O3SLM: Open Weight, Open Data, and Open Vocabulary Sketch-Language Model

Rishi Gupta[1], **Mukilan Karuppasamy**[1*], **Shyam Marjit**[1*], **Aditay Tripathi**[1†], **Anirban Chakraborty**[1]

[1]Indian Institute of Science, Bangalore
rishig@iisc.ac.in, mukilan.nitt@gmail.com, shyammarjit@iisc.ac.in, aditaytr@gmail.com, anirban@iisc.ac.in

Abstract

While Large Vision Language Models (LVLMs) are increasingly deployed in real-world applications, their ability to interpret abstract visual inputs remains limited. Specifically, they struggle to comprehend hand-drawn sketches, a modality that offers an intuitive means of expressing concepts that are difficult to describe textually. We identify the primary bottleneck as the absence of a large-scale dataset that jointly models sketches, photorealistic images, and corresponding natural language instructions. To address this, we present two key contributions: (1) a new, large-scale dataset of image-sketch-instruction triplets designed to facilitate both pretraining and instruction tuning, and (2) **O3SLM**, an LVLM trained on this dataset. Comprehensive evaluations on multiple sketch-based tasks: (a) object localization, (b) counting, (c) image retrieval *i.e.*, (SBIR and fine-grained SBIR), and (d) visual question answering (VQA); while incorporating the three existing sketch datasets, namely QuickDraw!, Sketchy, and Tu Berlin, along with our generated **SketchVCL** dataset, show that **O3SLM** achieves state-of-the-art performance, substantially outperforming existing LVLMs in sketch comprehension and reasoning.

Project page — https://vcl-iisc.github.io/O3SLM/

1. Introduction

The landscape of Large Vision Language Models (LVLMs) is advancing rapidly, with models becoming increasingly powerful and accessible. Foundational open-source models such as LLaVA (Liu et al. 2023), Qwen-VL (Wang et al. 2024), DeepSeek-VL2 (Wu et al. 2024), and Molmo (Deitke et al. 2025) have demonstrated remarkable vision-language capabilities on benchmark tasks like Visual Question Answering (VQA) and Document VQA. This success has catalyzed a wave of development in fine-grained LVLMs (Kuckreja et al. 2024; Zhang et al. 2024; Cheng et al. 2024) that target specialized domains and tasks. For instance, recent work has focused on adapting these models for object detection (Zhang et al. 2024) and depth estimation (Cheng et al. 2024). While text is the primary modality for LVLMs, it struggles to convey complex or nuanced visual ideas efficiently, more precisely tasks involving fine-grained images. Communicating intricate spatial arrangements or specific object attributes through text alone can be cumbersome and ambiguous. Hand-drawn sketches offer a powerful solution by enabling intuitive visual prompting. Furthermore, sketches transcend linguistic barriers, making them a more accessible and universal communication tool compared to text, which demands descriptive proficiency from the user.

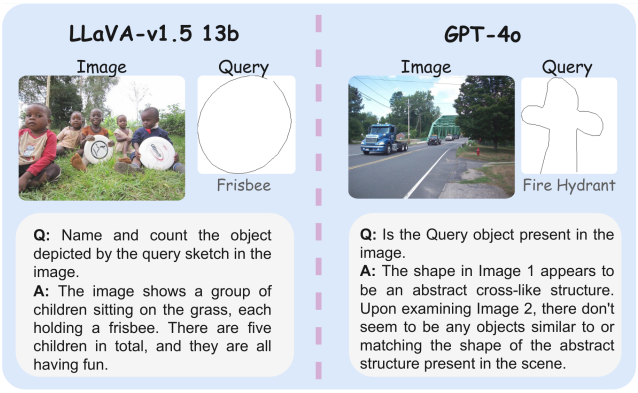

Figure 1: **Limitations of LVLMs in Sketch Understanding.** Although current LVLMs can interpret sketches to some level of abstraction, they struggle in sketch understanding for downstream tasks like detection and reasoning.

Despite their expressive power, the inherent nature of sketches makes them challenging for machine perception (Mathur et al. 2025). They are highly abstract and exhibit significant variability based on artistic style, cultural background, and drawing skill (Alaniz et al. 2022). This complexity has made sketches a long-standing subject of study in computer vision, driving research in tasks like sketch-based image retrieval (SBIR) (Koley et al. 2024a,b), object detection (Chowdhury et al. 2023b; Tripathi, Mishra, and Chakraborty 2024; Tripathi et al. 2020), classification (Tiwari, Biswas, and Lladós 2024; Bandyopadhyay et al. 2024; Koley et al. 2025a), segmentation (Koley et al.

*These authors contributed equally.
†Major contributions made during association with IISc.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

