FOCUS: Unified Vision-Language Modeling for Interactive Editing Driven by Referential Segmentation

Fan Yang[1,2,3] **Yousong Zhu**[4†], **Xin Li**[2], **Yufei Zhan**[1,3], **Hongyin Zhao**[1], **Shurong Zheng**[1,2,3], **Yaowei Wang**[2], **Ming Tang**[1,3], **Jinqiao Wang**[1,2,3,5†],

[1]Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Haidian District, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]School of Artificial Intelligence, University of Chinese Academy of Science, Beijing, China
[4]School of Artificial Intelligence, China University of Mining and Technology-Beijing, Beijing, China
[5]Wuhan AI Research, Wuhan, China
{yangfan_2022, zhanyufei2021, zhengshurong2023, zhaohongyin2020}@ia.ac.cn
{tangm, jqwang}@nlpr.ia.ac.cn, zhuyousong@cumtb.edu.cn, {lix07, wangyw}@pcl.ac.cn

Abstract

Recent Large Vision Language Models (LVLMs) demonstrate promising capabilities in unifying visual understanding and generative modeling, enabling both accurate content understanding and flexible editing. However, current approaches treat ***"what to see"*** and ***"how to edit"*** separately: they either perform isolated object segmentation or utilize segmentation masks merely as conditional prompts for local edit generation tasks, often relying on multiple disjointed models. To bridge these gaps, we introduce FOCUS, a unified LVLM that integrates segmentation-aware perception and controllable object-centric generation within an end-to-end framework. FOCUS employs a dual-branch visual encoder to simultaneously capture global semantic context and fine-grained spatial details. In addition, we leverage a MoVQGAN-based visual tokenizer to produce discrete visual tokens that enhance generation quality. To enable accurate and controllable image editing, we propose a progressive multi-stage training pipeline, where segmentation masks are jointly optimized and used as spatial condition prompts to guide the diffusion decoder. This strategy aligns visual encoding, segmentation, and generation modules, effectively bridging segmentation-aware perception with fine-grained visual synthesis. Extensive experiments across three core tasks, including multimodal understanding, referring segmentation accuracy, and controllable image generation, demonstrate that FOCUS achieves strong performance by jointly optimizing visual perception and generative capabilities.

1 Introduction

Large Vision-Language Models (LVLMs) are becoming a transformative paradigm in artificial intelligence [45, 2, 38]. Through large-scale pretraining, they unify visual and textual modalities and have achieved remarkable progress across a variety of tasks. These models can jointly process

*is the corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-1123

Figure 1: **FOCUS** enables fine-grained segmentation and editing for both **images** and **videos** through multimodal user instructions. It supports various region specification formats including **clicks**, **scribbles**, **boxes**, and **masks**, and for videos, **annotations on any single frame** suffice to guide full-clip editing. FOCUS can perform detailed region segmentation (e.g., identifying individual people), and supports diverse editing operations such as **removal, replacement**, and **scene transformation** across spatial and temporal domains.

images, videos, and natural language within a single architecture, demonstrating strong performance in visual question answering, image captioning, referring segmentation, and conditional generation [33, 75, 50, 38, 61]. The effectiveness of LVLMs largely stems from their ability to align multimodal semantics and generalize across diverse tasks with minimal task-specific adaptation.

At the same time, image and video generation technologies [47, 7, 46, 85] have advanced rapidly, enabling the synthesis of high-quality visual content and reshaping how we approach artistic creation and visual communication. These developments have driven increasing demands for controllability and fidelity across domains such as art, industry, and education [19]. To meet these demands, some approaches [60, 15, 10] as is shown in fig. 2(a) attempt to combine image generation models with segmentation decoders and text processing modules through modular design. However, such methods often rely on manually engineered pipelines and lack unified modeling and deep feature-level interaction. To address these limitations, others method [30] as in fig. 2(b) leverage the instruction understanding capabilities of LVLMs to dynamically dispatch pretrained expert models through task routing, enabling flexible multi-task adaptation. While these approaches improve tool orchestration, they lack deep cross-modal fusion and joint feature optimization, making it difficult to unify perception and generation effectively. More recent methods[55, 56] as is show in fig. 2(c) have begun to construct unified vision-language frameworks that combine image understanding and generation, using the generalization capability of LVLMs to bridge the semantic gap between high-level understanding and low-level synthesis. However, most of these methods still operate at a coarse level of text-driven control and struggle to support fine-grained editing or object-level manipulation.

To overcome these challenges, we propose FOCUS, an end-to-end LVLM framework that unifies segmentation-aware perception and region-controllable generation under natural language guidance, as is shown in fig. 1. The core of FOCUS features a dual-branch visual encoder, where a CLIP-like or QwenViT-style encoder extracts global semantic representations, while a hierarchical encoder based on ConvNeXt-L focuses on fine-grained local perception. This structure provides stable multi-scale segmentation support and improves adaptability to varying image resolutions. To improve visual generation, FOCUS adopts a VQGAN-based visual tokenizer, inspired by [62, 57, 21], which separately models semantic concepts and texture information. To mitigate the inevitable information loss during quantization, we retain the continuous pre-quantization features from the tokenizer as visual inputs to the language model, enabling fine-grained multimodal understanding. FOCUS is trained through a progressive multi-stage strategy, gradually increasing the input and output resolutions from low to high to ensure stable convergence and final performance. Segmentation

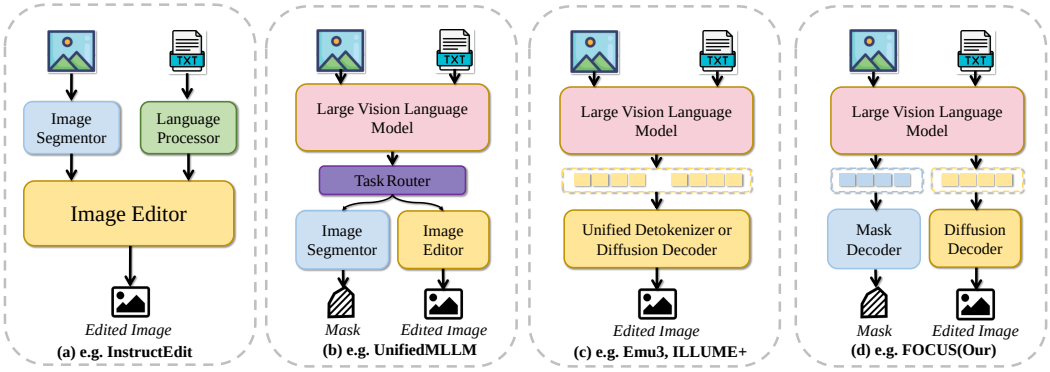

Figure 2: Comparison of controllable image editing paradigms. (a) Modular methods rely on separately trained components. (b) Task-routing LLMs orchestrate existing tools without joint modeling. (c) Unified models combine perception and generation but lack fine-grained control. (d) Our FOCUS jointly models segmentation and generation for precise, region-level editing.

masks are jointly optimized and used as spatial condition prompts to guide a diffusion-based generator for pixel-level editing. During both the multimodal pretraining and instruction tuning stages, we introduce diverse and increasingly complex task distributions, with carefully designed instruction formats, to fully enhance the model's perception understanding and generation capabilities.

Experimental results show that FOCUS achieves significant improvements across three core tasks: multimodal understanding, controllable image generation and editing, and referring segmentation. By jointly modeling pixel-level perception and generation, FOCUS demonstrates higher semantic precision and stronger spatial controllability. Further analysis reveals that pixel-level perception plays a crucial role in bridging the gap between high-level understanding and low-level synthesis. These findings confirm the feasibility and effectiveness of unifying segmentation-aware perception and controllable generation within a single LVLM framework, paving the way for interactive, precise, and generalizable multimodal editing systems.

- We propose FOCUS, a unified large vision language model that integrates pixel-level perception with region-controllable image generation and editing within an end-to-end framework.
- We develop a progressive multi-stage training pipeline that gradually increases image resolution and task complexity. This pipeline enables effective alignment and interaction between the segmentation decoder and the generation module across different scales and modalities.
- Extensive experiments on multimodal understanding, referential segmentation, and controllable image editing demonstrate that FOCUS consistently outperforms existing state-of-the-art models in both visual understanding and generation quality.

2 Related work

The landscape of large vision–language models (LVMs) has rapidly evolved, aiming to unify multimodal understanding and generation across images, videos, and texts. A representative example is the Emu series, which introduces autoregressive modeling to predict the next visual or textual token, thus enabling a generalist interface for diverse multimodal tasks such as image captioning, video event understanding, and cross-modal generation. This contrasts with earlier LVM designs (e.g., BLIP [27], Flamingo [1]) that bridge frozen vision and language models using separate modality connectors, often focusing solely on text prediction and neglecting direct supervision over visual signals. Emu2 [56] further extends this paradigm by positioning itself as an in-context learner, exploiting interleaved video–text data to deliver fine-grained temporal and causal reasoning over multimodal inputs.

