Alpha-CLIP: A CLIP Model Focusing on Wherever You Want

Zeyi Sun*[1,4], Ye Fang*[2,4], Tong Wu[3], Pan Zhang[4], Yuhang Zang[4],
Shu Kong[5], Yuanjun Xiong[6], Dahua Lin[3,4], Jiaqi Wang†[4]
[1]Shanghai Jiao Tong University [2]Fudan University [3]The Chinese University of Hong Kong
[4]Shanghai AI Laboratory [5]University of Macau [6]MThreads, Inc.

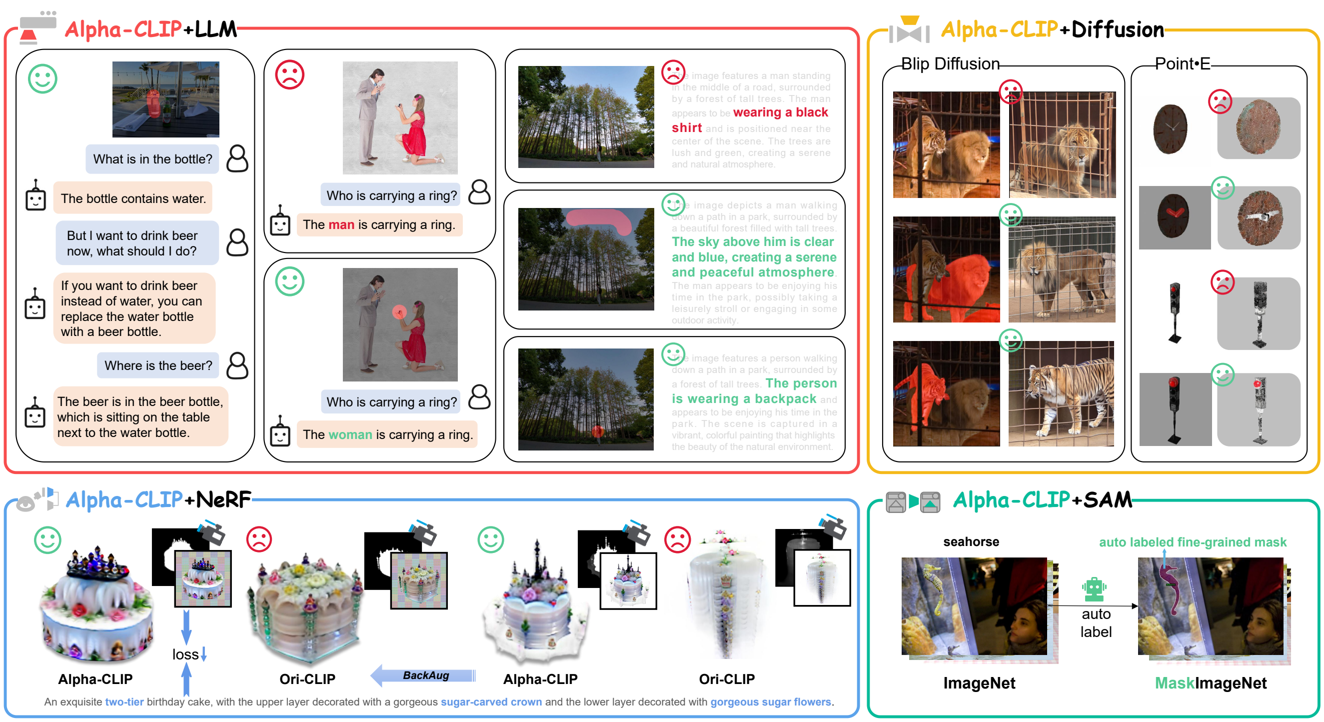

Figure 1. **Usage of our proposed Alpha-CLIP**. Our Alpha-CLIP can seamlessly replace the original CLIP in a wide range of tasks to allow the whole system to focus on any specified region given by points, strokes or masks. Cases marked with ☹ are generated with the original CLIP. Cases marked with ☺ are generated with our Alpha-CLIP. All cases shown here are made simply by replacing the original CLIP of the system with a plug-in Alpha-CLIP without further tuning.

Abstract

Contrastive Language-Image Pre-training (CLIP) plays an essential role in extracting valuable content information from images across diverse tasks. It aligns textual and visual modalities to comprehend the entire image, including all the details, even those irrelevant to specific tasks. However, for a finer understanding and controlled editing of images, it becomes crucial to focus on specific regions of interest, which can be indicated as points, masks, or boxes by humans or perception models. To fulfill the requirements, we introduce Alpha-CLIP, an enhanced version of CLIP with an auxiliary alpha channel to suggest attentive regions and fine-tuned with constructed millions of RGBA region-text pairs. Alpha-CLIP not only preserves the visual recognition ability of CLIP but also enables precise control over the emphasis of image contents. It demonstrates effectiveness in various tasks, including but not limited to open-world recognition, multimodal large language models, and conditional 2D / 3D generation. It has a strong potential to serve as a versatile tool for image-related tasks. Our project is with codes and models available is linked to https://aleafy.github.io/alpha-clip/.

* Equal contribution. † Corresponding author.