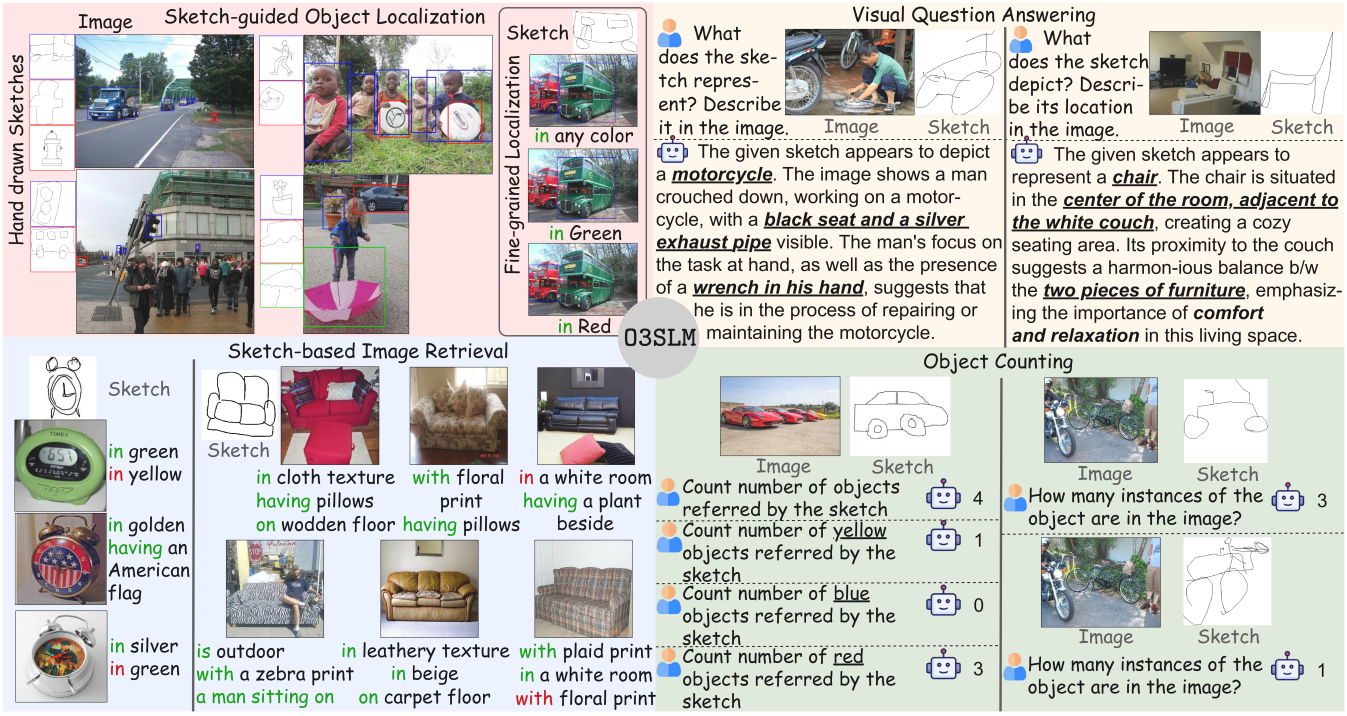

Figure 2: **Capabilities of our model - O3SLM.** Our model is the first Large Vision-Language Model (LVLM) to demonstrate advanced alignment between sketches, images, and text—where existing LVLMs consistently fail (see Table 2). Through extensive pretraining on our proposed **SketchVCL** dataset, the model develops a robust understanding of crude hand-drawn sketches and how they relate to the visual and textual modalities in which current LVLMs already excel. This training enables cross-modal transfer, allowing the model to handle fine-grained queries using sketch-text pairs, even though it was originally trained with sketches alone. **O3SLM** is trained across multiple tasks, including Visual Question Answering (VQA), Sketch-Based Image Retrieval (SBIR), sketch-based counting, and sketch-based object detection.

2025b,c), and generation (Liu et al. 2025; Koley et al. 2023; Banerjee et al. 2024). It is this variability and abstraction that current open-source LVLMs struggle to comprehend.

Consider a case where we try to query the existing LVLMs with crude hand-drawn sketches, to give a detailed description of the sketch. Despite excelling at understanding natural images and structured visual inputs, current models consistently fail to make sense of crude sketches. More importantly, even when they grasp some visual cues, they are unable to leverage this information to perform tasks as shown in Figure 1. Such a scenario highlights a critical gap: while sketches reside in the image domain, their abstraction and ambiguity make them fundamentally different. As LVLMs evolve, understanding sketches alongside text and natural images is increasingly crucial. Yet, sketches remain a major blind spot in open-weight models - primarily due to the lack of a large-scale, diverse, open-source dataset combining all three modalities.

To address the aforementioned gap, we introduce **SketchVCL**, a comprehensive multi-task dataset of image-sketch-instruction triplets to train LVLMs for four fundamental sketch-based reasoning tasks: **(a)** object localization, **(b)** image retrieval, **(c)** counting, and **(d)** sketch-aware VQA. We use this dataset to train our model **O3SLM** - a novel LVLM engineered for sketch-based reasoning. Our experiments show that this model achieves state-of-the-art (SOTA) performance on all the aforementioned tasks against open-weight models. By releasing both the dataset and a high-performing model, we aim to unlock the potential of LVLMs to reason with sketches as fluently as natural images.

Our key contributions are as follows:

1) We construct **SketchVCL**, a large-scale, multi-task instruction-tuning dataset composed of ⟨image, sketch, instruction⟩ triplets (refer to Table 1). This is enabled by two key innovations: **(a)** A novel sketch generation pipeline to generate fine-grained sketches paired with specific object instances within an image. **(b)** A two-stage training structure. First, a large-scale sketch-alignment stage, developing sketch understanding for an open vocabulary setup. Second, an instruction alignment stage to tune our model to task-specific instructions.

2) We introduce **O3SLM**, a unified LVLM designed to handle diverse sketch-based grounding and reasoning tasks within a single framework. **O3SLM** is fine-tuned on our curated dataset, leveraging the multi-task curriculum to achieve superior sketch-image feature alignment.

3) We demonstrate through comprehensive evaluation that **O3SLM** establishes new SOTA performance on sketch-guided object localization, image retrieval, and VQA (counting too) against open-sourced LVLMs. Notably, our single

model significantly outperforms general-purpose zero-shot LVLMs, even surpassing closed source models above its weight class like GPT-4o and Gemini 1.5 Pro.

2. Related Works

Sketches as a Visual Modality. Sketches offer a uniquely abstract and expressive visual modality, widely explored in tasks such as SBIR (Koley et al. 2024a,b), object detection (Tripathi et al. 2020; Tripathi, Mishra, and Chakraborty 2024; Chowdhury et al. 2023b), image synthesis (Koley et al. 2024c), and even video generation (Liu et al. 2025). Although rich in expressiveness, sketches present challenges for machine perception due to their high variability, abstraction, and inherent noise. These characteristics often result in limited generalization performance, especially when deployed in open-world settings with novel classes or unseen domains (Tripathi, Mishra, and Chakraborty 2024). A major cause of this is the lack of large-scale, open-vocabulary pretraining using sketch modalities—most prior works rely on narrowly scoped datasets and task-specific architectures. Our work addresses this limitation through a unified pretraining and instruction tuning framework that aligns sketches with both natural images and text, enabling scalable sketch understanding across multiple vision-language tasks.

