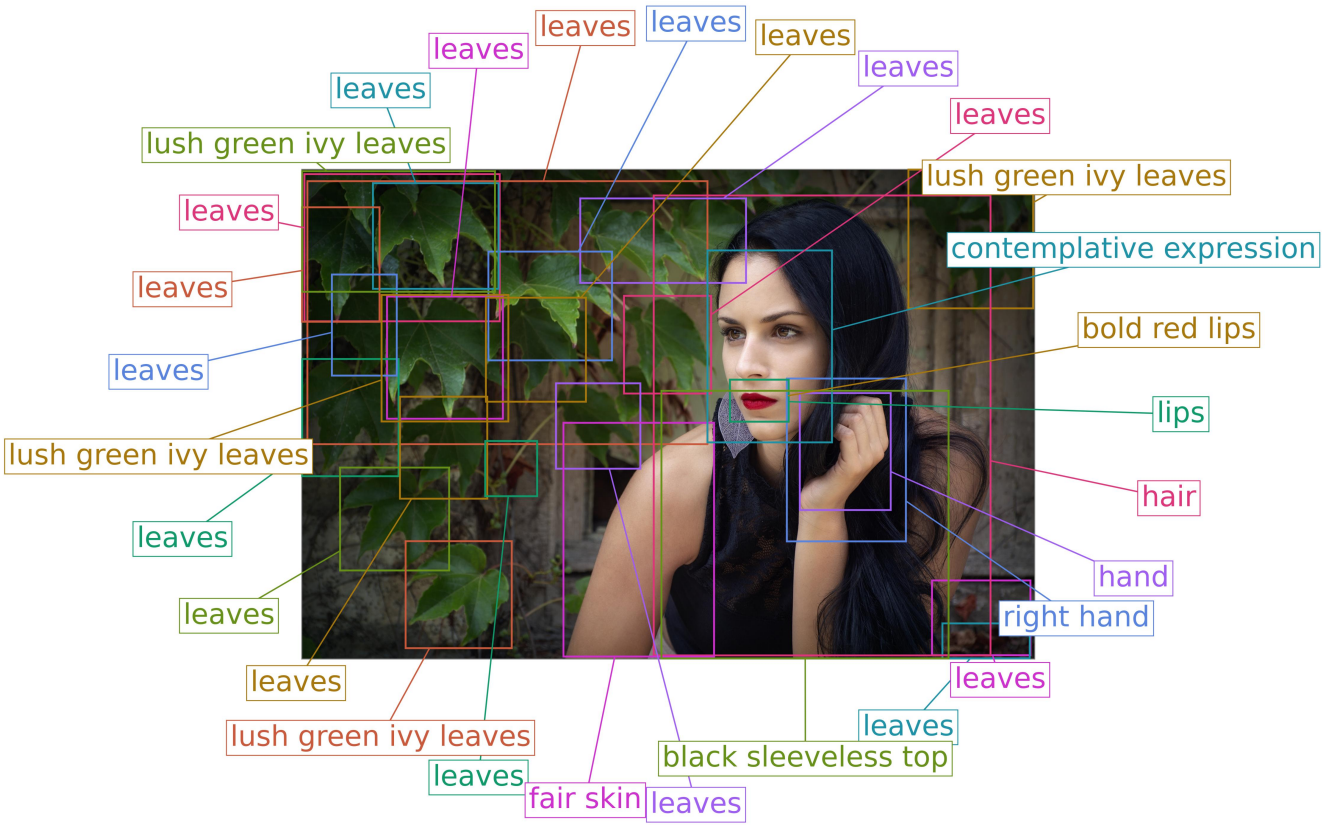
Regarding the regional overlap that is almost inevitable with OVDs, we identify a crucial factor: even in images with complex hierarchical relationships, the number of bounding boxes that can be effectively stacked for description remains limited. Therefore, when selecting candidate boxes for VLM inspection, we set our sampling target to five layers with lower overlap. The probability of sampling each instance is computed by penalizing several factors: previously sampled boxes, already sampled captions, distance from the image center, and small box sizes. Details of implementation could be found in the codes coming.