Domains	Components	Tasks	Methods	Advantages over the original CLIP
Image Recognition	Alpha-CLIP	Zero-shot Classification Zero-shot REC	-	Superior classification accuracy Excellent region-text comprehension ability
	Alpha-CLIP + SAM	Data Engine for OVD	Detic [71]	Higher OVD mAP
MLLM	Alpha-CLIP + LLM	VQA, Captioning	BLIP-2 [25], LLaVA-1.5 [29]	Region-focused captioning / VQA Eliminating hallucinations Reducing model bias
2D Generation	Alpha-CLIP + Diffusion	Image Variation	BLIP-Diffusion [24]	Controllable generation Enabling subject-driven generation in complex images
3D Generation	Alpha-CLIP + Diffusion	Generalized Image-to-3D	Point-E [34]	Rectifying absent parts
	Alpha-CLIP + NeRF	Optimized Image-to-3D	PureCLIPNeRF [22]	Improved 3D optimization results

Table 1. **Downstream tasks of Alpha-CLIP and their advantages over the original CLIP**

1. Introduction

Recent advances in Contrastive Language-Image Pre-training (CLIP) [16, 38] and its diverse variants [8, 27, 50] have established a robust framework for extracting semantically coherent features from both images and texts. These features aim to capture all the semantic details within images, exhibiting potent representation capabilities and exceptional generalizability, making them versatile in a variety of downstream tasks, such as open-world recognition [5, 11, 57, 59, 60], Multimodal Large Language Models (MLLMs) [4, 15, 23, 25, 29, 30, 36, 51, 66], and 2D / 3D generation [17, 22, 24, 33, 34, 39, 63].

While CLIP captures the content of the entire image, it is also crucial to focus on the regions of interest to enable a finer understanding [14, 18, 21, 37, 48, 72] and controllable content generation [22, 33, 46, 55]. These regions can be specified by points, masks, or boxes via human interaction or perception models (e.g., SAM [19], GLIP [26] and proposal networks [65]).

To fulfill the demands of downstream tasks, researchers have attempted to acquire region-focused CLIP features using two primary strategies. The first method is to exclude non-relevant areas by cropping the regions of interest into distinct patches [5, 49, 68, 69] or applying masking to the irrelevant parts of images [28], features [28, 55], and attention masks [60, 70]. However, this approach disrupts (in cropping) and omits (in masking) contextual information, which is crucial for precise image understanding and reasoning. The second method is to highlight the regions of interest by circles [47] or mask contour [61] on the images fed to CLIP. Although user-friendly, it changes the original content of the images, which will result in undesirable recognition and generation results (cf. Fig. 2).

To achieve region focus without hurting original image, we propose Alpha-CLIP, which improves CLIP [38] by incorporating regions of interest through an additional alpha channel input. Along with the RGB channels, the introduced alpha channel enables the Alpha-CLIP to focus on designated areas while maintaining an awareness of the contextual information. While initialized with the CLIP [38] model, the training of Alpha-CLIP still requires a large set of region-text paired data. By harnessing the Segment Anything Model (SAM) [19] and multimodal large models for image captioning, such as BLIP-2 [25], we develop an effective pipeline to generate millions of region-text pairs that are readily convertible to RGBA-text data. After training with a mixture of region-text pairs and image-text pairs, Alpha-CLIP can focus on the specific regions while maintaining the visual recognition accuracy of CLIP [38].

Alpha-CLIP can enhance CLIP across a wide array of downstream tasks, applying a plug-and-play methodology that permeates diverse domains, spanning from perception to generation in 2D and 3D applications, as shown in Fig. 1 and Tab. 1. Specifically, **1) Image Recognition**: Alpha-CLIP not only maintains the visual recognition ability of the original CLIP but also boosts the capability of region-based recognition. Specifically, when provided with ground-truth region to focus on, Alpha-CLIP achieves 4.1% improvement in top-1 accuracy on zero-shot ImageNet classification task. This superior region-based recognition ability helps downstream tasks like Referring Expression Comprehension(REC) [49] or serves as data engine for Open Vocabulary Detection(OVD) [71]. **2) Serving as vision backbone for MLLM**: In conjunction with a large language model, Alpha-CLIP becomes capable of facilitating region level captioning and VQA within a MLLM framework. This integration significantly mitigates the occurrences of hallucinations (*e.g.*, black shirt) and diminishes model bias (*e.g.*, man carrying a ring). **3) 2D generation**: When integrated with a diffusion model, Alpha-CLIP enhances the controllability of BLIP-Diffusion [24] in image variation tasks. In addition, it enables the extraction of subjects from complex images for subject-driven generation, surmounting an obstacle encountered when deploying BLIP-Diffusion with the original CLIP, which only supports single subjects in simplistic images. **4) 3D generation**: In addition to the capabilities in 2D generation, Alpha-CLIP exhibits proficiency in 3D generation as well. It can be effectively deployed in conjunction with a diffusion model, such as Point-E [34], to enhance the

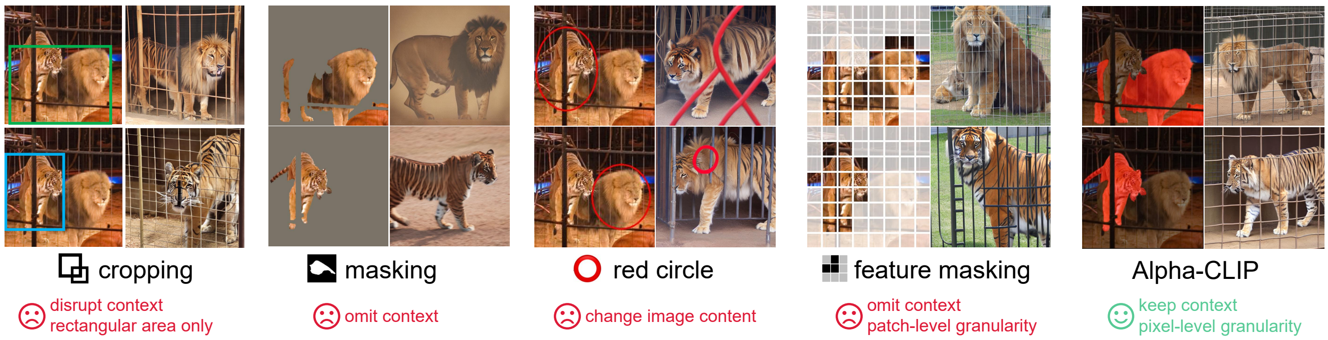

Figure 2. **Alpha-CLIP vs. other methods of region-focusing** for image generation using BLIP-Diffusion [24]. The fine-grained region focusing ability of Alpha-CLIP produces better results than these methods that adopt the original CLIP.

quality of 3D object generation. Additionally, it can be utilized with NeRF [32], exemplified by PureCLIPNeRF [22], to optimize the creation of superior 3D objects.