While many of these works achieve remarkable integration of multimodal comprehension and generation, they often fall short in interactive and controllable editing. For instance, the LLaVA-Interactive [11] system explores multi-turn, multimodal interactions combining visual chat, seg-

https://doi.org/10.52202/085713-1123

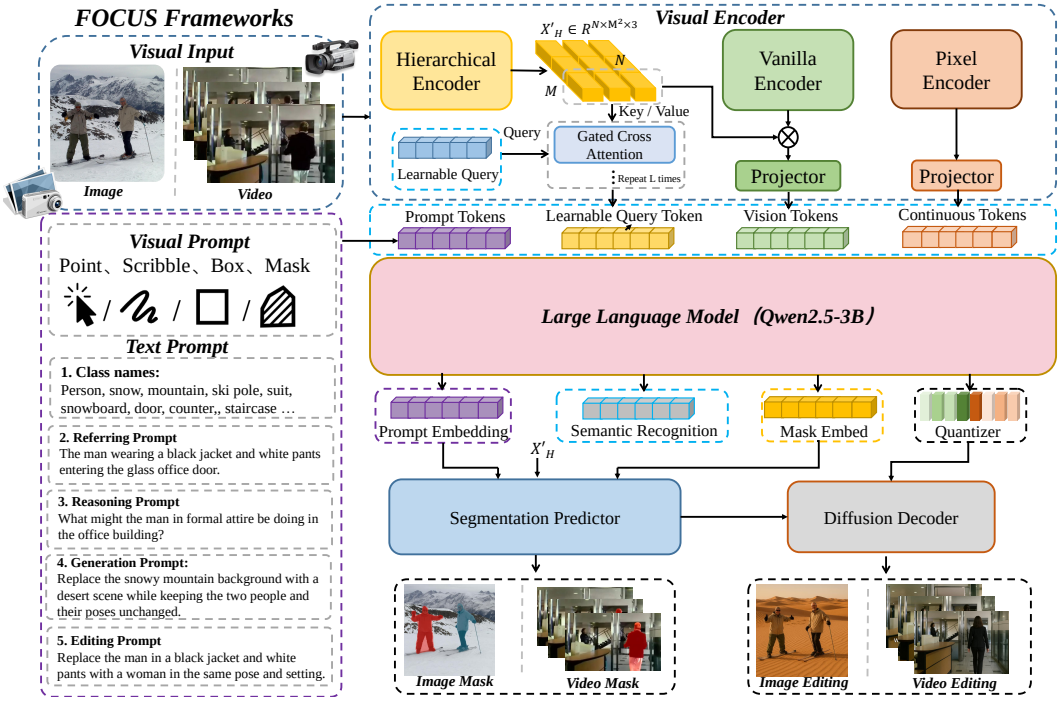

Figure 3: Overview of the FOCUS framework and training pipeline. The left part shows the unified architecture, which integrates a vanilla encoder, a hierarchical encoder, a pixel encoder, a large language model, a segmentation predictor, and a diffusion decoder for fine-grained perception and controllable image generation.

mentation, and editing but largely depends on the synergy of pretrained components without true end-to-end learning. More critically, recent editing frameworks such as InstructEdit emphasize the importance of high-quality segmentation masks by leveraging external Grounded-SAM masks to guide diffusion-based image editing. However, these approaches are not fully end-to-end; they separate mask generation from the editing pipeline, highlighting a major gap in unifying segmentation awareness with controllable generation.

In parallel, dataset-centric efforts like Localized Narratives [48] and Video Localized Narratives [59] have emerged, providing rich multimodal annotations by aligning every word or phrase with specific image or video regions. These resources enable granular referential grounding and enhance the evaluation and training of models that need to capture both spatial and temporal semantics across modalities. Additionally, Describe Anything [31] offers detailed localized captioning, helping bridge the gap between referential understanding and region-specific generation in both image and video domains.

Collectively, these prior works lay the foundation for advancing multimodal understanding, generation, and editing. Yet, there remains an unmet need for a unified architecture that couples segmentation-aware decoding with mask-driven diffusion editing in a fully end-to-end manner. This motivates the development of the proposed framework, which seeks to integrate referential localization, structured mask generation, and controllable editing into a seamless multimodal pipeline.

3 Methodology

As illustrated in Fig.3, FOCUS is a unified large vision-language model that integrates pixel-level perception and controllable image generation in an end-to-end framework. The model comprises four core components, covering the entire process from visual encoding to image synthesis. Section 3.1 introduce a dual-branch visual encoder and a generation-oriented visual tokenizer are employed to extract global semantic features and multi-scale fine-grained representations from inputs at different resolutions, discretizing the visual content into high-level tokens suitable for downstream editing tasks. Next, in the Section 3.2 a large language model based on Qwen2.5 is introduced to unify

multimodal input formats through task prompts, enabling support for a wide range of vision-language tasks. Building on this, in the Section 3.3 a segmentation decoder is designed to perform high-precision fine-grained object segmentation. A diffusion-based image generator then synthesizes high-fidelity images guided by spatial segmentation masks and conditioned on discrete visual tokens in the Section 3.4. In the Section 3.5 further details the progressive training strategy and the construction of multi-source datasets.

3.1 Dual Visual Encoders and Generative Tokenizer

A fundamental challenge in building unified large vision-language models for both image understanding and generation lies in the significant discrepancy between high-level semantic understanding and low-level fine-grained image synthesis. This discrepancy often leads to mutual interference and optimization conflicts during training. To address this issue, we propose a novel architecture that combines a dual-branch visual encoder with a generative visual tokenizer. This design preserves the ability to model global high-level semantic features while enhancing the representation of fine-grained, multi-scale low-level visual details. Meanwhile, the visual tokenizer discretizes continuous visual information, effectively extending the representational capacity of the image generation module.

Specifically, we process each image at two resolutions. A low-resolution image $X_L \in \mathbb{R}^{3\times256\times256}$ is fed into a **Vanilla Encoder** (based on QwenViT [3]) to capture global semantic information. Simultaneously, a high-resolution image $X_H \in \mathbb{R}^{3\times768\times768}$ is encoded by a **Hierarchical Encoder** (ConvNeXt-L [41]) to obtain high-resolution, multi-scale visual features. These two branches are fused using a cross-attention mechanism to enhance low-resolution features with local detail:

$$X'_H = \text{ConvNeXt}(X_H), \quad X'_L = \text{QwenViT}(X_L) \tag{1}$$

$$E'_{\text{img}} = \text{CrossAttn}(X'_L, X'_H), \quad E_{\text{img}} = \text{MLP}(E'_{\text{img}}) + E'_{\text{img}} \tag{2}$$

To further inject fine-grained visual information into the language model, we introduce a gated cross-attention adapter that enhances learnable queries with high-resolution visual features at multiple scales. For the l-th layer query $h^{(l)}$, and j-th scale feature $f^{(j)}_{\text{img}}$, the fusion is formulated as:

$$h^{(l)'} = h^{(l)} + \tanh(\gamma^{(l)}) \cdot \text{CrossAttn}(h^{(l)}, G_p(f^{(j)}_{\text{img}})) \tag{3}$$

$$h^{(l)}_{\text{Adapter}} = h^{(l)'} + \tanh(\beta^{(l)}) \cdot \text{FFN}(h^{(l)'}) \tag{4}$$

where $G_p(\cdot)$ is a projection function aligning the visual feature space, and $\gamma^{(l)}, \beta^{(l)}$ are learnable scaling parameters initialized to zero. This mechanism injects spatial detail while maintaining the efficiency of learnable queries, enhancing both segmentation and generative capabilities.

To enhance the upper bound of image generation quality, we introduce a **generative visual tokenizer** based on the MoVQGAN [84] architecture to discretize visual information. The visual tokenizer serves as a key component for unified autoregressive image generation, but its quantization process inevitably introduces information loss. To address this, we use the continuous features before quantization as the visual input to the large language model (LLM), which provides a more informative representation for fine-grained multimodal understanding. The LLM then outputs discrete visual tokens that guide the diffusion model for high-quality image synthesis.

Specifically, we use the fused dual-branch features E'_{img} as the global semantic representation for generation, which are quantized into discrete semantic tokens and decoded through a lightweight decoder. During the visual tokenizer's pretraining stage, this semantic branch is supervised with a cosine similarity loss against the original encoder features. A downsampling rate of $28\times$ is adopted to align with mainstream vision-language models and focus on high-level semantic concepts.

We further extend the generative visual tokenizer with a **pixel branch**, following the standard MoVQGAN design. This branch uses a $16\times$ downsampling rate to preserve fine textures. After quantization, the semantic and pixel tokens are concatenated along the channel dimension and passed to the decoder for image reconstruction. The pixel branch is trained with a combination of L1 loss, perceptual loss, and adversarial loss. We use large codebooks for both branches, with a size of 32,768 for the semantic branch and 98,304 for the pixel branch.

Together, the generative visual tokenizer and dual-branch visual encoder construct a unified and expressive visual representation that enables **FOCUS** to support both fine-grained perception and controllable image generation within a single framework.

https://doi.org/10.52202/085713-1123

3.2 Large Vision-Language Model and Input Schema

In **FOCUS**, we adopt **Qwen2.5-3B** [71] as the large language model (LLM) to model unified multimodal inputs from the visual encoder and task instructions. The model takes as input a four-tuple $(E_{\text{img}}, X_S, X_T, h^{(l)}_{\text{Adapter}})$, where E_{img} denotes the fused visual features from the dual-branch encoder, X_S represents the structured textual prompts, X_T refers to the continuous pixel-level visual features from the pixel encoder, and $h^{(l)}_{\text{Adapter}}$ is the learnable query feature obtained through multi-scale cross-modal adaptation.