Multimodal Fusion of Sketch and Text. Sketch and language provide complementary cues for visual reasoning - while sketches capture spatial and shape-based priors (Koley et al. 2024b), text encodes semantic and contextual information (Chowdhury et al. 2023a; Koley et al. 2025a). Prior works have explored sketch-text fusion for tasks like fine-grained SBIR (Baldrati et al. 2023; Saito et al. 2023) and scene captioning, yet these approaches are often limited to single-task pipelines or rely on handcrafted feature fusion mechanisms. Importantly, existing systems do not operate under open-vocabulary constraints, nor do they support unified pipelines that can generalize across task boundaries. To the best of our knowledge, ours is the first work to demonstrate native support for joint sketch-text queries in tasks such as object detection and counting in LVLMs.

Large Vision-Language Models (LVLMs). Open-sourced LVLMs such as LLaVA (Liu et al. 2023), Qwen-VL2 (Wang et al. 2024), DeepSeek-VL2 (Wu et al. 2024), Pixtral (Agrawal et al. 2024), and Molmo (Deitke et al. 2025) have shown impressive multimodal reasoning capabilities. However, these models fail in the presence of abstract visual inputs such as sketches. MiniGPT-v2 (Chen et al. 2023) is one of the few LVLMs spatially trained for grounding tasks; however, it lacks native support for multi-image input, disallowing usage of sketches to query images. On the other hand, closed-source models like GPT-4o (Hurst et al. 2024) and Gemini 1.5/2.5 (Team et al. 2024) show some initial sketch understanding, but their multimodal grounding remains weak. Moreover, inaccessibility and lack of interpretability limit their applicability in open scientific research. Our model, **O3SLM**, is built to bridge this gap by supporting joint reasoning over sketches, natural images, and text, and demonstrates robust generalization in retrieval, localization, and reasoning tasks, even under zero-shot and OOD settings. Recently, (Lee et al. 2024; Fu et al. 2025b) attributed the text-to-image retrieval task to LVLMs, yet to adhere to the sketch understanding capabilities.

Sketch generations from Images. Recent works on sketch generation from images or text explore various approaches. Early GAN-based methods (Li et al. 2019) are fast but limited in quality and data diversity. Later methods (Mathur et al. 2025; Vinker et al. 2022; Xing et al. 2024; Vinker et al. 2023) use Bézier curves and VLMs like CLIP (Radford et al. 2021) for alignment, but produce overly simple sketches with clean backgrounds. Diffusion-based models (Xing et al. 2023; Arar et al. 2025) generate high-quality sketches but are extremely slow. For our large-scale dataset, we adopt the Photo2Sketch (Li et al. 2019) method, which offers better quality than CLIP-based methods.

Sketch Datasets. While several popular sketch datasets exist—such as QuickDraw! (Jongejan et al. 2016), Sketchy (Sangkloy et al. 2016), ShoeV2, QMUL-ChairV2 (Yu et al. 2016), TU-Berlin (Eitz, Hays, and Alexa 2012), and SketchyCOCO (Gao et al. 2020)—none are well-suited for large-scale training of LVLMs. QuickDraw and TU-Berlin consist only of class-wise sketches without paired images. Sketchy, ShoeV2, and QMUL-ChairV2 include image-sketch pairs, but these are limited to simple, single-object scenes designed primarily for the SBIR task. Crucially, none of these datasets include textual descriptions or question-answer pairs, which are essential for training LVLMs. SketchyCOCO provides paired sketches for the COCO dataset, but it is limited to instance-level sketches from only 14 object categories and also lacks associated text. We have summarized the limitations of current sketch datasets in supplementary. Our proposed dataset, **SketchVCL**, addresses these limitations by incorporating a diverse set of images and sketches from multiple sources, along with rich textual annotations. This makes it a valuable resource for integrating sketches into multimodal LLMs at scale.

3. SketchVCL Dataset

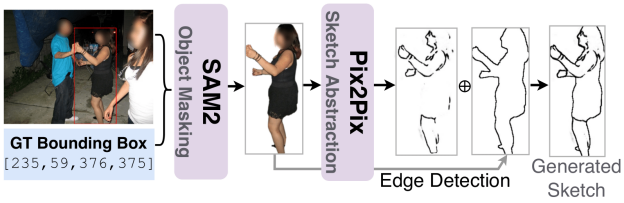

Figure 3: **Automated Large-Scale Sketch Generation Pipeline.** For each object instance, we use the SAM2-generated segmentation maps to mask the background and pass the foreground through Pix2Pix (Li et al. 2019) for sketch generation. These sketches are enhanced using edge detection using morphological gradients. The final sketch is an aggregation of the edges and the Pix2Pix sketch.

Due to the absence of instruction-aligned sketch-image datasets, we propose **SketchVCL** - a large-scale dataset for sketch-image-text alignment and instruction tuning.

SketchVCL comprises 600K instructions for pretraining on OpenImages (Kuznetsova et al. 2020) and Object365 (Shao et al. 2019) datasets, and 50K instructions on COCO (Lin et al. 2014) for finetuning.

OpenImages and Objects365 contain 600 and 365 classes, respectively, which offer broad domain and vocabulary coverage. However, obtaining sketches for such a large class distribution remains a challenge. To address this, we develop a sketch curation pipeline, which generates instance-level sketches for these datasets. This allows us to get paired sketches for large-scale pretraining datasets. Similarly, for finetuning, we generate sketches over COCO to represent the classes absent from Sketchy and QuickDraw!.

Sketch Curation Pipeline

We assume sketches as pixel information abstraction, where the sketches usually consist of contours and edges of the objects. This accumulation of edges will lead to a meaningful yet difficult to comprehend attribute, which is often complementary for explaining objects, as it stores the fine grained details like pose and shape.

To leverage this, we curate object specific sketch datasets: SketchVCL-OI, SketchVCL-O365, and SketchVCL-C, derived from single object instances within the training subsets of OpenImages, Object365, and COCO, respectively. We have summarized the pipeline in Figure 3, through which, we curated 19M and 14M sketches from Object365 and OpenImages, respectively.