In summary, we propose Alpha-CLIP, which equips the original CLIP model with the capability of region awareness. Through fine-tuning on millions of RGBA region-text pairs, Alpha-CLIP demonstrates significant advantages over the original CLIP across various tasks, including but not limited to image recognition [38, 49, 71], multimodal large language models [25, 30], 2D generation [24, 39] and 3D generation [22, 34].

2. Related Work

Empowering CLIP with region awareness. To enable CLIP [38] to disentangle regions from the whole image for more targeted processing and understanding, various methods have been explored in the field of *segmentation*. Among them, MaskCLIP [70] uses a 1x1 convolution layer to extract CLIP's final 2D features to obtain semantic information for different regions. SAN [59] trains a side network alongside CLIP to assist the model in local semantic perception. MaskCLIP [7] and ODISE [57] use attention masks to make CLIP focus more on local regions. These methods do not alter the weights of the CLIP model itself. RegionCLIP [69] generate region box-text pairs for local region and fine-tune CLIP model for box level recognition. MaskAdaptedCLIP[28] generates mask-text pairs for local masks through a pseudo-labeling process and fine-tunes the CLIP model to make it more adaptable to masked images. MaskQCLIP[60] fine-tunes attention layer for new mask `[CLS]` tokens to make it more fit for mask object classification. These two methods attempt to enhance CLIP's ability to focus on local features and exclusively fine-tune CLIP on specific downstream datasets, resulting in poor generalization ability beyond detection or segmentation tasks.

Another approach is to *change the input image* by simply cropping or masking the image to leave only the foreground object. ReCLIP [49] and OvarNet [5] crop the original image using bounding box from object proposal network [65] and are applied on Referring Expression Comprehension and Open Attribute Recognition tasks. MaskAdaptedCLIP [28] sets the background area to pure color in pixel space and uses the masked image as input for open-vocabulary segmentation. However, the valuable context information is lost except for using complex post-process proposed in ReCLIP [49]. Some other approaches prompt the CLIP by modifying the input image, guiding CLIP to focus on the area of interest. For example, Red-Circle [47], FGVP [61] use a circle or mask contour to tell CLIP where to focus. Overall, the quality of these approaches that *change the original content of input image* is heavily contingent upon the symbols in CLIP's pre-training dataset. Another limitation is directing modification of images causes a domain gap with CLIP pertaining images. Unlike previous approaches that rely on *segmentation* or *changing the input image*, our Alpha-CLIP incorporates an additional alpha channel, which does not change the image content and preserves the generalization performance (cf. Fig. 2).

Region-level image annotation. Existing CLIP models are pretrained on large-scale datasets like LAION-400M [44] and LAION-5B [45], while fine-grained mask-level labels are not available due to high manual labor costs. Recently, Kosmos-2 [36] introduced a pseudo-labeling pipeline that uses the pre-trained GLIP [26] model to automatically generate fine-grained pseudo-labels of region boxes and their associate expressions. By using this pseudo-labeling baseline, Kosmos-2 releases the GRIT dataset and equips multimodal model [15] with local perception capabilities. Similarly, the All-Seeing [54] project also generates fine-grained text labels via the pseudo-labeling pipeline. Meanwhile, the recent SAM [19] model is trained on massive vision modality data with strong zero-shot abilities for downstream tasks like box-to-mask conversion and automatic mask generation. These developments have made it possible to generate pseudo-masks with region captions at a large scale and have opened up the potential for greater adjustments to CLIP for region-level recognition. Therefore, We build upon GRIT[36] and SAM [19] to propose a method for generating RGBA region-text pairs from grounding data.

CLIP in MLLM. At the age of Multi-modal Large Lan-

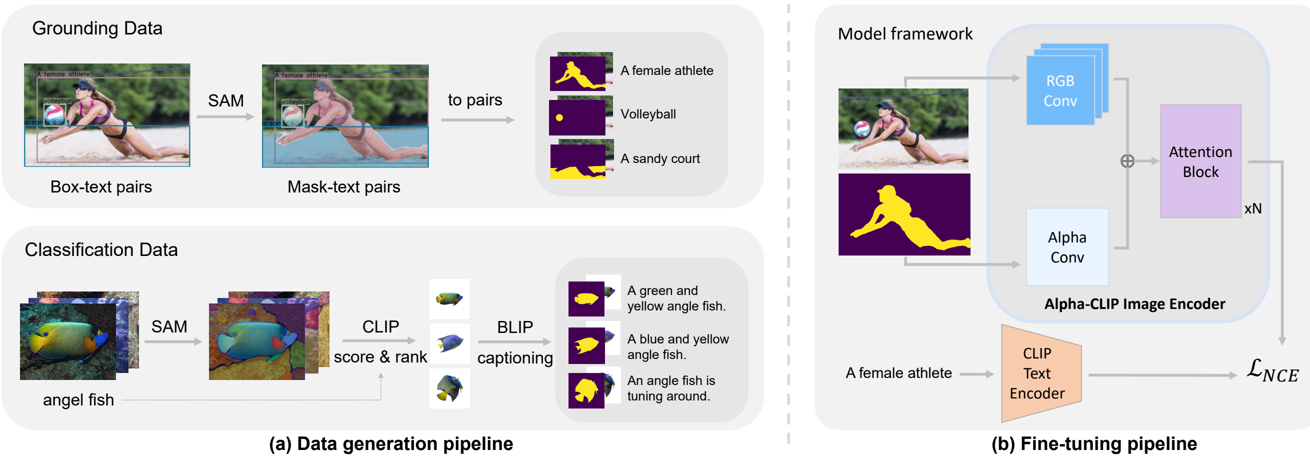

Figure 3. **The pipeline of our data generation method and model architecture**. **(a)** Our method generates millions of RGBA-region text pairs. **(b)** Alpha-CLIP modifies the CLIP image encoder to take an additional alpha channel along with RGB.