Fig. 12 illustrates the sampled layers, while Fig. 13 encompasses the primary bounding boxes that successfully passed cross-checking by the VLM. This selection is sufficient to provide a comprehensive summary of the image. In this example, the original results from four OVDs yielded a total of 107 bounding boxes; however, only 30 boxes were sent for VLM cross-checking after resampling, with 21 ultimately passing. This approach significantly reduces potential information leakage while considerably conserving computational resources.

D. Limitations in ROVI's Annotations

This section explores some seemingly questionable features and limitations within ROVI based on several typical examples. Based on Fig. 14, the following points can be analyzed:

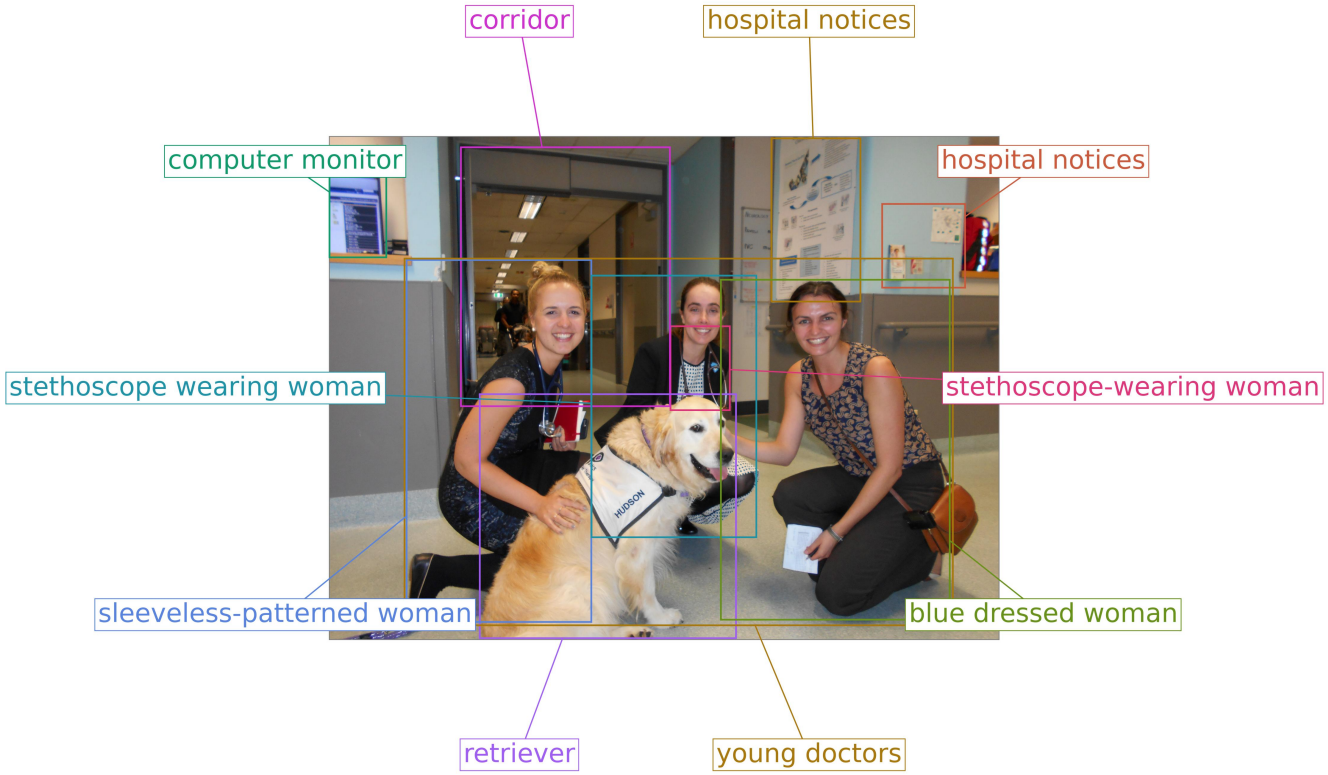
- As demonstrated by *leaves* or *lush green ivy leaves*, we did not address singular and plural forms; instead, we generally maintained their original appearance in the VLM descriptions. This decision is based on several considerations: first, due to the inherent limitations of OVDs, if both "leaves" and "leaf" were treated as detection targets, the detector would struggle to differentiate between them. Second, our objective is to assist in text-to-image generation, where the text encoder, serving as a foundational module, can easily bridge the differences between singular and plural forms. Consequently, the training outcomes for text-to-image generation can broadly be shared across singular and plural forms based on the capabilities of the text encoder. As for inference, a more practical approach for users is to input multiple boxes in the singular form.
- In the context of open-vocabulary, overlapping bounding boxes of different categories are almost inevitable. Sec. C previously mentioned the potential misdetection issues related to medals of different materials; based on similar considerations, even when detecting "lush green ivy leaves," the original form "leaves" is not discarded.
- The overlap among bounding boxes of the same category can be somewhat mitigated using techniques such as Non-Maximum Suppression (NMS), although we still retain some overlaps. This is particularly true for combinations of categories that exhibit weaker responses and present detection challenges, where it can be difficult to determine which bounding box is more appropriate or whether to include as many similar objects as possible within a single box.