Stage	Task	Image Dataset	Sketch Dataset	# Size
Pretraining	Detailed description	Objects365	SketchVCL-O365	300k
	with bounding box	OpenImages	SketchVCL-OI	300k
Finetuning	Object Detection	COCO	SketchMIX	110k
	VQA	COCO	SketchMIX	50k
	Counting	PixMo Count	SketchMIX	30k
	SBIR	Sketchy	SketchMIX	25k

Table 1: **Training Data Composition.** The distribution of data for each task and corresponding datasets is shown. The total pretraining size is 600k, while the total finetuning size is 215k. Instruction tuning data is curated based on the downstream tasks. More in supplementary.

Stage I: Sketch Alignment

As shown in Table 2, current LVLMs struggle to comprehend hand-drawn sketches. To address this issue, we introduce a large-scale pretraining stage. The goal of this stage is to make our model align sketches with the image and text.

The pretraining phase is designed such to teach the model correspondences across the three modalities: hand-drawn sketches, natural images and text. Specifically, we keep the following goals in mind for the model: i) recognize the sketch, ii) associate the object in the natural image with the sketch, iii) develop fine-grained spatial understanding required for object detection - an ability most LVLMs lack, and iv) retain its natural language capabilities, such as describing visual content through text. In this way, we ensure 3-way alignment of sketches, images, and text.

To curate our pretraining dataset, we first randomly sample 250K images each from Objects365 and OpenImages. Sketches for these datasets are synthetically generated from our pipeline. For each image, we randomly sample one of the annotated object classes. Further, to ensure adequate representation of all classes, we identify *tail classes* (those with fewer than 5000 instances) and select an additional 50K images, with a balanced sampling strategy over these tail classes, from either datasets. This results in 600K images, each paired with a target object class. For each image, we generate a descriptive caption of the target objects using DeepSeek-VL2. These captions are further refined and aligned with our task format using LLaMA-3-8B Instruct. Each response begins by identifying the sketch, followed by a detailed description of the object's appearance, its relation to surrounding elements, and concludes with bounding box coordinates around each instance of the target category.

Stage II: Instruction Tuning

While Stage I is responsible for aligning sketches to the images, in Stage II, we align the model to sketch-based tasks and multi-round conversation. Specifically, we train the model across four tasks: Counting, Object Localization, Visual Question Answering (VQA), and Sketch-based Image Retrieval (SBIR). To facilitate instruction tuning, we curate a task-specific dataset, summarized in Table 1. Following (Deitke et al. 2025), we used task specific natural language descriptors as a prefix to the prompt rather than adding tokens to the vocabulary. For all the tasks, the sketches are randomly sampled from SketchMIX for the corresponding class, as described later in this section. Task prompts are also randomly selected to avoid prompt overfitting and improve robustness. Additional implementation details are in the supplementary.

The curated instruction set covers four tasks: **(1) Counting:** We use the descriptor string *COUNT* as the prefix to the prompt. We use 30K images from the training subset of Pixmo-count dataset, each annotated with ground truth counts. **(2) Object Detection:** For this task, the descriptor *BBOX* is used as a prefix. We use the training split of COCO to get 110K instructions each with a unique image-class pair. Each instance is an aggregation of all the bounding boxes for a class in an image. **(3) VQA:** The descriptor string *VQA* is added as a prefix for VQA prompts. Detailed image descriptions are taken from ShareGPT4V (Chen et al. 2024), and Llama-3 (Grattafiori et al. 2024) generates multi-round, complex reasoning-based questions using the descriptions from ShareGPT4V. To balance sketch-based and general reasoning, we create 25K sketch-based QA pairs and 25K without sketches. **(4) SBIR:** We use the descriptor *SBIR* as a prefix to the prompt for this task. Sketchy (Sangkloy et al. 2016) images, each containing a single object class are used to curate 12.5K positive and 12.5K negative sketch-image pairs. The answer to this task is strictly *yes* or *no*, which is used to compute accuracy as discussed in Section 5.

SketchMIX. For Stage II training, we construct a diverse sketch pool by combining sketches from multiple sources; throughout this paper, we refer to this combination

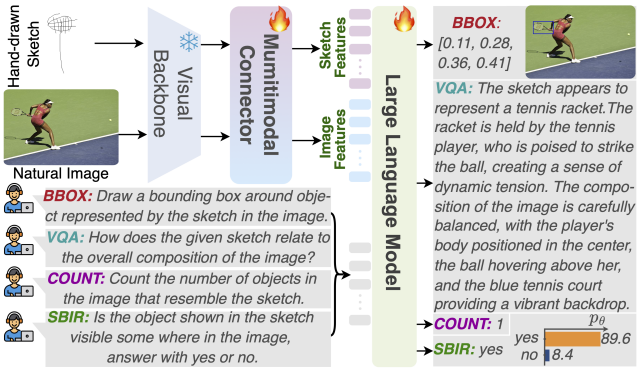

Figure 4: **Summary of O3SLM.** We use CLIP-L-336 as the visual backbone. The hand-drawn sketch and natural image are encoded using this backbone, then the multimodal connector projects the sketch and image features to the input space of the LLM. Finally, the sketch, image, and text tokens are concatenated and passed through the LLM.

as SketchMIX. To adapt our model to hand-drawn sketches, we include sketches from Sketchy and QuickDraw!, which represent higher quality and abstract sketches, respectively. For more diversity, we also include sketches generated from our pipeline on COCO and Objects365 images. To assess the model's ability to generalize to an unseen style of sketches, we intentionally exclude the TU-Berlin (Eitz, Hays, and Alexa 2012) dataset from SketchMIX. TU-Berlin offers a wide range of abstraction levels in its sketches, which serves well for evaluation. While mixing sketches from various datasets, we need to combine the classes across datasets. Hence, we adopt the Objects365 class taxonomy as a common set and map classes from each of the sketch datasets to this set; we have between 200-350 sketches in every class. We have described the method of curating SketchMIX in detail in the supplementary[1].