guage Models (MLLMs) [1–4, 15, 23, 25, 29, 30, 35, 36, 66], CLIP [38] has been widely used as the vision backbone for its semantic representative feature and promising scalability. To make MLLM focus on the specific region, Kosmos-2 [36] uses millions of region-caption data to train the model with the guidance of box corner points. GPT4ROI [67] propose to apply the ROI Align [13] operator on the CLIP image feature to refer to the specific region. GLaMM [40] further adds an extra region encoder. Different from previous methods that only support box-level focusing and rely on training additional networks, our work achieves more fine-grained mask-level region focusing and merely uses the CLIP model.

CLIP in 2D image variation. CLIP image encoder is widely used in 2D image variation (*e.g*., DALLE-2 [39], Diffusers [53] and IP-Adapter [63]) to achieve better quality or controllability. As for subject-driven image variation pioneered by DreamBooth [42], extraction of pure single object feature from the whole image is more important as the following method ELITE [55] proposes to use feature-level masking to eliminate background information to generate better subjects. Similarly, BLIP-Diffusion [24] uses text to extract the most relevant object features. All these subject-driven image variation methods require the image to have a single foreground object in the center of the image and cannot achieve variation by focusing on user-specified objects in more complex images while maintaining original context information. Such limitations highlight the importance of our Alpha-CLIP that enables subject-driven generation in complex scenes and achieves user-defined region focusing in image variation tasks.

CLIP in 3D generation. Some existing 3D object generation methods involve CLIP [38] model. In diffusion based 3D generation, Point-E [34] uses the point cloud diffusion model to generate the point cloud directly conditioned by the CLIP feature from a single view image or text. Another approach in the field of text-to-3D is pioneered by Dream Fields [17], which uses the CLIP model to provide supervision loss. Following works include PureCLIPNeRF [22], CLIP-Mesh [33], CLIP-Forge [43] and Dream3D [58] also use CLIP image encoder to extract rendered image features. Our Alpha-CLIP can enhance CLIP in 3D object generation, enable Point-E with user-defined region focus ability and help optimization based text-to-3D models to yield high-quality generation results.

3. Method

This section describes the data pipeline and framework of Alpha-CLIP. As illustrated in Fig. 3, we first design a data pipeline to generate RGBA-region text pairs data (Sec. 3.1). Using our generated data, we then train our Alpha-CLIP with additional Alpha-channel inputs (Sec. 3.2).

3.1. RGBA Region-Text Pair Generation

To fine-tune the CLIP model with an additional alpha channel input, we first design a data generation pipeline (cf. Fig. 3a) to create millions of RGBA-region text pairs. Our d pipeline consists of the following two components.

Grounding data pipeline. As depicted in the upper part of Fig. 3a, this branch is dedicated to generating region-text pairs, which include natural images with foreground alpha channels and corresponding referring expressions for specific regions. The natural images are from the GRIT dataset [36], which employs GLIP and CLIP to automatically extract labels of *box* region-text pairs. Building upon GRIT, we take a further step of generating *mask* region-text pairs. Specifically, we use SAM [19] to automatically generate high-equality pseudo-masks for each box region.

Classification data pipeline. As illustrated in the lower part of Fig. 3a, this branch is utilized for generating region-text pairs where the foreground objects are highlighted while the original background is removed. We employ the ImageNet [6] dataset for this purpose. Firstly, we use SAM to automatically generate several masks for each image in ImageNet. Subsequently, we crop the foreground object of each mask, center it, and enlarge it. CLIP is then used to calculate scores with the corresponding class label of the image to which each mask belongs. Following this, we sort the masks by class based on their scores and select the top-ranked masks with the highest scores. Regarding the text component, to ensure that the caption for each mask is not merely the ImageNet [6] class label, we place the foreground object on a pure white background. Then we use BLIP-2 [25] to annotate these masks with captions. Finally, we merge the fine-grained ImageNet class label with the image-specific captions generated by BLIP-2 [25], resulting in millions of RGBA region-text pairs.

3.2. Alpha-CLIP

Model structure. Our Alpha-CLIP implements subtle structural modifications to the CLIP image encoder to preserve CLIP's prior knowledge. In the CLIP image encoder's ViT [9] structure, an RGB convolution is applied to the image in the first layer. As shown in Fig. 3b, we introduce an additional Alpha Conv layer parallel to the RGB Conv layer, which enables the CLIP image encoder to accept an extra alpha channel as input. The alpha channel input is set to range from $[0, 1]$, where 1 represents the foreground and 0 indicates the background. We initialize the Alpha Conv kernel weights to zero, ensuring that the initial Alpha-CLIP ignores the alpha channel as input.

Training method. During training, we keep the CLIP text encoder fixed and entirely train the Alpha-CLIP image encoder. Compared to the first convolution layer that processes the alpha channel input, we apply a lower learning rate to the subsequent transformer blocks. To preserve CLIP's global recognition capability for full images, we adopt a specific data sampling strategy during training. We set the sample ratio, denoted as $r_s = 0.1$ to occasionally replace our generated RGBA-text pairs with the original image-text pairs and set the alpha channel to full 1. Please refer to the supplementary for ablation studies such as the number of unfreeze Transformer blocks and value of r_s.

Alpha-CLIP for downstream tasks. After the training, Alpha-CLIP possesses the capability to focus on a specified region and controlled editing. Alpha-CLIP can enhance CLIP's performance on various baselines in a plug-and-play fashion, across various downstream tasks like recognition, MLLM, and 2D/3D generation (see Tab. 1 in Sec. 1).