WEB CAP. Photo in She | Author Петя Лазарова - petial | PHOTO FORUM
VLM DESC. A photograph of a woman with long, dark hair and fair skin, wearing a black sleeveless top with a lace-like pattern. She is positioned in front of a backdrop of lush green ivy leaves, and her right hand is gently touching her hair. The woman is looking off to the side with a contemplative expression, and her lips are painted a bold red color.

Figure 14. Effect of dense open-vocabulary detection based on VLM-LLM re-captioning.

- *right hand* is an error in the VLM description. We found that the VLM does not recognize a person's mirrored relationship within a photograph; thus, positional judgments such as "right" are primarily based on the object's location within the image. Fortunately, such errors are quite rare and have minimal impact.



WEB CAP. Huddie and young docs
VLM DESC. A photograph taken inside a hospital corridor. Three smiling women are crouching down to interact with a golden retriever wearing a vest labeled "HUDSON." The woman on the left is wearing a blue dress and a stethoscope around her neck. The woman in the middle is wearing a black jacket and glasses, and the woman on the right is wearing a sleeveless top with a pattern. Behind them, there are various hospital-related notices on the walls and a computer monitor on a desk.

Figure 15. LLM summarized questionable open-vocabulary categories.

Fig. 15 illustrates instances where the LLM generates some unreasonable open-vocabulary categories, primarily related to human attire. Among these, *stethoscope-wearing woman*, *sleeveless-patterned woman*, and *blue dressed woman* are not particularly rational categories.

Our experiments revealed that the LLM's tendency to synthesize categories across discontinuous text is generally beneficial; for example, *bold red lips* in Fig. 14 demonstrates this advantage. Although the category *blue dressed woman* deviates somewhat from typical human conventions, it can still be utilized to some extent after processing through the text encoder. After weighing these factors, we accepted the LLM's inductive tendencies.

Another issue pertains to the misidentification of human identities. The example in the figure is derived from the web caption *young doctor*. However, due to a lack of distinct visual elements, we cannot be certain whether the three women belong to this category (the woman wearing a stethoscope may be more closely aligned). Similar issues may arise with NER categories; for instance, in Fig. 10, OWLv2's detection results include *Schröder*, but we cannot confirm which of the three awardees this individual represents.