4. O3SLM

Our proposed **O3SLM** aims to align hand-drawn sketches to the visual information learnt by open LVLMs. We use CLIP ViT-L/336 as the visual backbone to encode both sketches and natural images; the 336px variant allows a higher input resolution for better fine-grained spatial understanding. The extracted embeddings are mapped to the LLM's input space via the multimodal connector (a two-layered MLP). Our base language model is Vicuna v1.5. We initialize the model weights from LLaVA-1.5 to leverage its strong text-image alignment and large-scale training. In summary, our architecture has three components: i) a visual backbone (CLIP-L), ii) a multimodal projector (MLP), and iii) a large language model (LLM). We concatenate the sketch, image, and text tokens before feeding them to the LLM. Since concatenating two sequences and applying self-attention is a more powerful alternative to cross-attention across these sequences, we avoid introducing additional alignment modules. This yields a conceptually simpler yet effective framework, relying on the LLM's internal self-attention and large capacity to learn alignment implicitly. The architecture is summarized in Figure 4.

Tasks. We explicitly train **O3SLM** for four tasks during the fine-tuning stage. For each task, we provide a sketch, an image, and a text prompt. Following Molmo (Deitke et al. 2025), we prefix each task with a small string unique to each task. This aligns the model's output to a consistent format, which is helpful during evaluation. We train the model for multi-round visual question answering by prompting the model with an image, a textual question, and a sketch to refer to a specific object. The questions ask the model to describe the object in the image referred to by sketches. Apart from visual attributes like color, appearance, surroundings, etc, we also question the model about things like the purpose of objects, etc. For counting, we ask the model to count how many instances of a sketch exist in the image and train it to return a single integer. For the localization task, we train the model to return bounding boxes in $\{[x_1, y_1, x_2, y_2]\}$ format. We also introduce a simple yet effective approach for image retrieval using LVLMs, which fits into current frameworks for training and inferring LVLMs. We have summarized these tasks in Figure 2 and discuss them in detail in Section 5. Notably, our model is able to generalize across tasks and is able to utilize fine-grained queries.

5. Results and Discussion

In this section, we present the performance of **O3SLM** on sketch-based object detection, counting, and image retrieval tasks, followed by ablation studies and analysis.

5.1. Performance on Downstream Tasks

In this section, we evaluate **O3SLM** across various sketch-based tasks. Further, we demonstrate that **O3SLM** can handle combined text-sketch queries, showing its fine-grained multimodal understanding.

Counting. We introduce the sketch-based object counting task to evaluate a model's ability to identify and count objects specified solely through a sketch. Unlike full object detection or localization, this task focuses on counting instances, making it a comparatively lighter task. Given the natural image, sketch pair, and ground-truth count of the class: $X_i = (I_i, S_i, c_i)$, the goal is to predict how many instances of the sketched object appear in the image. The natural images $I_i \in D_v$ are sampled from the validation subsets of COCO and Pixmo-count, while $S_i \in D_s$ is sampled from 4 Sketch datasets: QuickDraw, TU Berlin, Sketchy, and SketchVCL-C. The accuracy for the counting task is computed as, Accuracy $= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\,[\,\hat{c}_i = c_i\,]$ here, $\hat{c}_i$ represents the prediction for the i^{th} instance.

As shown in Table 2, our model generalizes well to complex counting scenarios, performing strongly on COCO-Count, which averages three object classes per image. Pixmo-Count, with mostly single-class images, is simpler, and the model remains competitive. In a pure zero-shot set-

[1]Supplementary material is provided in the extended arXiv version, available at: https://vcl-iisc.github.io/O3SLM/.

Large Vision-Language Models	PixMo-Count Sketchy (153)	PixMo-Count QuickDraw! (194)	PixMo-Count Tu Berlin† (453)	PixMo-Count SketchVCL-C (457)	PixMo-Count Avg.	COCO Sketchy (180)	COCO QuickDraw! (299)	COCO Tu Berlin† (430)	COCO SketchVCL-C (472)	COCO Avg.
API call only										
GPT-4o-mini (Hurst et al. 2024)	48.4	40.2	20.5	17.5	31.7	14.4	19.1	11.6	14.2	14.8
GPT-4o (Hurst et al. 2024)	45.1	45.9	24.7	18.8	33.6	13.9	20.1	12.8	18.9	16.4
Gemini 1.5 Flash (Team et al. 2024)	33.3	30.4	15.2	10.9	22.5	17.2	24.0	11.6	16.7	17.4
Gemini 1.5 Pro (Team et al. 2024)	37.3	42.3	27.6	22.8	32.5	16.1	19.1	16.0	16.7	17.0
Open weights (≈ 7B Model Size)										
LLaVA-1.5-7B (Liu et al. 2023)	22.2	18.0	13.2	10.6	16.0	16.1	12.7	10.5	9.1	12.1
Qwen2.5-VL-7B (Wang et al. 2024)	25.5	25.3	13.3	6.8	17.7	2.2	6.0	4.2	5.9	4.6
DeepSeek-VL2-small (Wu et al. 2024)	40.5	20.6	12.4	8.1	20.4	8.9	7.4	2.8	5.3	6.1
Molmo-7B-D (Deitke et al. 2025)	32.7	36.6	32.7	19.3	30.3	19.4	20.4	2.3	5.9	12.0
O3SLM-7B (Ours)	41.8	33.0	50.6	48.6	**43.5**	35.6	29.8	30.2	29.7	**31.3**
Open weights (≈ 13B Model Size)										
LLaVA-1.5-13B (Liu et al. 2023)	1.3	8.8	2.0	9.9	5.5	2.8	10.0	2.6	6.1	5.4
Pixtral-12B (Agrawal et al. 2024)	39.2	26.8	34.4	12.7	28.3	16.7	14.0	18.4	8.1	14.3
DeepSeek-VL2 (Wu et al. 2024)	21.6	9.8	5.5	2.6	9.9	0.6	0.7	0.0	1.3	0.6
O3SLM-13B (Ours)	45.1	37.6	47.0	46.4	**44.0**	36.7	29.8	30.5	29.9	**31.7**

Table 2: **Evaluation on Sketch-Based Counting.** We evaluate performance on images from COCO and PixMo-Count datasets (Deitke et al. 2025). COCO presents a more challenging setting, with a more object categories per image, forcing the model to rely more on the sketches as a query. We sample sketches from four datasets representing various levels of abstraction and difficulty of hand-drawn sketches, for example QuickDraw! has highly abstract and often incomplete sketches. † indicates sketch datasets which are unseen by our model during training; they assess our model's ability to generalize to sketch styles.

ting on TU Berlin sketches, it achieves the best performance, demonstrating robust cross-modal transfer.