These inputs are fed into the LLM to produce the output representation:

$$E_O = F_{\text{LLM}}(E_{\text{img}}, X_S, X_T, h^{(l)}_{\text{Adapter}})$$

From the LLM output E_O, we extract two key components. First, we introduce a decoding constraint that enforces the model to predict the names of all present objects before generating the mask tokens. This prior step encourages the generation of semantically enriched mask tokens, which incorporate global semantic context from the image and are subsequently used by the segmentation module to produce masks.

In addition, the LLM supports autoregressive prediction of discrete visual tokens. Following a semantic-first strategy, the model first generates semantic tokens to define global content structure, then generates pixel tokens to recover fine visual textures. This two-stage decoding improves the alignment between textual instructions and generated visual content.

Prompt Design. The input prompt schema consists of two components: the *task instruction prompt* S_I and the *condition prompt* S_C. The task instruction prompt specifies the objective of the model in natural language, guiding the LLM to perform segmentation, generation, or editing accordingly. For example, in class-based segmentation tasks such as panoptic, open-vocabulary, or video instance segmentation, the instruction can be *"Please segment all the positive objects by the following candidate categories."* In referring or reasoning segmentation tasks (e.g., RES [23, 76, 44, 34], R-VOS [28], ReasonVOS [70]), the instruction becomes *"Please segment the target referred to by the language description."* For visual-guided tasks such as interactive or video object segmentation, the instruction is phrased as "Please segment according to the given visual reference regions."

The condition prompt provides additional information specific to each task. In class-based segmentation, it lists the candidate category labels; in referring segmentation, it provides a natural language expression; in visual-guided settings, it includes pooled region features sampled from CLIP embeddings at specified coordinates. For image generation tasks, the condition prompt consists of a text description such as *"a cat sitting on a couch,"* while for image editing tasks, it contains language-based or spatial referring expressions like *"replace the person on the right with a dog."* The condition prompt not only supplies contextual information but also serves as an implicit classifier for category-aware mask prediction.

3.3 Segmentation Module

The segmentation predictor F_p takes three types of inputs: task-specific prompt embeddings $\{E^k_P\}^K_{k=1}$, semantic-enhanced mask token embeddings $\{E^j_Q\}^N_{j=1}$ from the LLM output, and multi-scale visual features $f_{\text{img}} = X'_H$ from the visual encoder. Here, K denotes the number of candidate categories, and N represents the number of mask proposals. These inputs are fused to predict segmentation outputs in the following form:

$$\{(m_j, z_j, e_j)\}^N_{j=1} = F_p\left(\{E^k_P\}^K_{k=1}, \{E^j_Q\}^N_{j=1}, f_{\text{img}}\right), \tag{5}$$

where $m_j \in \mathbb{R}^{H\times W}$ denotes the j-th predicted binary mask, $z_j \in \mathbb{R}^K$ is the associated category score vector, and $e_j \in \mathbb{R}^D$ is an optional instance-level embedding produced by an auxiliary embedding head to enable temporal association in video segmentation tasks.

For video tasks, we adopt a frame-by-frame processing strategy to generate frame-level segmentation results, ensuring efficient training and inference while maintaining temporal consistency.

To support multi-scale supervision in the subsequent image generation stage, the segmentation module consistently produces fixed-resolution masks based on X'_H, providing stable spatial guidance. These

masks are then rescaled to match the resolution used by the diffusion model, enabling effective alignment between segmentation outputs and the image synthesis pipeline.

3.4 Diffusion Decoder for Controllable Generation

To enable high-quality and region-controllable image generation, FOCUS incorporates a latent diffusion decoder initialized from SDXL. The generation process is formulated as denoising over latent variables, conditioned on structured visual prompts. We utilize semantic tokens z_{sem} and pixel-level tokens z_{pix} generated by the large language model. These tokens are mapped into continuous embeddings via a learned codebook and concatenated with noisy latents before being passed into a UNet-based denoising backbone.

To achieve spatial control, we leverage predicted segmentation masks $m_j \in \mathbb{R}^{H \times W}$ from the segmentation module. These masks are first downsampled to match the latent spatial resolution, yielding $\tilde{m}_j \in \mathbb{R}^{H' \times W'}$, where $H' \times W'$ denotes the latent feature resolution. The downsampled mask is then flattened and passed through a linear projection layer to obtain a spatial guidance sequence $f_m \in \mathbb{R}^{L \times C}$, where $L = H' \times W'$ and C is the channel dimension. This sequence is injected into the UNet through cross-attention at intermediate layers to guide regional generation. The interaction is formalized as:

$$\hat{z}_t = \text{CrossAttn}(\phi(z_t), f_m), \quad z_{t+1} = \text{UNet}(\hat{z}_t), \tag{6}$$

where $\phi(z_t)$ denotes the projected latent features at denoising timestep t, and f_m provides the spatial condition derived from $\tilde{m}_j$.

We also experimented with using mask token embeddings $\{E_Q^j\}$ as alternative conditioning inputs, but found that direct spatial control through explicit masks yields better localization and more faithful region editing in our experiments.

3.5 Training Procedure and Data Composition

To support unified pixel-level understanding, high-level multimodal reasoning, and spatially controllable image generation, FOCUS adopts a progressive four-stage training paradigm that incrementally builds visual tokenization, vision-language alignment, and conditional generation capabilities.

Stage 1: Pretraining of Dual-Branch Visual Tokenizer and Diffusion Decoder. We first train a semantic branch and a pixel branch to discretize input images into semantic and texture tokens using SimVQ quantizers. A progressive resolution strategy is adopted, starting from 256×256 and gradually increasing to 512×512. A large-scale image-text dataset of approximately 58M pairs is constructed based on resolution constraints and visual diversity. On top of this, we pretrain a latent diffusion decoder using a 10M image subset, enabling high-fidelity reconstruction from discrete tokens. This stage focuses purely on image modeling and does not involve segmentation or controllable generation.

Stage 2: Visual-Language Adapter Warmup. To align visual features with the input space of the large language model (LLM), we train projection heads for both the vanilla encoder and the pixel encoder, along with learnable queries and gated cross-attention modules in the hierarchical encoder branch. All visual backbone encoders are kept frozen. Training is conducted at 256×256 resolution with standard language modeling loss.

Stage 3: Multimodal Pretraining with Segmentation-Aware Alignment. In this stage, we jointly train the LLM, visual adapters, and the mask decoder to improve the model's perception of segmentation structures and enhance multimodal generation capability. Training is performed in two phases with input resolutions of 256×256 and 512×512. Supervision includes token-level cross-entropy loss, segmentation mask supervision (Dice + BCE), and image reconstruction loss (L2) from diffusion outputs. The training corpus includes multimodal data spanning image-text pairs, segmentation tasks, and vision-language reasoning. Full dataset details are presented in appendix.

Stage 4: Instruction Tuning for Region-Controlled Editing. This stage introduces region-level controllability for editing and generation tasks. We jointly train the LLM, the mask decoder, and

 https://doi.org/10.52202/085713-1123

cross-attention layers within the diffusion model, while freezing the visual encoders, visual tokenizer, and VAE components. Input images are resized using a bucketed aspect ratio cropping strategy with total pixel counts ranging from 512^2 to 1024^2. Supervision includes token prediction (cross-entropy), segmentation mask prediction (Dice + BCE), and image reconstruction (L2 loss). Detailed task and dataset configurations are summarized in appendix.

4 Experiments

We conduct comprehensive experiments to evaluate the performance of FOCUS across three representative tasks: multimodal understanding, referential segmentation, and generation and controllable editing. All experiments are designed to validate the model's ability to unify perception and generation in an end-to-end framework, with strong alignment to natural language instructions, fine-grained localization, and object-aware editing fidelity.

4.1 Implementation Details

In our experiments, we adopt Qwen2.5 as the backbone large language model (LLM). The visual encoder consists of two branches: the semantic decoder is composed of four attention blocks with 2D relative positional encoding (2D-RoPE), while the pixel encoder and decoder follow a MoVQGAN-based architecture with base channel dimensions of 128 and 384, respectively. The codebook size is set to 32,768 for the semantic branch and 98,304 for the pixel branch, with both using a code dimension of 32. We employ the AdamW optimizer without weight decay and use a constant learning rate across the visual encoder, diffusion decoder, and the large vision language model. Detailed training hyperparameters for each component are summarized in the supplementary materials. The training of the Dual-Branch Visual Tokenizer and the diffusion decoder each took approximately 3 days on a computing cluster, while the 3B-parameter MLLM required around 13 days to complete the three-stage training process.

Table 1: FOCUS performance on general and document-oriented benchmarks.