4. Experiments

Data. We train Alpha-CLIP on RGBA region-text pairs using grounding data pipeline from GRIT-20m [36] for zero-shot ImageNet classification. We combine it with 460k RGBA region-text pair from ImageNet [6] using classification data pipeline to train Alpha-CLIP for other tasks including REC, OVD, region-level captioning, 2D image variation, and 3D generation. Ablation on data volume and mixture of data are in the supplementary material.

Methods	ViT-B/16		ViT-L/14	
	Top-1	Top-5	Top-1	Top-5
Original CLIP [38]	66.48	88.90	73.48	91.60
MaskAdaptedCLIP [28]	57.86	79.12	63.50	86.34
Red Circle [47]	65.37	88.68	73.37	92.09
MaskCLIP* [70]	67.86	89.40	77.04	93.39
Alpha-CLIP(ours)	**68.89**	**90.51**	**77.41**	**94.45**

Table 2. **Zero-shot classification on ImageNet-S [10].** When given the foreground object on the alpha channel, our Alpha-CLIP significantly improves zero-shot classification and surpasses previous baselines such as MaskCLIP [70].

Model	Alpha Map	Top-1	Top-5
CLIP [38]	-	73.48	91.60
Alpha-CLIP	whole image	73.37	91.75
	rectangular box	75.62	93.34
	mask	77.41	94.45

Table 3. **Zero-shot classification on ImageNet-S [10] with different alpha map levels.** Alpha-CLIP is comparable to the original CLIP when the foreground mask is not available, and further boosts the performance with rectangular box or mask alpha maps.

4.1. Alpha-CLIP in Image Recognition

Zero-shot image classification. We select the ImageNet-S [10] dataset for zero-shot classification analysis, which comprises 919 classes with semantic segmentation annotations selected from ImageNet-1k. We prepare the image-level semantic segmentation masks as the alpha channel input. We select representative baseline methods designed for making CLIP focus on the specific region: MaskCLIP [70], MaskAdaptedCLIP [28], and Red Circle [47]. Note that MaskCLIP is designed for mask generation rather than recognition. We make necessary modifications to MaskCLIP to adapt it for the recognition task (please refer to the supplementary for our implementation details). We use the mean of per-class accuracy as the evaluation metric.

Tab. 2 presents the zero-shot classification comparison on ImageNet-S *validation* set. This experiment effectively demonstrates that when provided with a foreground object mask through the alpha channel, our Alpha-CLIP generates visual features that are more focused on the foreground object, leading to better image-level classification compared to the original CLIP and other baseline approaches. It is worth noticing that Although MaskCLIP [70] achieves good results without needing to fine-tune the CLIP model, it is not directly compatible with methods that require the whole feature map instead of just the `[CLS]` token. This limitation is particularly relevant when considering methods like BLIP-2[25], BLIP-Diffusion [24], LLaVA [30] and Point-

Method	RefCOCO			RefCOCO+			RefCOCOg	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
CPT [62]	32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5
ReCLIP [49]	45.8	46.1	47.1	47.9	50.1	45.1	59.3	59.0
Red Circle [47]	49.8	58.6	39.9	55.3	**63.9**	45.4	59.4	58.9
Alpha-CLIP	**55.7**	**61.1**	**50.3**	**55.6**	62.7	**46.4**	**61.2**	**62.0**

Table 4. **Comparison with state-of-the-art on zero-shot REC.** We report top-1 accuracy (%). Replacing CLIP in ReCLILP [49] with Alpha-CLIP outperforms other zero-shot approaches on most datasets, including Red Circle[47], ReCLIP[49] and CPT[62].

E [34]. In contrast, our Alpha-CLIP is more general and can be applied to these approaches effectively.

We also evaluate Alpha-CLIP in scenarios where the foreground mask is unavailable. As shown in Tab. 3, when foreground prior is not available, we set alpha channel input to all one. We observe that the recognition ability of Alpha-CLIP (second row) remains on par with the original CLIP (top row). When provided foreground box (third row) or foreground mask (bottom row), Alpha-CLIP can significantly improve classification accuracy.

Zero-shot referring expression comprehension. In addition to the zero-shot image classification task, we also conducted experiments on zero-shot Referring Expression Comprehension (REC). zero-shot REC is the task of localizing objects in an image given a textual reference expression in a zero-shot manner. We follow previous works to select the RefCOCO [64], RefCOCO+ [64], and RefCOCOg [31] datasets for evaluation. We select three representative approaches CPT [62], ReCLIP [49], and Red-Circle [47] as our baselines. We replace the CLIP model in this task with our Alpha-CLIP. Specifically, we use object proposals predicted by a pretrained detector [65] and employ SAM to obtain masks for each proposal. Instead of cropping the object by bounding box, we input the original image with an alpha map into our Alpha-CLIP. This modification has proven beneficial in preserving global contextual information as we find cropping only lead to worse result. Please refer to the supplementary material for more implementation details.

As shown in Tab. 4, Alpha-CLIP achieves competitive zero-shot results on the REC task, surpassing ReCLIP and RedCircle by an average of 6.8% and 3.0% accuracy across RefCOCO, RefCOCO+ and RefCOCOg benchmarks. The experimental results demonstrate that Alpha-CLIP enhances CLIP's ability to focus on the relevant region and such enhancement is also beneficial for the REC task that requires image-text understanding and reasoning capabilities.

Open vocabulary detection. The Open-Vocabulary Detection (OVD) task aims to detect novel classes that are not available during training. Detic [71] is a pseudo-labeling baseline that proposes to use the ImageNet dataset for OVD. Specifically, Detic first trains the detector on the base classes of LVIS [12], then uses the detector to generate pseudo bounding boxes on ImageNet. These pseudo boxes may cover the novel objects and help improve the detector's performance in novel classes. Such a semi-supervised pipeline is not data-efficient and Detic uses 1.2M images from ImageNet in the OV-LVIS [12] benchmark.