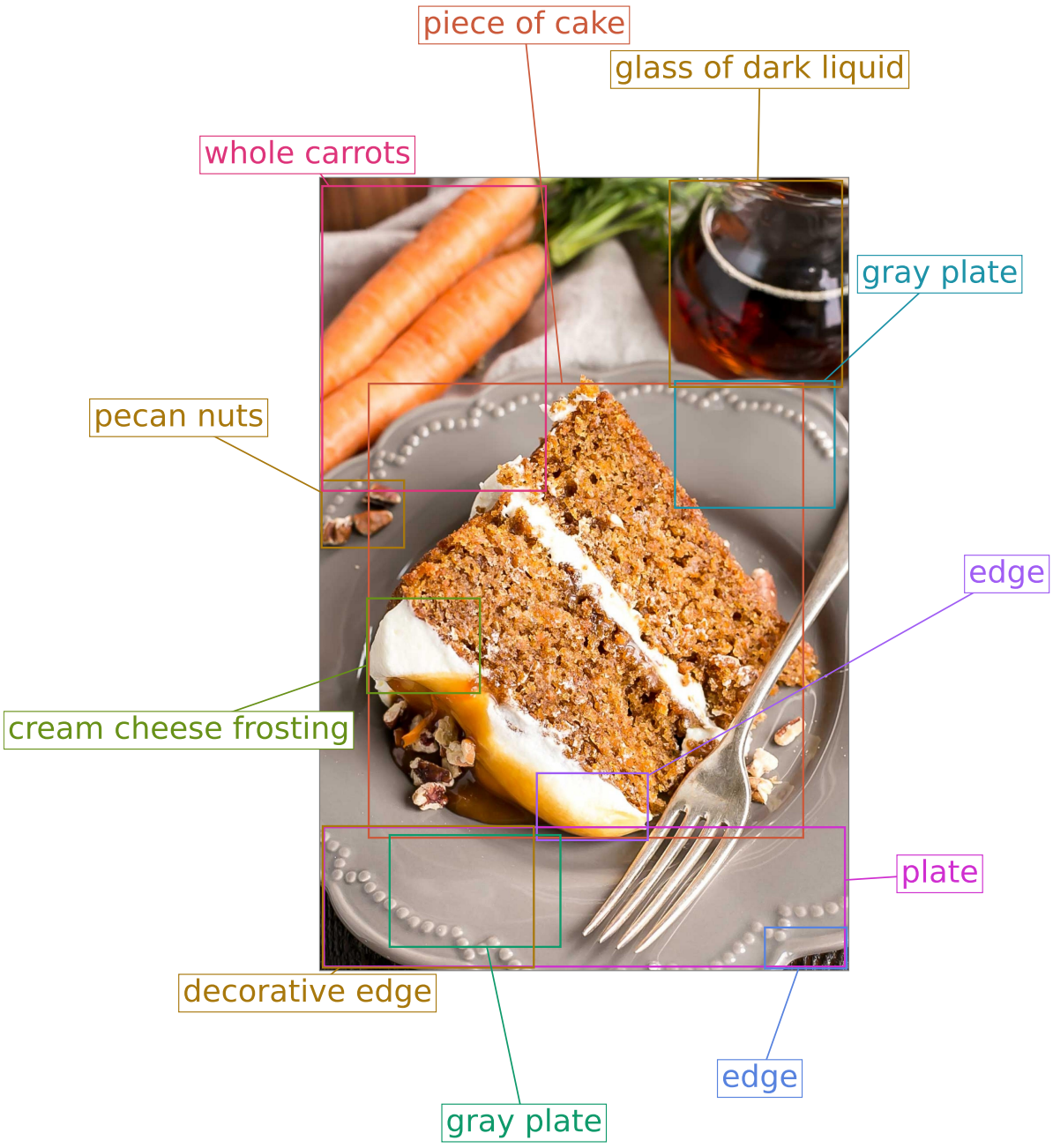


Figure 16. Discontinuous bounding box segments of *edge* and *gray plate*.

Fig. 16 illustrates the issue of discontinuous bounding boxes. Due to visual occlusion, the *plate* and *edge* in this example are not continuous, ultimately resulting in the detection results being fragmented into undesirable small segments. This problem arises from the tendencies inherited from OVDs.