Method	LLM	General							Doc				
		POPE	MMBench	SEED	MME-P	MM-Vet	MMMU	AI2D	VQA-text	ChartQA	DocVQA	InfoVQA	OCRBench
Understanding Only													
InstructBLIP [16]	Vicuna-7B [33]	-	36.0	53.4	-	26.2	30.6	33.8	50.1	12.5	13.9	-	276
Qwen-VL-Chat [3]	Qwen-7B	-	60.6	58.2	1487.5	-	35.9	45.9	61.5	66.3	62.6	-	488
LLaVA-1.5 [38]	Vicuna-7B	85.9	64.3	58.6	1510.7	31.1	35.4	54.8	58.2	18.2	28.1	25.8	318
ShareGPT4V [9]	Vicuna-7B	-	68.8	69.7	1567.4	37.6	37.2	58	60.4	21.3	-	-	371
LLaVA-NeXT [25]	Vicuna-7B	86.5	67.4	64.7	1519	43.9	35.1	66.6	64.9	54.8	74.4	37.1	532
Emu3-Chat [62]	8B from scratch	85.2	58.2	68.2	-	37.2	31.6	70.0	64.7	68.6	76.3	43.8	687
Unify Understanding and Generation													
Unified-IO [43]	6.8B from scratch	87.7	-	61.8	-	-	-	-	-	-	-	-	-
Chameleon [57]	7B from scratch	-	-	-	-	8.3	22.4	-	-	-	-	-	-
LWM [28]	LLaVA-2-7B	75.2	-	-	-	9.6	-	-	18.8	-	-	-	-
Show-o [68]	Phi-1.5B	73.8	-	-	948.4	-	25.1	-	-	-	-	-	-
VILA-U (256) [67]	LLaMA-2-7B	83.9	-	56.3	1336.2	27.7	-	-	48.3	-	-	-	-
VILA-U (384) [67]	LLaMA-2-7B	85.8	-	59	1401.8	33.5	-	-	60.8	-	-	-	-
Janus [66]	DeepSeek-LLM-1.3B	87.0	69.4	63.7	1338.0	34.3	30.5	-	-	-	-	-	-
Janus-Pro-1B [13]	DeepSeek-LLM-1.3B	86.2	75.5	68.3	1444.0	39.8	36.3	-	-	-	-	-	-
Janus-Pro-7B [13]	DeepSeek-LLM-7B	87.4	79.2	72.1	1567.1	50.0	41.0	-	-	-	-	-	-
ILLUME+ [21]	Qwen2.5-3B	87.6	80.8	73.3	1414.0	40.3	44.3	74.2	69.9	69.9	80.8	44.1	672
FOCUS	Qwen2.5-3B	**88.0**	**81.5**	**73.9**	**1570.3**	**50.2**	**44.9**	**74.5**	**70.4**	**70.3**	**81.1**	**44.3**	678

4.2 Multimodal understanding

To evaluate the multimodal understanding capabilities of our model, we conduct systematic evaluations on two categories of widely-used benchmarks, as is show in table 1: (1) General benchmarks, including POPE, MMBench, SEED, MME-P, MM-Vet, MMMU, and AI2D; and (2) Document-oriented benchmarks, including VQA-text, ChartQA, DocVQA, InfoVQA, and OCRBench. Experimental results show that, despite using only a 3B-parameter model, FOCUS achieves performance comparable to state-of-the-art unified models such as Janus-Pro-7B and ILLUME-7B, and notably outperforms ILLUME+ with the same parameter scale. This performance gain is largely attributed to the incorporation of multi-scale high-resolution features and segmentation masks, which significantly enhance pixel-level perception.

4.3 Controllable Generation and Editing

Multimodal image generation. To evaluate the multimodal visual generation capability, we use the MJHQ-30K, GenAI-bench and GenEval benchmarks in the table 2. For MJHQ-30K, we adopt the

Fréchet Inception Distance (FID) metric on 30K generated images compared to 30K high-quality images, measuring the generation quality and diversity. GenAI-bench and GenEval are challenging text-to-image generation benchmarks designed to reflect the consistency between text descriptions and generated images. We compare FOCUS with previous state-ofthe-art multimodal generation-only and unified models. This highlights the superior generation quality and diversity enabled by our diffusion-based approach. Additionally, FOCUS achieves competitive results on the GenAI-bench and GenEval benchmarks and attains the highest accuracy in advanced categories on GenAI-bench, demonstrating its ability to understand and generate images from complex text descriptions. Figure 7 shows more results of FOCUS on generating flexible resolution images.

Table 2: **Evaluation results on multimodal image generation benchmarks.**

Method	**Params.**	**Type**	***MJHQ30K***	***GenAI-bench***		***GenEval***						
			FID	Basic	Adv.	Overall	Single Obj	Two Obj.	Counting	Colors	Position	Color Attri.
Generation Only												
SDv1.5 [51]	0.9B	Diffusion	-	-	-	0.43	0.97	0.38	0.35	0.76	0.04	0.06
PixArt-α [7]	0.6B	Diffusion	6.14	-	-	0.48	0.98	0.45	0.44	0.08	0.07	0.07
SDXL [47]	2.6B	Diffusion	9.55	**0.83**	0.63	0.55	0.98	0.41	0.48	0.15	0.17	0.23
Emu3-Gen [62]	8B	Autoreg.	-	-	-	0.54	0.98	0.71	0.34	0.81	0.17	0.21
Unify Understanding and Generation												
Chameleon [57]	7B	Autoreg.	-	-	-	0.39	-	-	-	-	-	-
LWM [35]	7B	Autoreg.	17.77	0.63	0.53	0.47	0.93	0.41	0.46	0.79	0.09	0.15
Show-o [68]	1.5B	Autoreg.	15.18	0.70	0.60	0.45	0.95	0.52	0.49	0.83	0.28	0.30
VILA-U(256) [67]	7B	Autoreg.	12.81	0.72	0.64	-	-	-	-	-	-	-
VILA-U(384) [67]	7B	Autoreg.	7.69	0.71	0.66	-	-	-	-	-	-	-
Janus [66]	7B	Autoreg.	10.1	-	-	-	-	-	-	-	-	-
Janus-Pro-1B [13]	1.3B	Autoreg.	-	0.73	0.68	0.61	0.99	0.82	0.51	0.86	0.39	0.26
Janus-Pro-7B [13]	7B	Autoreg.	-	0.80	0.69	0.59	**0.90**	0.59	**0.90**	**0.79**	0.66	0.25
ILLUME+ [21]	3B	Autoreg.	6.00	0.72	0.71	0.72	1.00	**0.99**	0.88	0.62	**0.84**	0.53
FOCUS	3B	Autoreg.	**6.05**	**0.83**	**0.72**	**0.75**	**1.20**	0.98	0.87	0.66	0.81	**0.57**

Multimodal image editing. To assess the multimodal image editing capability of our method, we evaluate it on the Emu Edit benchmark and report the CLIP-I, CLIP-T, CLIP-DIR and DINO scores. The CLIP-I and DINO scores measure the model's ability to preserve elements from the source image, while the CLIP-T and CLIP-DIR score measures the consistency between the output image and the target caption. As illustrated in the table 4 , our model demonstrates strong performance in image editing tasks, surpassing specialized models, particularly in the CLIP-T metric. This indicates that the unified model's superior understanding enhances its ability to interpret editing instructions, resulting in more precise modifications.

Table 3: **Comparisons with other referring segmentation.**

Method	RefCOCO			Refcoco+			Refcocog		gRefCOCO		
	Val	testA	testB	Val	testA	testB	Val	Test	Val	testA	testB
Segmentation Specialist											
CRIS [80]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	55.3	63.8	51.0
LAVT [74]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	57.6	65.3	55.0
PolyFormer-B [39]	74.8	76.6	71.1	67.6	72.9	58.3	67.8	69.1	-	-	-
LVLM-based Segmentation Network											
LISA-7B [24]	74.9	79.1	72.3	65.1	70.9	58.1	67.9	70.6	38.7	52.6	44.8
PixelLM7B [50]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	-	-
PSALM [82]	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	42.0	52.4	50.6
HyperSeg [64]	**84.8**	85.7	**83.4**	**79.0**	83.5	**75.2**	**79.4**	78.9	47.5	57.3	52.5
FOCUS	84.1	**86.3**	82.7	78.5	**84.1**	74.3	79.3	**79.8**	**48.7**	**58.5**	**53.0**

Table 4: **Quantitative results on image editing benchmarks.** The performance with top-1 and top-2 value are denoted in bold and underline.

Method	Type	Tasks	*Emu Edit [53]*			
			DINO	CLIP-I	CLIP-T	CLIP-DIR
InstructPix2Pix [1]	Diffusion	Edit only	0.762	0.834	0.219	0.078
MagicBrush [81]	Diffusion	Edit only	0.776	0.838	0.222	0.09
OmniGen [55]	Diffusion	Edit only	0.804	0.836	0.233	-
Emu Edit [53]	Diffusion	Edit only	0.819	0.859	0.231	**0.109**
PUMA [18]	AR	Edit only	0.785	0.846	0.270	-
ILLUME	AR	Und, Gen, Edit	0.791	**0.879**	0.260	-
ILLUME+	AR	Und, Gen, Edit	0.826	0.872	0.275	0.101
FOCUS	AR	Und, Gen, Edit	**0.831**	0.876	**0.278**	0.105

4.4 Referring Segmentation Accuracy

We evaluate the referential segmentation performance of **FOCUS** on four standard benchmarks: RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO, using mean Intersection-over-Union (mIoU) as the evaluation metric. As shown in table 3, **FOCUS** achieves competitive or superior performance compared to both segmentation-specific and LVLM-based methods. These results demonstrate the effectiveness of **FOCUS** in pixel-level target localization and its strong capacity to align complex referring expressions with visual semantics in an end-to-end framework. More empirical evidence is presented in the appendix.