Dataset	$\text{mAP}_{\text{novel}}$	mAP
Detic-ImageNet	24.6	32.4
MaskImageNet (ori CLIP)	**27.9**	**32.5**
MaskImageNet (Alpha-CLIP)	**28.6**	**32.9**

Table 5. **Open-vocabulary detection on OV-LVIS [12].** Using MaskImageNet and our Alpha-CLIP can significantly improve $\text{mAP}_{\text{novel}}$ on novel classes.

To demonstrate the effectiveness of Alpha-CLIP on OVD, we transfer the top-ranked ImageNet (460K) into a collection, dubbed as MaskImageNet. Specifically, we apply our data generation pipeline, as detailed in Sec. 3.1 to generate pseudo-labeled bounding boxes and foreground masks for each image. We replace the ImageNet used in Detic's pseudo-labeling steps with our MaskImageNet. We also remove the background category loss and adjust the blending ratios of LVIS and MaskImageNet. Experimental results are presented in Tab. 5. Compared to the Detic baseline using ImageNet (top row), The second row demonstrates that using our MaskImageNet already enhances OVD capabilities. Furthermore, our Alpha-CLIP (bottom row) further improves OVD performance. Remarkably, our method (460K in MaskImageNet) is more data efficient than Detic (1.2M in ImageNet).

4.2. Alpha-CLIP in MLLM

We replace CLIP used in BLIP-2 [25] and LLaVA-1.5 [29] with our Alpha-CLIP to make MLLM directly focus on user-defined region in vision-language tasks such as region level captioning and VQA.

Region level captioning. As shown in Fig. 4, simply replacing CLIP with Alpha-CLIP enables MLLM to generate captions more focused on user-defined areas. In the third row cases about the telephone and mushroom, the original CLIP generates the wrong caption. This error may arise due to the CLIP vision feature mixing different objects and their properties in images with too many foreground objects. Alpha-CLIP guides MLLM to generate the correct caption by providing the area to focus on. We also visualize the CLIP attention map marked in the upper right to confirm our findings. More visualizations and implementation details of the attention map are in the supplementary material.

Besides qualitative results, we also provide the quantitative region level captioning results of Alpha-CLIP with LLaVA-1.5 [29] on Visual Genome [20] and RefCOCOg [31]. We fine-tune Alpha-CLIP+LLaVA-1.5 [29] with vicuna-7b [52] on these datasets with the same setting in [40, 67] and task prompts in [67] is adopted. Alpha-

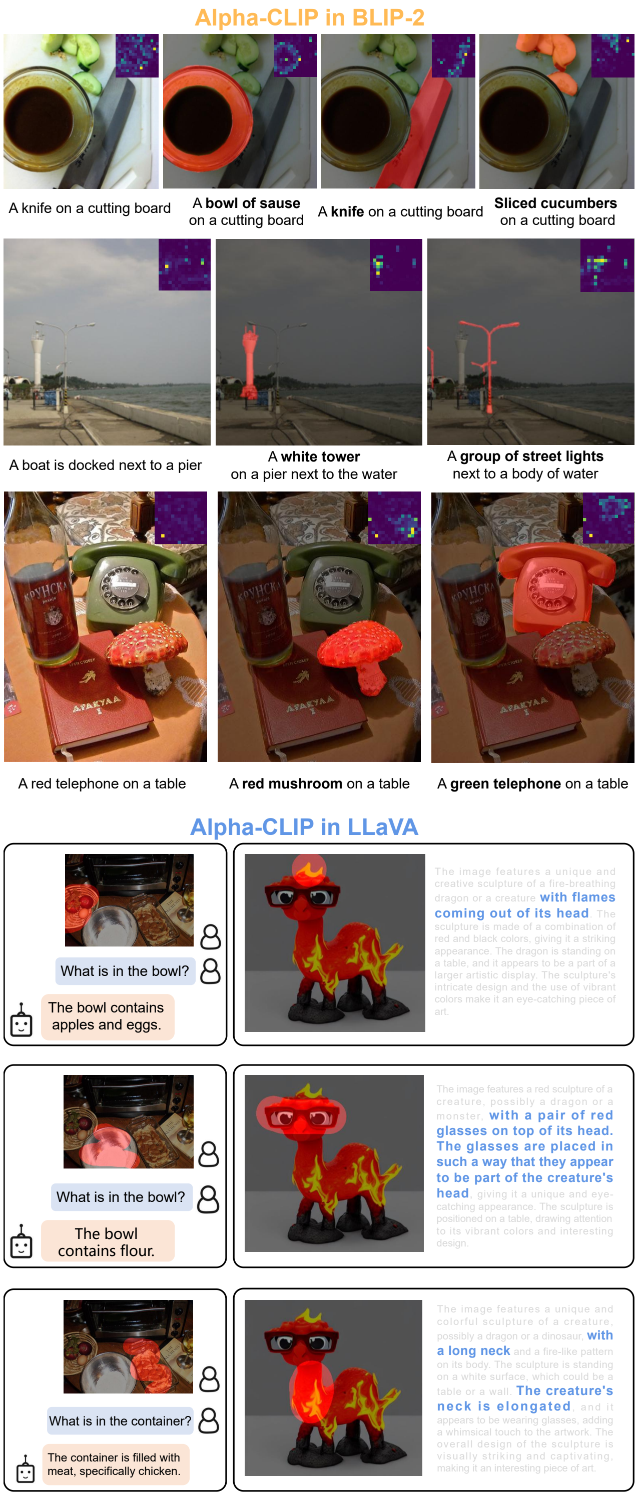

Figure 4. **Some results of Alpha-CLIP used in MLLMs** The upper half is image captioning result with BLIP-2 [25]. The first column is the original CLIP generated captions. Other columns represent the outcomes of Alpha-CLIP with highlighted region marked in red. The lower half is region focused VQA and image captioning result with LLaVA-1.5 [29].