Different OVDs may exhibit varying inclinations when faced with non-contiguous objects; they might either segment according to precise boundaries or attempt to summarize them into a cohesive whole that spans invisible portions, and sometimes they may exhibit both tendencies. The former can lead to overly fragmented bounding boxes, commonly seen in examples such as mountains, walls, water, flooring, etc. Conversely, the latter may result in overestimating size, as is often the case with items like a shirt worn underneath a coat.

Our approach is to leverage a resampling strategy that focuses on overlapping relationships to enhance candidate boxes' quality by utilizing knowledge from multiple OVDs. However, the results still fall short of expectations in a few instances.

E. Failure Cases in Generation

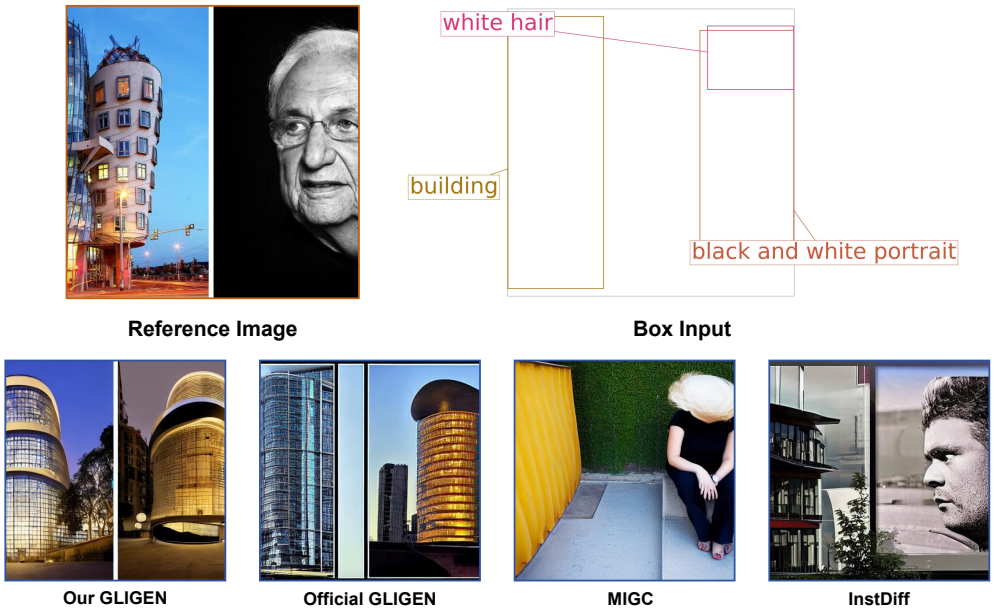


Figure 17. In the image above, the left side of the bounding box input represents a building, while the right side depicts a person, resembling a collage with two independent sections. Due to being trained on global VLM descriptions, our model places a strong emphasis on coherence, making it less likely to generate results that appear disjointed and lack natural transitions. Comparative results indicate that official GLIGEN is similar to our model, whereas MIGC and InstanceDiffusion successfully generated mutually independent objects. This observation reflects the latter two models' focus on independent instance grounding.