4.5 Ablation Studies

4.5.1 The effect of multi-stage training

The following table shows the results of using different image resolutions (256, 512, and 1024) for all-stage task training. We observe a clear trend: higher image resolutions improve the model's performance in image generation and segmentation, but lead to a decline in language understanding.

https://doi.org/10.52202/085713-1123

Table 5: The effect of different image resolutions and multi-stage training strategy on model performance across various tasks. ↑ indicates higher is better, ↓ indicates lower is better.

Image Size	MJHQ30K	GenAI-Bench		Image Understanding			RefCOCO		
	FID(↓)	Basic(↑)	Adv.(↑)	POPE(↑)	MMB(↑)	SEED(↑)	testA(↑)	TestB(↑)	Valid(↑)
256	11.85	0.63	0.56	85.3	73.1	70.1	78.5	79.2	75.4
512	7.39	0.70	0.61	87.3	80.6	71.3	81.3	82.4	79.8
1024	6.03	0.76	0.69	77.6(↓)	68.2(↓)	68.8(↓)	83.4	84.9	81.5
256->1024(Our)	6.05	0.83	0.72	88.0	81.5	73.9	84.1	86.3	82.7

Our analysis suggests that this decline is mainly due to an increase in hallucination issues at higher resolutions—when processing high-resolution images, the model is more likely to generate text that is inconsistent with the visual input. This increased hallucination negatively impacts the model's language comprehension, highlighting the challenge of balancing visual precision and robust language understanding in multimodal models.

To address this, our multi-stage training strategy progressively increases image resolution and task complexity, starting from simpler tasks at lower resolutions. This approach enables strong and balanced performance across all tasks.

4.5.2 The effect of multi-scale features in Gated Cross-Attention

Table 6: Ablation study on the effect of multi-scale feature fusion in Gated Cross-Attention.

Model Variant	Scales	RefCOCO mIoU(↑)	MMBench Acc.(↑)	CLIP-T(Edit)(↑)
FOCUS-SingleScale (Fine)	1	79.5	72.8	0.265
FOCUS-SingleScale (Coarse)	1	78.2	73.9	0.260
FOCUS-DualScale	2	80.5	73.5	0.270
FOCUS (Ours)	3	81.4	73.9	0.275

To effectively fuse multi-scale features, our model employs a Gated Cross-Attention module to process outputs from the ConvNeXt-L backbone, achieving a balance between fine-grained detail and global contextual understanding. Specifically, we design a 3-layer Gated Cross-Attention adapter that hierarchically integrates three distinct feature scales: the first layer focuses on fine features (f_2), the second on mid-level features (f_3), and the third on coarse semantic features (f_4).

As illustrated in the table, we compare our three-scale model (f_2, f_3, f_4) with variants that utilize only a single scale (f_4), dual scales (f_3, f_4), and all four scales (f_2-f_4). The results (mIoU on RefCOCO) demonstrate that our multi-scale fusion strategy is essential for accurately segmenting fine details, providing the optimal balance between performance and computational efficiency.

4.5.3 The effect of continuous visual tokens

Table 7: Ablation study on the effect of continuous visual token input on model performance.

Continuous Input	MJHQ30K	GenAI-Bench		Image Understanding		
	FID↓	Basic↑	Adv.↑	POPE↑	MMB↑	SEED↑
✗	6.85	0.68	0.65	82.1	50.4	56.0
✓	6.05	0.83	0.72	85.3	70.9	66.6

Our ablation study demonstrates that feeding continuous visual features directly into the LLM is essential for optimal performance. By bypassing discrete tokenization, this approach effectively reduces information loss and retains fine-grained visual details. As a result, the LLM receives a richer and more precise visual representation, which significantly enhances its ability to comprehend complex multimodal scenarios and accurately interpret referential instructions.

5 Conclusion

FOCUS demonstrates robust generalization and strong task performance across multimodal dialogue, referential segmentation, and controlled visual editing. The consistent improvements over baseline models affirm the effectiveness of our unified architecture, which tightly integrates segmentation-aware perception with instruction-guided generation. These results further support the practical value of FOCUS in real-world multimodal interaction scenarios.

6 Acknowledgement

This work was supported by the National Key Research and Development Program of China (No. 2024YFE0210600), the National Key R&D Program of China under Grant No. 2022ZD0160601, the National Natural Science Foundation of China under Grant Nos. 62176254 and 62276260, and the Beijing Natural Science Foundation (No. L247028).

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.

[2] Anthropic. Claude 3: Next-generation ai models. https://www.anthropic.com/news/claude-3, 2024. Accessed: 2025-05-12.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

[5] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[6] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation, 2018. URL https://arxiv.org/abs/1803.00557.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. URL https://arxiv.org/abs/2310.00426.

[8] Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, Dingdong Wang, Kun Xiang, Haoyuan Li, Haoli Bai, Jianhua Han, Xiaohui Li, Weike Jin, Nian Xie, Yu Zhang, James T. Kwok, Hengshuang Zhao, Xiaodan Liang, Dit-Yan Yeung, Xiao Chen, Zhenguo Li, Wei Zhang, Qun Liu, Jun Yao, Lanqing Hong, Lu Hou, and Hang Xu. Emova: Empowering language models to see, hear and speak with vivid emotions, 2025. URL https://arxiv.org/abs/2409.18042.

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. URL https://arxiv.org/abs/2311.12793.

[10] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing, 2023. URL https://arxiv.org/abs/2311.00571.

[11] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*, 2023.

[12] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014.

[13] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811.

[14] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022. URL https://arxiv.org/abs/2112.01527.

[15] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. URL https://arxiv.org/abs/2210.11427.

 https://doi.org/10.52202/085713-1123

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

[18] Rongyao Fang, Chengqi Duan, Kun Wang, Hao Li, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Hongsheng Li, and Xihui Liu. Puma: Empowering unified mllm with multi-granular visual generation, 2024. URL https://arxiv.org/abs/2410.13861.

[19] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving, 2024. URL https://arxiv.org/abs/2309.17452.

[20] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022.

[21] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, and Hang Xu. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement, 2025. URL https://arxiv.org/abs/2504.01934.

[22] Siming Huang, Tianhao Cheng, J. K. Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J. Yang, Jiaheng Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Zhaoxiang Zhang, Jie Fu, Qian Liu, Ge Zhang, Zili Wang, Yuan Qi, Yinghui Xu, and Wei Chu. Opencoder: The open cookbook for top-tier code large language models, 2025. URL https://arxiv.org/abs/2411.04905.

[23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Jan 2014. doi: 10.3115/v1/d14-1086. URL http://dx.doi.org/10.3115/v1/d14-1086.

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

[26] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. URL https://arxiv.org/abs/2307.16125.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

[28] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Towards robust referring video object segmentation with cyclic relational consensus, 2023. URL https://arxiv.org/abs/2207.01203.

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. URL https://arxiv.org/abs/2305.10355.

[30] Zhaowei Li, Wei Wang, YiQing Cai, Xu Qi, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model, 2024. URL https://arxiv.org/abs/2408.02503.

[31] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and Yin Cui. Describe anything: Detailed localized image and video captioning, 2025. URL https://arxiv.org/abs/2504.16072.

[32] Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/datasets/Open-Orca/OpenOrca, 2023.

[33] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[34] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023.

[35] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2025. URL https://arxiv.org/abs/2402.08268.

[36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[39] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023.

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

[41] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[42] Dakuan Lu, Xiaoyu Tan, Rui Xu, Tianchu Yao, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. Scp-116k: A high-quality problem-solution dataset and a generalized pipeline for automated extraction in the higher education science domain, 2025. URL https://arxiv.org/abs/2501.15587.

[43] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks, 2022. URL https://arxiv.org/abs/2206.08916.

[44] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.9. URL http://dx.doi.org/10.1109/cvpr.2016.9.

[45] OpenAI. Chatgpt: Language model (gpt-4). https://chat.openai.com, 2023. Accessed: 2025-05-12.

[46] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

[47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

[48] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives, 2020. URL https://arxiv.org/abs/1912.03098.

[49] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.

[50] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023.

 https://doi.org/10.52202/085713-1123

[51] 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 (CVPR)*, pages 10684–10695, June 2022.

[52] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 208–223, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58555-6.

[53] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks, 2023. URL https://arxiv.org/abs/2311.10089.

[54] Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing, 2024. URL https://arxiv.org/abs/2411.06686.

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

[56] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[57] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

[58] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

[59] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives, 2023. URL https://arxiv.org/abs/2302.11217.

[60] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions, 2023. URL https://arxiv.org/abs/2305.18047.

[61] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. Nov 2023.

[62] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

[63] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. On the de-duplication of laion-2b, 2023. URL https://arxiv.org/abs/2303.12733.

[64] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model, 2024. URL https://arxiv.org/abs/2411.17606.

[65] Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision, 2025. URL https://arxiv.org/abs/2411.07199.

[66] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2410.13848.

[67] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2025. URL https://arxiv.org/abs/2409.04429.

[68] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2408.12528.

[69] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing, 2024. URL https://arxiv.org/abs/2406.08464.

[70] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models, 2024. URL https://arxiv.org/abs/2407.11325.

[71] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[72] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation, 2019. URL https://arxiv.org/abs/1905.04804.

[73] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023.

[74] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[75] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. Oct 2023.

[76] 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.

[77] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024.

[78] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. URL https://arxiv.org/abs/2308.02490.

[79] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

[80] Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi, and Paul Allen. From recognition to cognition: Visual commonsense reasoning.