CLIP image encoder is kept frozen with LLM fine-tuned to adapt region caption format. Results are shown in Tab. 6. Alpha-CLIP+LLaVA-1.5 achieves competitive results over the baseline methods, even surpassing previous expert models like GPT4ROI [67] and GLaMM [40] with ROI Align [13] or additional region encoder structure pretrained on a large volume of region-text pairs.

Model	RefCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [56]	15.2	71.6	17.1	142.0
Kosmos-2 [36]	14.1	62.3	-	-
GPT4RoI [67]	-	-	17.4	145.2
GLaMM [40]	16.2	105.0	18.6	157.8
Alpha-CLIP+LLaVA [29]	**16.7**	**109.2**	**18.9**	**160.3**

Table 6. **Performance of Alpha-CLIP in region level captioning.** We report METEOR and CIDEr metrics on Visual Genome and refCOCOg Datasets.

Region based VQA. MLLM can chat with users with simple reasoning. In this scenario, alpha channel input can act as the visual prompt defined by the user to highlight specific regions of interest. As shown in Fig. 4 and Fig. 1, the user can simply use stroke to tell MLLM the referring object or regions to focus on. More visualization results of VQA with Alpha-CLIP are in the supplementary material.

4.3. Alpha-CLIP in image variation.

Alpha-CLIP can be used in most image variation models that use CLIP image encoder [24, 39, 53, 55, 63]. For example, BLIP-Diffusion bridges CLIP [38] and stable-diffusion [41] with Q-former to generate and edit 2D images controlled by text. Since BLIP-Diffusion [24] is a typical method that maintains subject information, we use BLIP-Diffusion to demonstrate the effectiveness of Alpha-CLIP. By introducing Alpha-CLIP, we can add an additional set of vision prompts to allow the model to focus on specified regions for 2D generation. We replace the ViT-L/14 model in BLIP-Diffusion [24] with Alpha-CLIP while keeping the other parts unchanged. We set the empty text prompt to make results irrelevant with semantics. As shown in Fig. 1, Alpha-CLIP with alpha map on highlighted areas enables BLIP-Diffusion to generate region-focused results. We also compare our Alpha-CLIP with other CLIP region-focused approaches such as image cropping, pixel-level image masking, red circle, and feature-level masking (Please refer to supplementary for implementation details). As shown in Fig. 2, image cropping can not solve the occlusion problem. The red-circle solution will change the image content. Neither pixel-level nor feature-level masking can convey original background information. In contrast, our Alpha-CLIP that prompts CLIP with fine-grained region mask solves the above problems and generates cleaner results while maintaining original background information. More visualizations are in the supplementary material.

4.4. Alpha-CLIP in 3D Object Generation.

Alpha-CLIP can also apply to 3D Object Generation. We test it in two different approaches: 1) Point-E [34] that is a diffusion-based method for image-to-3D, and 2) PureCLIP-NeRF [22] that is an optimization-based approach for text-to-3D.

Alpha-CLIP in Point-E. Point-E [34] can achieve image-to-3D through conditioning diffusion model with CLIP image feature. We replace the CLIP ViT-L/14 image encoder of the Point-E base-40M model with our Alpha-CLIP. We demonstrate that Alpha-CLIP is helpful in two cases: 1) When Point-E generates the point cloud with some parts missing, users can highlight the missing part in the condition image to remind the diffusion model to pay more attention to that part and fix this missing parts problem. 2) Users can highlight the part that needs to be emphasized on the 2D image. Point-E will spend more points on the highlighted part (with 1024 points in total in the base model). The results are shown in Fig. 5 with more results in the supplementary material.

Alpha-CLIP in PureCLIPNeRF. We input the rendered images with alpha channels obtained from density integration of NeRF [32] into Alpha-CLIP. When optimizing the object with Alpha-CLIP, the gradient can flow back from the alpha channel to help generate better results. As shown in Fig. 5, we find that PureCLIPNeRF generates objects that closely align with the provided textual prompts(especially bolded text) in terms of shape and color when replacing CLIP with Alpha-CLIP. Furthermore, there is an enhancement in the overall coherence of the generated objects, coupled with notable aesthetic qualities. We attribute this phenomenon to Alpha-CLIP's enhanced capability in optimizing density parameters of 3D representations directly and focusing only on the foreground area, which helps to generate an object that is more coherent and closely matches the input text.

Background augmentation in PureCLIPNeRF [22] inherited from Dream Fields [17] is a vital step to improve the consistency of objects, making them less diffuse compared to the first column in Fig. 5. However, this process is time-consuming as each augmented image has to go through CLIP to get optimization direction. We thus test the capabilities of Alpha-CLIP without background augmentations. Results are presented in the second column of Fig. 5. We observe that in most cases, using Alpha-CLIP without background augmentation produces objects that are clearer and better aligned with the given text than the original CLIP with 2x faster speed. Quantitative results and More visualizations are in the supplementary material