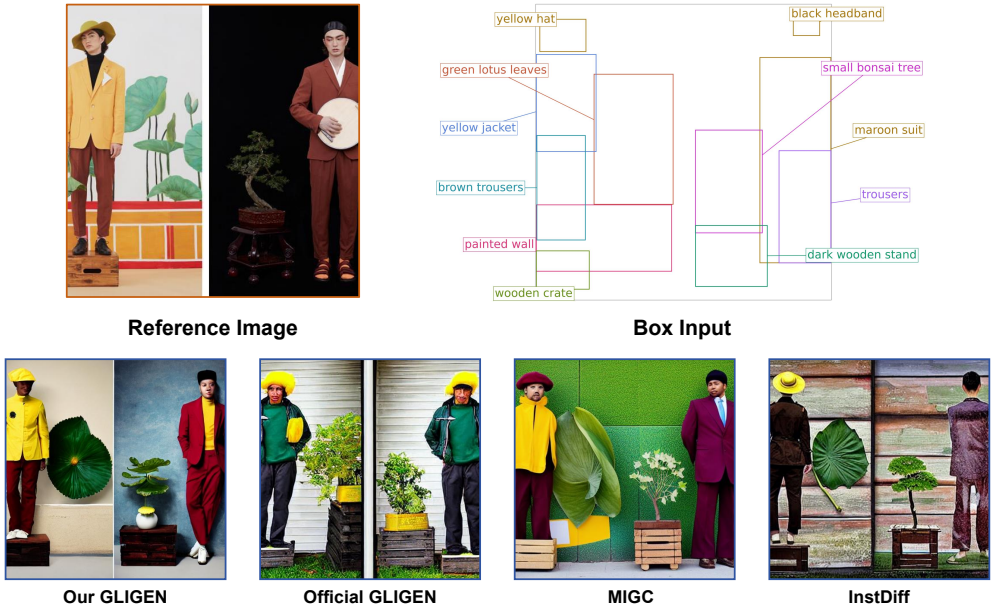
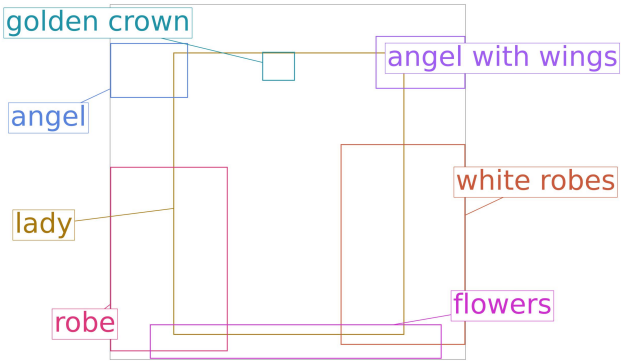
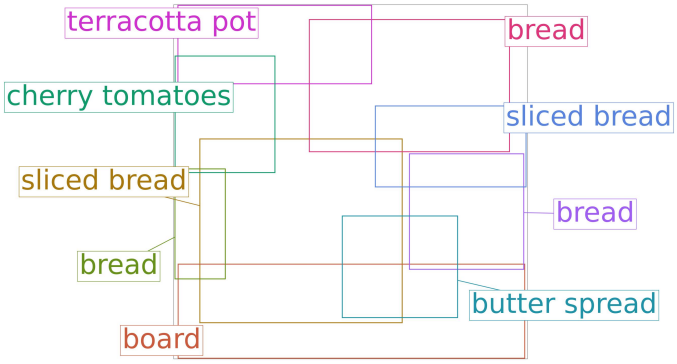


Figure 18. The category *small bonsai tree* is a low-frequency open-vocabulary category within our dataset, which may hinder the potentials of corresponding generative capabilities. In this example, the official GLIGEN's generation results for "bonsai" may be closer to expectations. However, the attribute binding for our other composite categories is noticeably superior, particularly in terms of color.



Reference Image | Box Input

Our GLIGEN | Official GLIGEN | MIGC | InstDiff



Reference Image | Box Input

Our GLIGEN | Official GLIGEN | MIGC | InstDiff

Figure 19. In this set of examples, the *angel* in the upper group and the *butter spread* in the lower group were not generated correctly. This is likely attributable to the relatively low-frequency visual associations present in the training dataset. For instance, all models in the upper group generated the bounding box labeled *angel* as a bird, which may reflect a common trend observed in the network images from the training set. In the lower group, both MIGC and InstanceDiffusion produced independent instances of *butter* that were unrelated to *spread*, implying a trend focus on independent instance grounding, similar to that seen in Fig. 17.