[81] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. URL https://arxiv.org/abs/2306.10012.

[82] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model, 2024. URL https://arxiv.org/abs/2403.14598.

[83] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale, 2024. URL https://arxiv.org/abs/2407.05282.

[84] Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation, 2022. URL https://arxiv.org/abs/2209.09002.

[85] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

[86] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. URL https://arxiv.org/abs/2304.10592.

 https://doi.org/10.52202/085713-1123

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline the primary contributions of the paper, including the proposal of a unified end-to-end large vision language model for pixel-level perception and controllable generation. These claims are consistently supported by the theoretical design and experimental results presented in the main body of the paper, ensuring alignment between stated contributions and actual findings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations" section that reflects on several key constraints of the proposed approach, such as limited generalization to out-of-distribution visual domains and the computational cost associated with joint training of segmentation and generation modules. Additionally, the authors discuss the assumptions made during pretraining and inference, and acknowledge potential issues in real-world deployment scenarios involving ambiguous or complex instructions.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification:The paper provides detailed information on the experimental setup, including model architecture, training hyperparameters, dataset splits, evaluation protocols, and hardware configurations. All critical implementation details that affect the main claims are included in the main text or appendix. This ensures that researchers can reproduce the results even without access to the code or pretrained models.

   Guidelines:

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

 https://doi.org/10.52202/085713-1123

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide partial core code in the supplemental material to demonstrate the openness and transparency of our method. Full code, pretrained models, and detailed instructions for reproducing the main results will be released after paper acceptance to ensure compliance with anonymity requirements during the review process.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper provides comprehensive experimental details, including dataset splits, model architecture configurations, training hyperparameters, optimizer types, learning rate schedules, and input resolutions. These are specified in the main text and appendix to ensure clarity and reproducibility. The rationale behind hyperparameter selection is also explained based on validation performance.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [NA]

   Justification: While the paper includes extensive quantitative results and comparisons, we do not report statistical significance metrics such as error bars or confidence intervals, as the primary experiments are deterministic and evaluated on standardized benchmarks. Variance analysis is left for future work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We report the GPU type (e.g., A100), memory size, number of training hours, and cluster environment used for each major experiment in the appendix. This ensures that others can estimate the required compute for reproduction.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

https://doi.org/10.52202/085713-1123

Justification: The research fully adheres to the NeurIPS Code of Ethics. All data used are publicly available or appropriately licensed, no personally identifiable information is involved, and no human subjects were used. We have taken care to ensure fairness, transparency, and reproducibility throughout the study, and all content has been anonymized in accordance with double-blind review requirements.

Guidelines:

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

10. **Broader impacts**

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

Answer: [NA]

Justification: The work focuses on developing foundational methods for multimodal vision–language modeling and does not directly target downstream applications with immediate societal implications. As such, potential societal impacts—whether positive or negative—are not discussed in the paper. We consider this research as a building block for future systems, and a more thorough societal impact assessment would be appropriate at the deployment stage.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer:[NA]

Justification: The paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: we cite the original paper that produced the code package or dataset

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/085713-1123

Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use an existing large language model (LLM) as part of our framework to support downstream vision tasks such as segmentation and image-conditioned generation. However, we do not modify the architecture or pretraining of the LLM itself. All fine-tuning is conducted on publicly available datasets, and the LLM is treated as a fixed backbone to support multimodal learning.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Dual Quantizer and Diffusion Decoder Training

To enable high-quality image generation, FOCUS pretrains dual quantizers with a three-branch visual encoder and subsequently trains a diffusion decoder conditioned on the tokens. This section details the training process for both the quantizers and the diffusion decoder.

Dual Quantizer Training. As illustrated in Fig. 4(a), the semantic quantizer is constructed from two visual branches: a *vanilla encoder* that captures global semantics from low-resolution inputs,

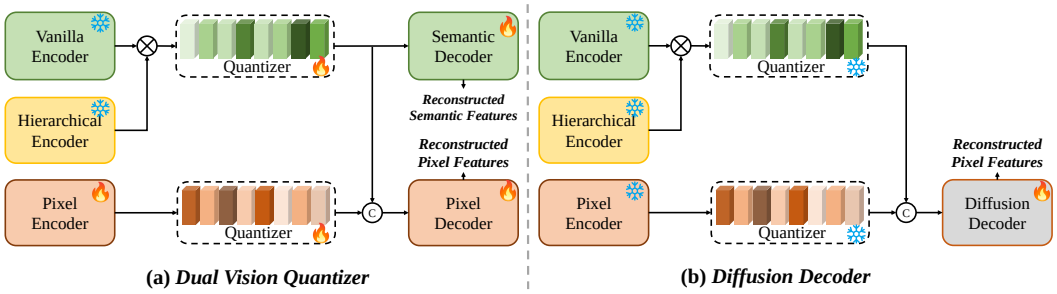

Figure 4: (a) Dual Vision Quantizer: Two separate branches encode semantic and pixel information via quantization and reconstruction. (b) Diffusion Decoder: Reuses pretrained tokens to condition UNet-based denoising for image reconstruction.

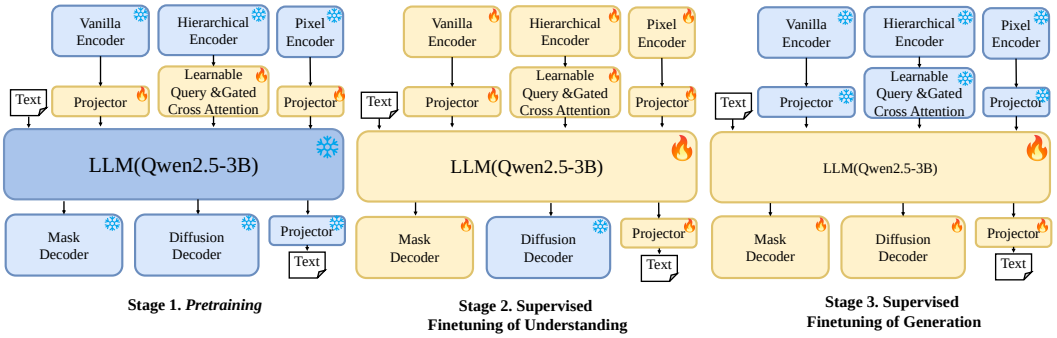

Figure 5: Three-stage progressive training strategy of FOCUS. Frozen modules are shown in blue, trainable modules in orange. Each stage incrementally activates relevant components to align semantic perception and generative editing.

and a *hierarchical encoder* that extracts high-resolution spatial features. These two streams cross attention to provide rich semantic representations, which are quantized via a SimVQ [86] module and reconstructed using a lightweight transformer as the semantic decoder.

In parallel, a dedicated *pixel encoder* processes high-resolution images to extract fine-grained, low-level textures. Following the MoVQGAN design [84], its features are quantized and decoded via a pixel-level decoder. This branch is supervised using L1, perceptual, and adversarial losses.

Each quantizer maintains a separate codebook: 32,768 entries for semantic tokens and 98,304 entries for pixel tokens as [21]. The quantizer is trained progressively from 256×256 to 512×512 resolution using a bucketed batching strategy, and optimized on a corpus of 45M diverse image-text pairs.

Diffusion decoder training. After quantizer training, we train a UNet-based diffusion decoder, initialized from SDXL [47], to reconstruct high-resolution images conditioned on the learned discrete tokens. As shown in Fig. 4(b), semantic and pixel tokens are embedded into continuous representations and concatenated with noisy latents to guide the denoising process.

To accommodate diverse image shapes, 11 aspect-ratio-specific canvas sizes are predefined: {1:1, 3:4, 4:3, 2:3, 3:2, 1:2, 2:1, 1:3, 3:1, 1:4, 4:1} [21], and samples with more than 20% content loss after cropping are discarded. The training proceeds in two stages: the first at 512×512 resolution and the second at 1024×1024 for super-resolution refinement. During this stage, all encoders and codebooks are frozen, and only the diffusion decoder is updated.

B Progressive Large Vision Language Model Training

A core challenge in building a unified large vision language model lies in the optimization conflict between high-level semantic understanding in language and low-level visual synthesis in image generation. FOCUS addresses this by adopting a progressive three-stage training paradigm (Fig. 5)

https://doi.org/10.52202/085713-1123

that incrementally activates model components to align pixel-level perception and controllable image generation.

Visual-Language Projector Warmup. This stage establishes initial alignment between visual features and the large language model (LLM). We train only the projection heads of the vanilla and pixel encoders, along with the learnable queries and gated cross-attention modules in the hierarchical encoder. All backbone encoders, the LLM, the mask decoder is Mask2Former [14], and the diffusion decoder are frozen. Training is performed on 256×256 image-text pairs using a standard language modeling objective, without any segmentation or generation supervision.

Multimodal Pretraining with Segmentation-Aware Alignment. We activate and jointly train the LLM, visual-language adapters, and the mask decoder to enhance multimodal understanding and segmentation capability. The diffusion decoder remains frozen. Supervision includes token-level cross-entropy, segmentation losses (Dice and BCE), and visual reconstruction losses using precomputed features. Training is conducted at both 256×256 and 512×512 resolutions on datasets covering image-text alignment, referring segmentation, and multimodal reasoning.