Object Detection. Object detection with LVLMs is challenging because next-token prediction does not align with spatial localization. Following (Kuckreja et al. 2024), we therefore report Accuracy as a softer and more interpretable metric instead of mAP. As shown in Table 3, our model substantially outperforms prior work. The very low accuracy of existing models highlights how introducing sketch queries severely degrades detection performance due to the large domain gap between sketches and natural images.

Sketch-based Image Retrieval. For a given sketch S and a gallery of images $I = \{I_1, I_2, \cdots I_N\}$, the task of sketch-based image retrieval (SBIR) aims to retrieve the top-k images from the gallery, which align closely with the sketch S. Let T represent the text prompt and $X_i = \{I_i, S, T\}$ be an input triplet. We define our training objective as:

$$\underset{\theta}{\text{argmin}} \quad -\sum_{i=1}^{N} [y_i \log(p_\theta(\texttt{<yes>}|X_i)) + (1-y_i)\log(p_\theta(\texttt{<no>}|X_i))] \tag{1}$$

Here, <yes> and <no> are tokens from the LLM's vocabulary. y_i represents the ground truth label; it is 1 for positive classes (i.e., the image I_i matches the sketch S) and 0 for negative classes. This objective is very similar to the binary cross-entropy, and it can directly be used to train LLMs without altering their training framework.

During inference, we simply select the top k images with the highest confidence for the <yes> token. This is achieved by sorting the probabilities in descending order and selecting the top-k entries: $\underset{i}{\text{argsort}}\, p_\theta(\texttt{<yes>}|X_i)[-k:]$.

We report SBIR results on the Sketchy dataset in Table 4. For evaluation, we rank the gallery images based on the probability assigned to the token <yes> that a given sketch corresponds to each of the gallery images. We compute top-K accuracy (Acc@K) by checking how many of the top-K retrieved images belong to the same class as the query sketch. We compute these metrics across all query sketches. It is important to note that since there are only 5 matching gallery images per query, the maximum achievable Acc@10 is 50. For each sketch, we perform forward passes with all the gallery images, resulting in a total of 10,000 forward passes for the 100 query sketches. Table 4 shows that the baseline LLaVA-1.5 struggles to interpret sketches and associate them with corresponding images. In contrast, **O3SLM** demonstrates a strong understanding of sketches.

5.2. Model Ablations

Figure 5 discusses the effect of skipping the pretraining stage, being highly sketch dependent SBIR benefits significantly from pretraining. We discuss these results for the detection task in the supplementary.

Freezing Multimodal Connector. As shown by the quantitative metrics in Figure 6, we get significant gains by tuning the projector along with the LLM. Notably, training only the projector on our 7B model outperforms using a 13B model with a frozen projector. This demonstrates the benefit of aligning sketches and images at the projector level.

5.3. Analysis

Image-Only Performance. A natural question to ask is whether **O3SLM** is still capable of original image only tasks performed by LLaVA. In Table 5, we see $< 5\%$ decrease in popular image-only benchmarks, and we see a large jump in text-only detection compared to LLaVA.

Models	Sketchy					QuickDraw!					Tu Berlin†					SketchVCL-C				
	Acc	Acc@0.5	Acc^S	Acc^M	Acc^L	Acc	Acc@0.5	Acc^S	Acc^M	Acc^L	Acc	Acc@0.5	Acc^S	Acc^M	Acc^L	Acc	Acc@0.5	Acc^S	Acc^M	Acc^L
LLaVA-1.5-7B	3.5	9.1	0.0	0.0	3.5	2.7	6.9	0.0	0.2	2.8	3.5	9.7	0.0	0.6	3.5	2.9	7.4	0.0	1.1	2.9
OneVision	2.8	9.2	0.0	2.2	2.8	3.8	9.5	0.1	0.9	3.9	2.7	9.4	0.0	1.8	2.8	3.6	8.9	0.0	1.1	3.7
DeepSeek-VL2-small	3.3	9.7	0.0	0.5	3.3	3.8	10.3	0.0	1.1	3.9	3.8	11.4	0.0	2.2	3.8	4.5	11.3	0.0	0.6	4.7
Molmo-7B-D	1.8	5.3	0.0	0.7	1.8	2.9	7.9	0.0	2.1	2.9	2.1	7.5	0.0	2.4	2.0	2.1	5.3	0.0	2.0	2.0
O3SLM-7B (Ours)	19.4	33.9	2.1	13.3	33.2	13.6	23.8	1.3	9.3	26.1	16.8	29.4	1.7	11.0	29.9	11.8	21.5	1.3	8.5	23.9
LLaVA-1.5-13B	4.2	11.4	0.0	0.3	4.2	2.9	7.1	0.0	0.6	2.9	3.6	10.1	0.5	0.2	3.6	2.8	7.4	0.3	0.9	2.9
DeepSeek-VL2	2.3	6.0	0.0	0.0	2.3	1.7	5.6	0.0	1.2	1.8	2.1	6.0	0.0	2.3	2.1	1.5	4.0	0.0	0.9	1.5
O3SLM-13B (Ours)	21.3	35.6	2.8	15.5	35.2	16.7	28.1	2.0	11.8	29.1	18.7	31.5	2.1	13.2	32.1	14.6	24.8	1.3	11.0	26.0

Table 3: **Sketch-Based Object Detection.** To evaluate the sketch-based object detection on images COCO val2017, and sketches from four different datasets, specifically: Sketchy, QuickDraw!, TU-Berlin, and SketchVCL-C. Following (Kuckreja et al. 2024), we report the Acc metric, we include mAP scores of our model in our supplementary. † indicates sketch datasets that are unseen by our model during training; they assess our model's ability to generalize to sketch styles.