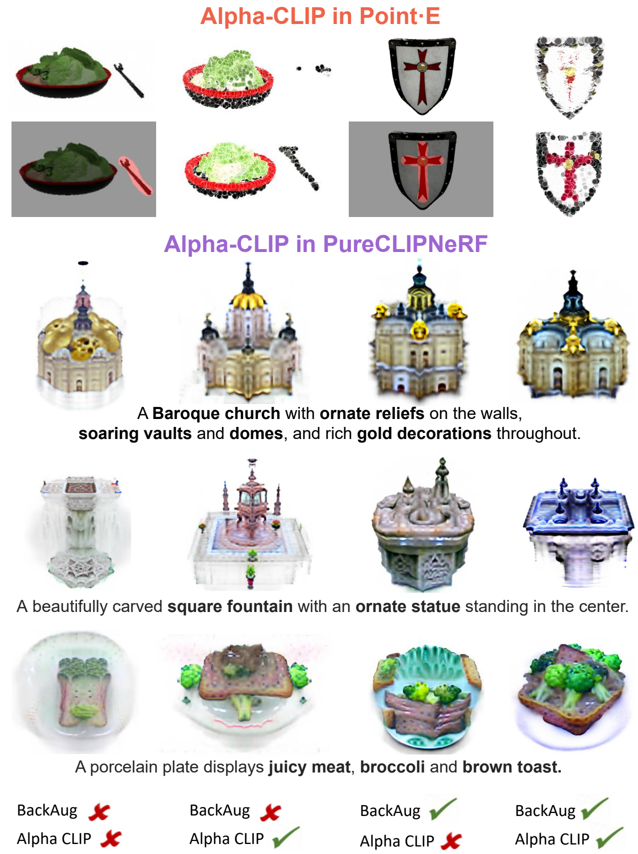

Figure 5. **Results of Alpha-CLIP in 3D generation.** The top part shows 3D point clouds generation using Point·E [34]. The first row displays objects generated by the original CLIP. The second row illustrates the results of Alpha-CLIP, with highlighted areas in red. The bottom part shows 3D objects generated by PureCLIPNeRF [22]. The CLIP model is replaced with Alpha-CLIP, and tests are conducted with and without background augmentation.

5. Conclusion

In this work, We propose the Alpha-CLIP model, which introduces an additional alpha channel to specify the regions of interest. Trained on millions of RGBA region-text pairs, Alpha-CLIP not only exhibits excellent region-focus capabilities but also ensures its output space remains consistent with the original CLIP model. This consistency allows seamless replacement in various downstream applications of CLIP. We demonstrate that when prompted with specific regions of interest, Alpha-CLIP shows improved zero-shot recognition abilities and verifies its usefulness in many downstream tasks. The applications of CLIP extend far beyond the scope of this article. We hope that Alpha-CLIP will be applicable in more scenarios when foreground regions or masks are available.

6. Acknowledgment

This work is partially supported by the National Key R&D Program of China (2022ZD0160201), the University of Macau (SRG2023-00044-FST), and Shanghai Artificial Intelligence Laboratory.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 4

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[3] Jinze Bai, Rui Men, Hao Yang, Xuancheng Ren, Kai Dang, Yichang Zhang, Xiaohuan Zhou, Peng Wang, Sinan Tan, An Yang, et al. Ofasys: A multi-modal multi-task learning system for building generalist models. *arXiv preprint arXiv:2212.04408*, 2022.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4

[5] Keyan Chen, Xiaolong Jiang, Yao Hu, Xu Tang, Yan Gao, Jianqi Chen, and Weidi Xie. Ovarnet: Towards open-vocabulary object attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23518–23527, 2023. 2, 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 5

[7] Zheng Ding, Jie Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *International Conference on Machine Learning*, 2022. 3

[8] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10995–11005, 2023. 2

[9] 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. 5

[10] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 5

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021. 2

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4, 7

[14] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 2

[15] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 2, 3, 4

[16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. https://github.com/mlfoundations/open_clip, 2021. 2

[17] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 2, 4, 8

[18] John R Kender, Parijat Dube, Zhengyang Han, and Bishwaranjan Bhattacharjee. G2l: A high-dimensional geometric approach for automatic generation of highly accurate pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1093–1102, 2023. 2

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3, 4

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[21] S Chandeesh Kumar, M Hemalatha, S Badri Narayan, and P Nandhini. Region driven remote sensing image captioning. *Procedia Computer Science*, 165:32–40, 2019. 2

[22] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. 2, 3, 4, 8

[23] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2, 4

[24] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 2, 3, 4, 5, 7

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training

with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 2, 3, 4, 5, 6, 7

[26] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 3

[27] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23390–23400, 2023. 2

[28] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2, 3, 5

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 4, 6, 7

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 3, 4, 5

[31] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3, 8

[33] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 2, 4

[34] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2, 3, 4, 6, 8

[35] OpenAI. Gpt-4 technical report, 2023. 4

[36] 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. 2, 3, 4, 5, 7

[37] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 2

[38] 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. 2, 3, 4, 5, 7

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3, 4, 7

[40] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model, 2023. 4, 6, 7

[41] 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. 7

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 4

[43] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 4

[44] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[46] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[47] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023. 2, 3, 5, 6

[48] Wei Su, Peihan Miao, Huanzhang Dou, Yongjian Fu, and Xi Li. Referring expression comprehension using language adaptive inference. *arXiv preprint arXiv:2306.04451*, 2023. 2

[49] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215, 2022. 2, 3, 6

[50] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue

Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023. 2

[51] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2

[52] Vicuna. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org/, 2023. 6

[53] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 4, 7

[54] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, Yushi Chen, Tong Lu, Jifeng Dai, and Yu Qiao. The all-seeing project: Towards panoptic visual recognition and understanding of the open world, 2023. 3

[55] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2, 4, 7

[56] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 7

[57] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2, 3

[58] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 4

[59] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 2, 3

[60] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masqclip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 887–898, 2023. 2, 3

[61] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *arXiv preprint arXiv:2306.04356*, 2023. 2, 3

[62] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 6

[63] 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. 2, 4, 7

[64] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 6

[65] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2, 3, 6

[66] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 2, 4

[67] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 4, 6, 7

[68] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022. 2

[69] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2, 3

[70] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2, 3, 5

[71] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 2, 3, 6

[72] Yijie Zhou, Likun Cai, Xianhui Cheng, Zhongxue Gan, Xiangyang Xue, and Wenchao Ding. Openannotate3d: Open-vocabulary auto-labeling system for multi-modal 3d data. *arXiv preprint arXiv:2310.13398*, 2023. 2