Instruction Tuning for Region-Controlled Editing. This stage focuses on segmentation perception and controllable generation for LVLMs. We instruct tuning the LLM, the mask decoder, and the cross-attention layers in the diffusion decoder, while freezing all visual encoders. Images are processed at resolutions up to 1024×1024. Segmentation outputs are used as spatial guidance and injected into the diffusion decoder's attention layers. Supervision includes text generation loss, segmentation loss, and L2 loss between denoised outputs and targets.

Table 8: Dataset distribution across training stages in FOCUS

Stage	Number	Task	Source
Stage I: Visual Quantizer and Diffusion Pretraining	45M	Image-to-Image	COYO [5], EMOVA [8], LAION-2B [63]
Stage II: Visual-Language Projector Warmup	30M	Image-to-Text	LLaVA-150K [37, 36], COYO, EMOVA-Pretrain
Stage III: Multimodal Pretraining with Generative and Structural Signals	35M	Image-to-Text & Editing	UltraEdit [83], AnyEdit [77], SEED-Edit [54]
	3M	Segmentation	RefCOCO-Series [76, 44], RefClef [23], Paco-LVIS [49], PartImageNet [20], DAVIS-2017 [6], Pascal-Part [12], YouTube-VIS2019 [72]
	5M	Dialog / QA	Magpie [69], OpenOrca [32], SCP-116K [42], OpenHermes [58], OPC-SFT-Stage1 [22]
Stage IV: Instruction Tuning for Controllable Editing	7M	Image Editing	EMOVA-SFT, Pixmo [17], M4-Instruct [36], OmniEdit [65], AnyEdit, UltraEdit, InstructPix2Pix [4], MagPie
	2M	Segmentation	ReasonSeg [24], Lisa++ Inst. Seg. & CoT [73], ReVOS [70], Ref-Youtube-VOS [52]
	0.5M	Interactive Editing	COCO-Interactive [82]

C Progressive Dataset Structuring

Each stage in FOCUS adopts dedicated datasets aligned with its training objectives, rather than relying on a uniform corpus across all phases. The dataset distribution and task assignments are summarized in Table 8.

Stage 1: Visual Quantizer and Diffusion Pretraining. This stage focuses on learning discrete representations for both semantic abstraction and fine-grained reconstruction. We utilize 45M image-text pairs from COYO, EMOVA, and LAION-2B. COYO contributes OCR-rich, language-aligned samples; EMOVA provides aesthetic-oriented captions; and LAION-2B enhances domain diversity. These datasets collectively support the training of dual visual quantizers and the diffusion decoder.

Stage 2: Visual-Language Adapter Warmup. To establish foundational alignment between visual embeddings and the language model, we train on 30M image-text pairs from COYO, EMOVA-Pretrain, and LLaVA-150K. These samples span natural scenes, documents, and instruction-following captions. This stage optimizes only the projection heads and adapter modules, using a language modeling loss while keeping all backbone components frozen.

Stage 3: Multimodal Pretraining with Generative and Structural Signals. This stage enhances the model's ability to generate, localize, and reason across modalities. We leverage 35M samples from UltraEdit, SEED-Edit, and AnyEdit for text-guided editing. An additional 3M segmentation

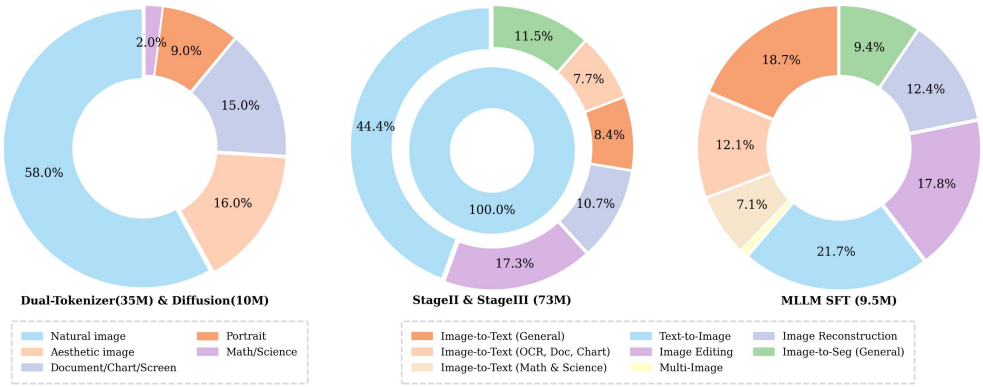

Figure 6: Summary of the data mixture in each stage. Our training data gradually covers a wide range of tasks and various image resoluton.

annotations are sourced from RefCOCO-series, RefClef, and video datasets like DAVIS-2017 and YouTube-VIS2019. For dialogue and question answering, we include 5M samples from Magpie, OpenOrca, SCP-116K, OpenHermes, and OPC-SFT-Stage1. This stage jointly supervises multimodal reasoning, structural grounding, and fine-grained visual tasks.

Stage 4: Instruction Tuning for Controllable Editing. The final stage fine-tunes the model for spatially guided, instruction-based editing. We use 7M samples from EMOVA-SFT, Pixmo, M4-Instruct, OmniEdit, AnyEdit, UltraEdit, InstructPix2Pix, and Magpie for complex editing tasks. We incorporate 2M segmentation-centric samples from ReasonSeg, Lisa++, and video segmentation sets like RVOS and RefYoutubeVOS. Additionally, COCO-Interactive provides 0.5M samples for interactive region-level editing. This stage improves instruction following, multi-object control, and editing consistency across visual contexts.

D Training Configurations Across Different Stages

Table 9: Training hyperparameters across different stages in FOCUS.

Settings	**Visual Quantizer** (Tokenizer)	**Diffusion Decoder** (Image Reconstruction)	**Projector Warmup** (Projector Warmup)	**Multimodal Pretraining** (Seg. Pretrain)	**Instruction Tuning** (Instruction Tuning)
Learning Rate	1e-4 (semantic) 2e-4 (pixel)	2e-5	1e-3	2e-5 (Visual encoder, LLM) 1e-3(Mask Decoder)	2e-5 (Visual encoder, LLM) 2e-6 (Mask Decoder, Diffusion)
Batch Size	256	128	512	128	256
Training Steps	136k (pixel) 28k (semantic)	220k	1epoch	3epoch	1epoch
Image Resolution	256 to 512	512 / 1024	256	256 / 512	512 to 1024
Frozen Modules	Vanilla encoder Hierarchical Encoder	All encoders Codebooks	Visual Encoder, LLM Mask Decoder, Diffusion	Diffusion	Visual Encoder

We adopt stage-specific training configurations to align with the objectives and resolution requirements of each module, as summarized in Table 9.

Dual Visual Tokenizer using a fixed learning rate of 1e-4 and batch size of 256. The training follows three progressive resolution stages: 136k steps at 256×256, 28k steps at 512×512. The semantic branch is optimized with a cosine similarity loss, while the pixel branch is trained with a combination of L1, perceptual, and adversarial losses.

Diffusion Decoder using a learning rate of 2e-5 and batch size of 128 for 265k steps. Training employs multi-aspect-ratio cropping up to 1024×1024 resolution. All visual encoders and tokenizers are frozen, and supervision is provided through L2 reconstruction loss.

LVLM Training is conducted in three distinct stages.

https://doi.org/10.52202/085713-1123

- **Projector Warmup**: Only the projection heads and gated visual adapters are trained using learning rates of $1e{-}3$ (projectors), with a batch size of 512 at 256×256 resolution for 1 epoch.
- **Multimodal Pretraining**: The visual encoder, large language model and mask decoder are optimized jointly. Phase one uses 256×256 resolution with a batch size of 256 for 1 epoch, and phase two uses 512×512 with a batch size of 128 for 2 epochs. A learning rate of $2e{-}5$ is used throughout, and $1e{-}3$ is used for the mask decoder.
- **Instruction Tuning**: The LLM, segmentation decoder, and diffusion model's attention layers are fine-tuned with learning rates of $2e{-}5$ (LLM) and $2e{-}6$ (mask decoder, diffusion), with a batch size of 256 for 1 epoch. Training involves input resolutions randomly sampled from 512×512 to 1024×1024 using bucketed cropping.

E Metrics.

We evaluate FOCUS across three core tasks using standard benchmarks and metrics. For multimodal understanding, we report accuracy on POPE [29], MM-Vet [78], MMBench [40], SEED [26], and MMMU [79] to assess the model's ability in vision-language reasoning, classification, and grounding. For referring segmentation, we use mean Intersection over Union (mIoU) on RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO, measuring the overlap between predicted masks and ground truth under language prompts. For controllable image generation and editing, we evaluate fidelity using FID (Fréchet Inception Distance), and alignment using CLIP-based scores (CLIP-I, CLIP-T, CLIP-DIR) to assess visual consistency and instruction compliance. Additional breakdowns from GenAI-bench and GenEval are used for fine-grained control metrics such as object count, color, and spatial placement.

F Unified Segmentation Capability of FOCUS

FOCUS is designed as a unified model capable of handling diverse segmentation tasks spanning spatial, temporal, and interactive modalities. We evaluate its generalization and adaptability across five key segmentation scenarios, each highlighting a distinct dimension of its fine-grained visual perception.