Models	Acc@1	Acc@5	Acc@10
LLaVA-1.5-7B	11.0	14.4	13.0
O3SLM-7B	**65.0**	**59.2**	**39.4**
LLaVA-1.5-13B	10.0	9.2	8.3
O3SLM-13B	**55.0**	**46.4**	**32.9**

Table 4: **SBIR.** Performance on Sketchy dataset. The substantial gains indicate that although the original LLaVA has very limited sketch understanding, our training data and methodology align sketches and text in **O3SLM**.

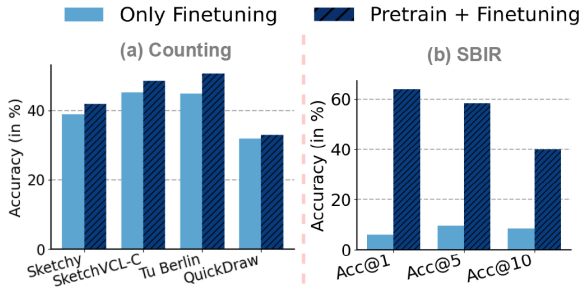

Figure 5: **Effect of Pretraining.** We assess the impact of our large-scale pretraining stage on two tasks. SBIR tasks significantly benefit from pretraining (Right), whereas the effect on counting is minimal (Left).

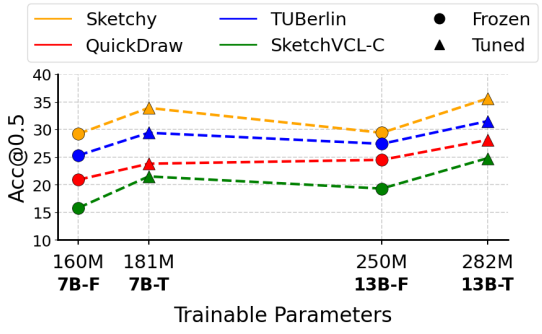

Figure 6: **Freezing Multimodal Connector.** We see the trade-off between the performance gain and the extra trainable parameters. We report Acc@0.5 on the detection task.

Models (13B)	VQA v2	MME	LLaVA in the Wild	SeedBench (Image)	Text-only Obj Det
LLaVA	80.0	1531	52.0	68.2	13.4
O3SLM	76.6	1414	48.5	65.5	21.0

Table 5: **Image-only tasks.** Comparison with our baseline LLaVA-v1.5 on the following image only bechmarks: VQAv2 (Goyal et al. 2017), MME (Fu et al. 2025a), LLaVA-in-the-Wild (Liu et al. 2023), SeedBench (Li et al. 2023). We report Acc for text-only object detection. Higher is better.

Emergent Fine-Grained Understanding. We qualitatively demonstrate that our model, **O3SLM**, can perform fine-grained SBIR tasks despite not being explicitly trained to use text descriptions as queries. This capability emerges from extensive multitask pretraining, where the VQA task serves as an auxiliary supervision signal, enhancing the model's ability to capture fine-grained semantics. By complementing the sketch with a textual description—capturing attributes such as color and texture that are typically difficult to convey through hand-drawn sketches—**O3SLM** is able to generalize to more nuanced instruction and incorporate textual understanding even in settings it was not explicitly trained for. As shown in Figure 2, objects can be queried not only through color or texture but also based on their surroundings or their interaction with other objects. We have included more examples in the supplementary materials.

Visual Question Answering. We present qualitative analysis on the VQA task in Figure 2 and the supplementary.

6. Conclusion

In this work, we introduced the novel task of sketch understanding in Large Vision-Language models. To this end, we proposed a novel sketch-based pretraining and visual instruction tuning dataset to incorporate sketch-to-image alignment. This was created using the automated sketch generation pipeline, enabling scalable and diverse training samples. Furthermore, we proposed our model trained on generated data, and through extensive evaluations on tasks like count, SBIR, and detection, we showed the effectiveness of our approach. Our model consistently outperforms existing LVLMs, underscoring the value of sketch-aware training.

Acknowledgments

We gratefully acknowledge Kotak-IISc AI/ML Centre (KIAC) for the generous conference travel grant and the GPU resources that enabled this research.

References

Agrawal, P.; Antoniak, S.; Hanna, E. B.; Chaplot, D.; Chudnovsky, J.; Garg, S.; Gervet, T.; Ghosh, S.; Héliou, A.; Jacob, P.; et al. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073*.

Alaniz, S.; Mancini, M.; Dutta, A.; Marcos, D.; and Akata, Z. 2022. Abstracting sketches through simple primitives. In *European Conference on Computer Vision*, 396–412. Springer.

Arar, E.; Frenkel, Y.; Cohen-Or, D.; Shamir, A.; and Vinker, Y. 2025. Swiftsketch: A diffusion model for image-to-vector sketch generation. In *ACM SIGGRAPH*, 1–12.

Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Zero-shot composed image retrieval with textual inversion. In *International Conference on Computer Vision*, 15338–15347.

Bandyopadhyay, H.; Chowdhury, P. N.; Sain, A.; Koley, S.; Xiang, T.; Bhunia, A. K.; and Song, Y.-Z. 2024. Do Generalised Classifiers Really Work on Human Drawn Sketches? In *European Conference on Computer Vision*, 217–235. Springer.

Banerjee, A.; Mathur, N.; LladÃğs, J.; Pal, U.; and Dutta, A. 2024. SVGCraft: Beyond Single Object Text-to-SVG Synthesis with Comprehensive Canvas Layout. *arXiv preprint arXiv:2404.00412*.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024. Sharegpt4v: Improving large multimodal models with better captions. In *European Conference on Computer Vision*, 370–387. Springer.

Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37: 135062–135093.

Chowdhury, P. N.; Bhunia, A. K.; Sain, A.; Koley, S.; Xiang, T.; and Song, Y.-Z. 2023a. Scenetrilogy: On human scene-sketch and its complementarity with photo and text. In *Computer vision and pattern recognition*, 10972–10983.