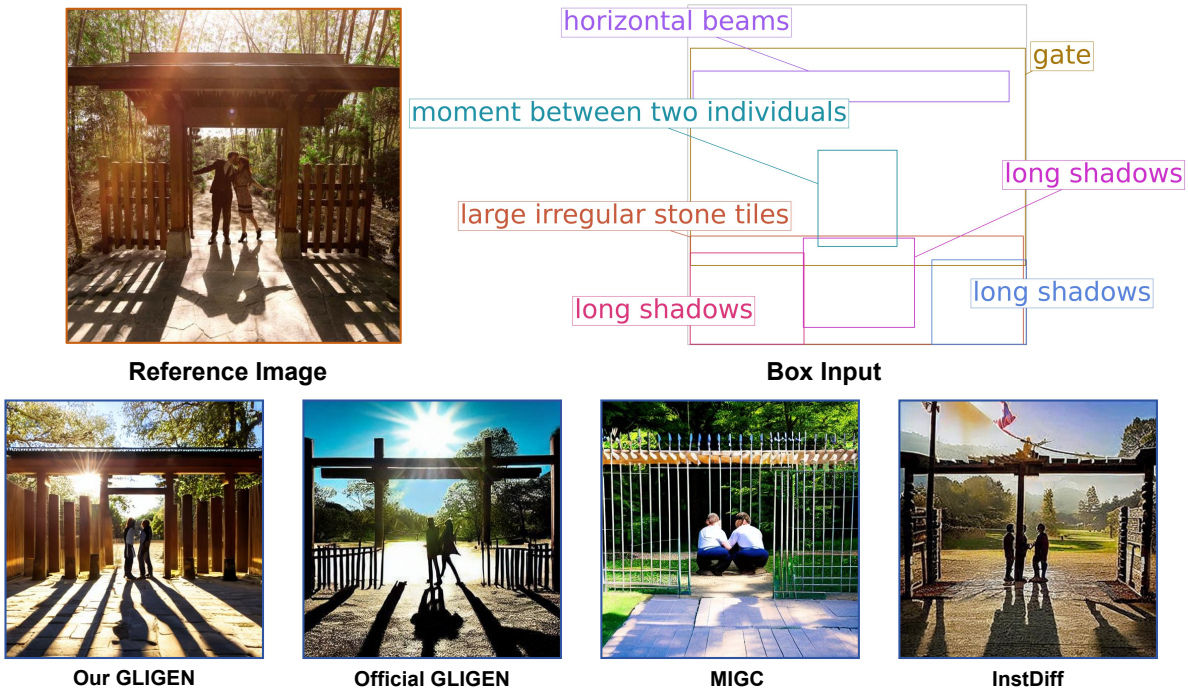


Figure 20. This set of examples illustrates several visual elements that extend beyond the spatial representation capabilities of bounding boxes. In the upper group, *shadows* are clearly better represented through segmentation masks, while merely providing bounding boxes significantly challenges the logical capabilities of the base model itself. In the lower group, the examples related to the *ski lift system* resulted in a complete failure of generation, indicating that isolated bounding boxes alone are insufficient to guide the model in capturing the suspended state of the cable cars along continuous tracks.

F. More Information about User Study

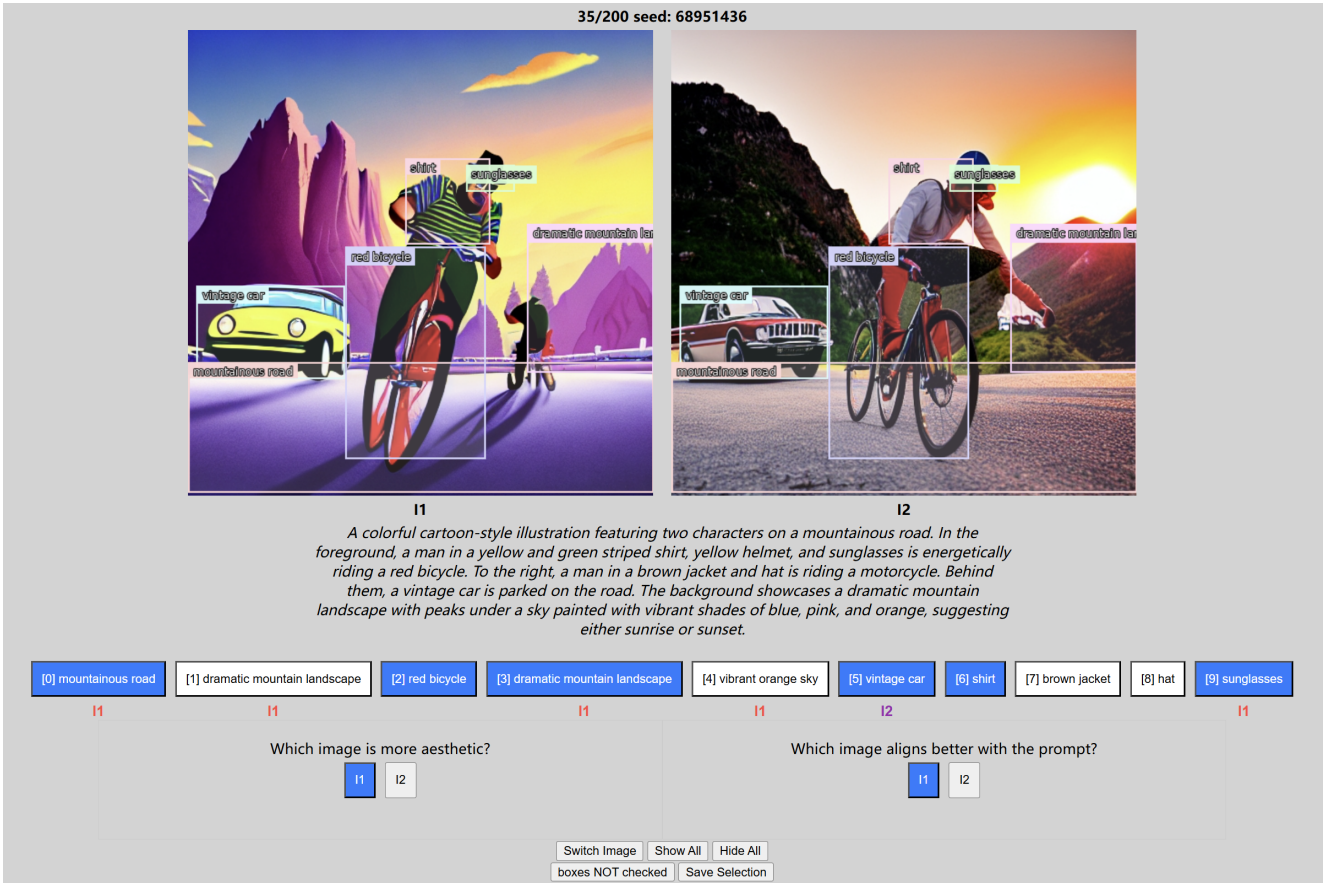


Figure 21. A screenshot of our user study interface.

Fig. 21 presents a screenshot of our evaluation interface, which shows a randomized pair of generated images alongside the input prompt and labeled bounding boxes. Participants were asked to choose a winner among the generated pairs based on: (1) aesthetic quality and (2) alignment with the provided prompt, without knowing the source generator. Optionally, participants could select a better instance alignment for each individual box-label pair. Users could toggle the bounding box overlays to examine occluded content. Most operations could be performed using either buttons or hotkeys.

Our initial assessment revealed that evaluating box-label alignment precision and complex prompt adherence (derived from VLM descriptions) was cognitively demanding, due to the large amount of text and boxes that have to be processed for each image. Considering the effort required and to prevent potential VLM misuse during evaluation, we prioritized evaluator expertise and trustworthiness. Due to the voluntary nature of this study, we recruited a group of colleagues experienced in image generation whom we could trust to complete the evaluations without resorting to LLM/VLM assistance.

Prompt alignment and aesthetic quality were assessed at the image level, while instance alignment was evaluated at the object level—comparing each box-label pair with its corresponding region in the generated images. This approach allowed us to measure both overall image coherence and the precision of individual object grounding within the synthesized outputs.

References

[1] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1

[2] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 11

[3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[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. 2023. 1, 2

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 11

[6] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 11

[7] 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. 1

[8] Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, et al. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844*, 2024. 11

[9] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation, 2024. 1, 2

[10] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis, 2024. 1, 3