Chowdhury, P. N.; Bhunia, A. K.; Sain, A.; Koley, S.; Xiang, T.; and Song, Y.-Z. 2023b. What can human sketches do for object detection? In *Conference on computer vision and pattern recognition*, 15083–15094.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Computer Vision and Pattern Recognition Conference*, 91–104.

Eitz, M.; Hays, J.; and Alexa, M. 2012. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31(4): 1–10.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2025a. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Neural Information Processing Systems Datasets and Benchmarks Track*.

Fu, C.; Wang, G.; Li, J.; Zhang, W.; Lu, R.; and Tang, S. 2025b. ITERATE: Image-Text Enhancement, Retrieval, and Alignment for Transmodal Evolution with LLMs. In *International Conference on Computational Linguistics*, 1365–1376.

Gao, C.; Liu, Q.; Xu, Q.; Wang, L.; Liu, J.; and Zou, C. 2020. Sketchycoco: Image generation from freehand scene sketches. In *Conference on computer vision and pattern recognition*, 5174–5183.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Conference on computer vision and pattern recognition*, 6904–6913.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jongejan, J.; Rowley, H.; Kawashima, T.; Kim, J.; and Fox-Gieg, N. 2016. The quick, draw!-ai experiment. *Mount View, CA, accessed Feb*, 17(2018): 4.

Koley, S.; Bhunia, A. K.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2023. Picture that sketch: Photorealistic image generation from abstract sketches. In *Conference on computer vision and pattern recognition*, 6850–6861.

Koley, S.; Bhunia, A. K.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2024a. How to handle sketch-abstraction in sketch-based image retrieval? In *Conference on Computer Vision and Pattern Recognition*, 16859–16869.

Koley, S.; Bhunia, A. K.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2024b. You'll Never Walk Alone: A Sketch and Text Duet for Fine-Grained Image Retrieval. In *Conference on Computer Vision and Pattern Recognition*, 16509–16519.

Koley, S.; Bhunia, A. K.; Sekhri, D.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2024c. It's All About Your Sketch: Democratising Sketch Control in Diffusion Models. In *Conference on Computer Vision and Pattern Recognition*, 7204–7214.

Koley, S.; Dutta, T. K.; Sain, A.; Chowdhury, P. N.; Bhunia, A. K.; and Song, Y.-Z. 2025a. SketchFusion: Learning Universal Sketch Features through Fusing Foundation Models. In *Computer Vision and Pattern Recognition Conference*, 2556–2567.

Koley, S.; Gajjala, V. R.; Sain, A.; Chowdhury, P. N.; Xiang, T.; Bhunia, A. K.; and Song, Y.-Z. 2025b. SketchYourSeg: Mask-Free Subjective Image Segmentation via Freehand Sketches. *arXiv preprint arXiv:2501.16022*.

Koley, S.; Gajjala, V. R.; Sain, A.; Nath Chowdhury, P.; Xiang, T.; Bhunia, A. K.; and Song, Y.-Z. 2025c. Freestyle Sketch-in-the-Loop Image Segmentation. *arXiv e-prints*, arXiv–2501.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Conference on Computer Vision and Pattern Recognition*, 27831–27840.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7): 1956–1981.

Lee, S.; Yu, S.; Park, J.; Yi, J.; and Yoon, S. 2024. Interactive textto-image retrieval with large language models: A plugand-play approach. In *Association for Computational Linguistics*, volume 1, 11–16.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, M.; Lin, Z.; Mech, R.; Yumer, E.; and Ramanan, D. 2019. Photo-sketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1403–1412. IEEE.

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

Liu, F.-L.; Fu, H.; Wang, X.; Ye, W.; Wan, P.; Zhang, D.; and Gao, L. 2025. SketchVideo: Sketch-based Video Generation and Editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23379–23390.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

Mathur, N.; Marjit, S.; Chaudhuri, A.; and Dutta, A. 2025. CLIPDraw++: Text-to-Sketch Synthesis with Simple Primitives. In *Computer Vision and Pattern Recognition Conference*, 6247–6256.

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

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition*, 19305–19314.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4): 1–12.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *International conference on computer vision*, 8430–8439.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tiwari, A.; Biswas, S.; and Lladós, J. 2024. Sketchgpt: Autoregressive modeling for sketch generation and recognition. In *International Conference on Document Analysis and Recognition*, 421–438. Springer.

Tripathi, A.; Dani, R. R.; Mishra, A.; and Chakraborty, A. 2020. Sketch-guided object localization in natural images. In *European Conference on Computer Vision*, 532–547. Springer.

Tripathi, A.; Mishra, A.; and Chakraborty, A. 2024. Query-guided attention in vision transformers for localizing objects using a single sketch. In *Winter Conference on Applications of Computer Vision*, 1083–1092.

Vinker, Y.; Alaluf, Y.; Cohen-Or, D.; and Shamir, A. 2023. Clipascene: Scene sketching with different types and levels of abstraction. In *International Conference on Computer Vision*, 4146–4156.

Vinker, Y.; Pajouheshgar, E.; Bo, J. Y.; Bachmann, R. C.; Bermano, A. H.; Cohen-Or, D.; Zamir, A.; and Shamir, A. 2022. CLIPasso: Semantically-Aware Object Sketching. *ACM Trans. Graph.*, 41(4).

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xing, X.; Wang, C.; Zhou, H.; Zhang, J.; Yu, Q.; and Xu, D. 2023. Diffsketcher: Text guided vector sketch synthesis through latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 15869–15889.

Xing, X.; Zhou, H.; Wang, C.; Zhang, J.; Xu, D.; and Yu, Q. 2024. Svgdreamer: Text guided svg generation with diffusion model. In *Conference on Computer Vision and Pattern Recognition*, 4546–4555.

Yu, Q.; Liu, F.; Song, Y.-Z.; Xiang, T.; Hospedales, T. M.; and Loy, C.-C. 2016. Sketch me that shoe. In *Computer vision and pattern recognition*, 799–807.

Zhang, H.; Li, H.; Li, F.; Ren, T.; Zou, X.; Liu, S.; Huang, S.; Gao, J.; Leizhang; Li, C.; et al. 2024. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, 19–35. Springer.