F.1 Contextual Reasoning and Referential Understanding

We evaluate the ability of FOCUS to perform segmentation under semantically complex and referential conditions using the ReasonSeg and ReVOS benchmarks. These tasks require the model to accurately localize and segment objects based on contextual or linguistic descriptions with varying temporal spans. As shown in Table 10, FOCUS achieves leading performance across both benchmarks. On ReasonSeg, it obtains a gIoU of 62.1 and a cIoU of 58.6, surpassing prior methods in semantic segmentation precision. On ReVOS, FOCUS ranks first in all reasoning and referring metrics, including a J&F of 57.2 in reasoning and 58.9 in referring, validating its unified capability in both spatial understanding and temporal reference tracking.

Table 10: Performance on Referring Video Object Segmentation (ReVOS) and ReasonSeg benchmarks. Bold: best; Underlined: second-best.

Method	Backbone	ReVOS-Reasoning			ReVOS-Referring			ReVOS-Overall			ReasonSeg	
		J	F	J&F	J	F	J&F	J	F	J&F	gIoU	cIoU
LMPM	Swin-T	13.3	24.3	18.8	29.0	39.1	34.1	21.2	31.7	26.4	-	-
ReferFormer	Video-Swin-B	21.3	25.6	23.4	31.2	34.3	32.7	26.2	29.9	28.1	-	-
LISA-7B	ViT-H	33.8	38.4	36.1	44.3	47.1	45.7	39.1	42.7	40.9	52.9	54.0
LaSagnA-7B	ViT-H	-	-	-	-	-	-	-	-	-	48.8	47.2
SAM4MLLM-7B	Efficient ViT-SAM-XL1	-	-	-	-	-	-	-	-	-	46.7	48.1
TrackGPT-13B	ViT-H	38.1	42.9	40.5	48.3	50.6	49.5	43.2	46.8	45.0	-	-
VISA-7B	ViT-H	36.7	41.7	39.2	51.1	54.7	52.9	43.9	48.2	46.1	52.7	**57.8**
VISA-13B	ViT-H	<u>38.3</u>	<u>43.5</u>	<u>40.9</u>	<u>52.3</u>	<u>55.8</u>	<u>54.1</u>	<u>45.3</u>	<u>49.7</u>	<u>47.5</u>	-	-
HyperSeg-3B	Swin-B	50.2	55.8	53.0	56.0	60.9	58.5	53.1	58.4	55.7	<u>59.2</u>	56.7
FOCUS	ConvNext-L	**51.6**	**56.3**	**57.2**	**56.8**	**61.0**	**58.9**	**54.3**	**59.1**	**56.7**	**62.1**	<u>58.6</u>

https://doi.org/10.52202/085713-1123

F.2 User-Guided Interactive Perception

To assess FOCUS's adaptability to user-driven segmentation, we evaluate it on the COCO-Interactive benchmark under four input modes: box, point, mask, and scribble. These interactions simulate real-time editing scenarios where users specify partial object regions.

As summarized in Table 11, FOCUS achieves the highest IoU across all modalities, including a box score of 83.4 and a point score of 78.6, demonstrating robust generalization across sparse and dense cues. This consistent performance under varied interaction types highlights FOCUS's potential for responsive and controllable editing applications.

Table 11: Performance on the COCO-Interactive benchmark. IoU (%) under different input modalities.

Method	Backbone	Box	Scribble	Mask	Point
SAM	ViT-B	68.7	-	-	33.6
SAM	ViT-L	71.6	-	-	37.7
SEEM	DaViT-B	42.1	44.0	65.0	57.8
PSALM	Swin-B	<u>80.9</u>	<u>80.0</u>	<u>82.4</u>	<u>74.0</u>
HyperSeg	Swin-B	77.3	75.2	79.5	63.4
FOCUS	ConvNext-L	**83.4**	**82.7**	**85.2**	**78.6**

F.3 Temporal Consistency in Dynamic Scenes

To evaluate FOCUS's capability in video-level segmentation, we benchmark it on both **Video Object Segmentation (VOS)** and **Video Instance Segmentation (VIS)** tasks. These tasks test the model's ability to consistently track and segment objects across time, even in the presence of occlusions, motion blur, or viewpoint shifts.

Table 12 shows that FOCUS achieves top performance across all datasets, including a J&F score of 79.1 on DAVIS17 and a mAP of 65.7 on YouTube-VIS. Its strong performance on both generic and referring-based video benchmarks confirms its robustness in temporally consistent, instance-aware segmentation.

Table 12: Video segmentation results across DAVIS17 (VOS), Ref-YT, Ref-DAVIS (R-VOS), and YouTube-VIS (VIS). Bold: best; Underlined: second-best.

Method	Backbone	DAVIS17 (J&F)	Ref-YT (J&F)	Ref-DAVIS (J&F)	YT-VIS (mAP)
SEEM	DaViT-B	62.8	-	-	-
OMG-Seg	ConvNeXt-L	74.3	-	-	56.4
ReferFormer	Video-Swin-B	-	62.9	61.1	-
OnlineRefer	Swin-L	-	63.5	64.8	-
UNINEXT	ConvNeXt-L	77.2	66.2	66.7	<u>64.3</u>
LISA-7B	ViT-H	-	53.9	64.8	-
VISA-13B	ViT-H	-	63.0	70.4	-
VideoLISA-3.8B	ViT-H	-	63.7	68.8	-
HyperSeg-3B	Swin-B	<u>77.6</u>	<u>68.5</u>	<u>71.2</u>	53.8
FOCUS	ConvNeXt-L	**79.1**	**69.3**	**72.4**	**65.7**

G The role of the segmentation mask

Table 13: Ablation study on the effect of segmentation mask guidance on EmuEdit benchmark.

Method	**EmuEdit**			
	DINO(↑)	CLIP-I(↑)	CLIP-T(↑)	CLIP-DIR(↓)
No Mask Guidance	0.812	0.866	0.268	0.113
Using a General-Purpose Mask	0.819	0.868	0.271	0.108
FOCUS(Ours)	0.826	0.872	0.275	0.101

To highlight the importance of segmentation masks in our framework, we design three comparative experiments. FOCUS (our full model) utilizes internally generated masks precisely aligned with

https://doi.org/10.52202/085713-1123

language instructions. Baseline 1 removes mask guidance entirely, relying solely on text instructions. Baseline 2 employs masks from an external general-purpose segmentation model (Grounding+SAM2). These experiments demonstrate that precise, instruction-aligned masks are essential for optimal editing performance.

H Visualization of Generation Quality

We present a set of images generated by FOCUS based on natural language prompts (see figure 7). These results span various styles including realistic scenes, conceptual designs, and artistic illustrations. They demonstrate the model's ability to produce high-quality and semantically consistent outputs.

Each image is generated solely from text input without using any visual masks or interactive guidance. The visualizations confirm FOCUS's strength in aligning language instructions with fine-grained visual content, showcasing both fidelity and controllability.

Figure 7: Visual examples of image generation results from FOCUS given only natural language prompts. The model produces diverse and high-fidelity outputs across a range of styles and scenes, demonstrating strong semantic alignment and visual controllability.

I Controllable Image Editing via Segmentation Guidance

To further demonstrate the controllability of FOCUS in localized image editing, we visualize representative examples where the model edits specific regions based on natural language instructions. As shown in Figure 8, the model takes an input image and a user-provided instruction describing the desired modification. It first predicts a segmentation mask that identifies the target region referenced in the text, then uses this mask to guide the generation of an edited output.

This pipeline illustrates the effectiveness of FOCUS in grounding language to spatial regions and applying precise, content-aware modifications. The examples cover diverse scenarios including object transformation, replacement, and contextual adjustment. The results validate that segmentation-aware diffusion guidance enables fine-grained, localized editing while preserving the global scene structure.

Table 14: Prompt schema design for different vision-language tasks. Each row shows how FOCUS handles a specific task by pairing a general instruction (task prompt) with a task-specific condition. This formulation supports unified handling of segmentation, generation, and editing tasks.

Task Type	**Task Instruction Prompt (SI)**	**Condition Prompt (SC)**
Class-based Segmentation	Please segment all the positive objects by the following candidate categories.	`["person", "dog", "car", "tree", ...]`
Referring Segmentation	Please segment the target referred to by the language description.	`"The man wearing a red hat standing beside the yellow car."`
Reasoning Segmentation	Please segment the target referred to by the reasoning-based description.	`"The object that the man is reaching for in the office."`
Interactive Segmentation	Please segment according to the given visual reference regions.	Pooled CLIP region features (e.g., clicks, scribbles, boxes)
Image Generation	Please generate an image according to the following description.	`"A tiny brown dog with white patches, eagerly holding a blue and black Frisbee."`
Image Editing	Please edit the image according to the following instruction.	`"Replace the man in a black jacket with a woman in the same pose."`

J Prompt Design for Multi-Task Vision-Language Modeling

To support a wide range of vision-language tasks within a unified framework, FOCUS adopts a structured prompt schema consisting of two components: a task instruction prompt and a condition prompt in Table. 14. The task instruction prompt defines the model objective in natural language, such as segmentation, generation, or editing. The condition prompt provides task-specific contextual information, such as category labels, referential descriptions, or visual cues.

This design enables the model to flexibly adapt to diverse tasks including class-based segmentation, referring and reasoning segmentation, interactive segmentation with visual cues, text-to-image generation, and fine-grained image editing. By standardizing task formulation through prompt schema, FOCUS achieves better generalization across modalities and applications.

https://doi.org/10.52202/085713-1123

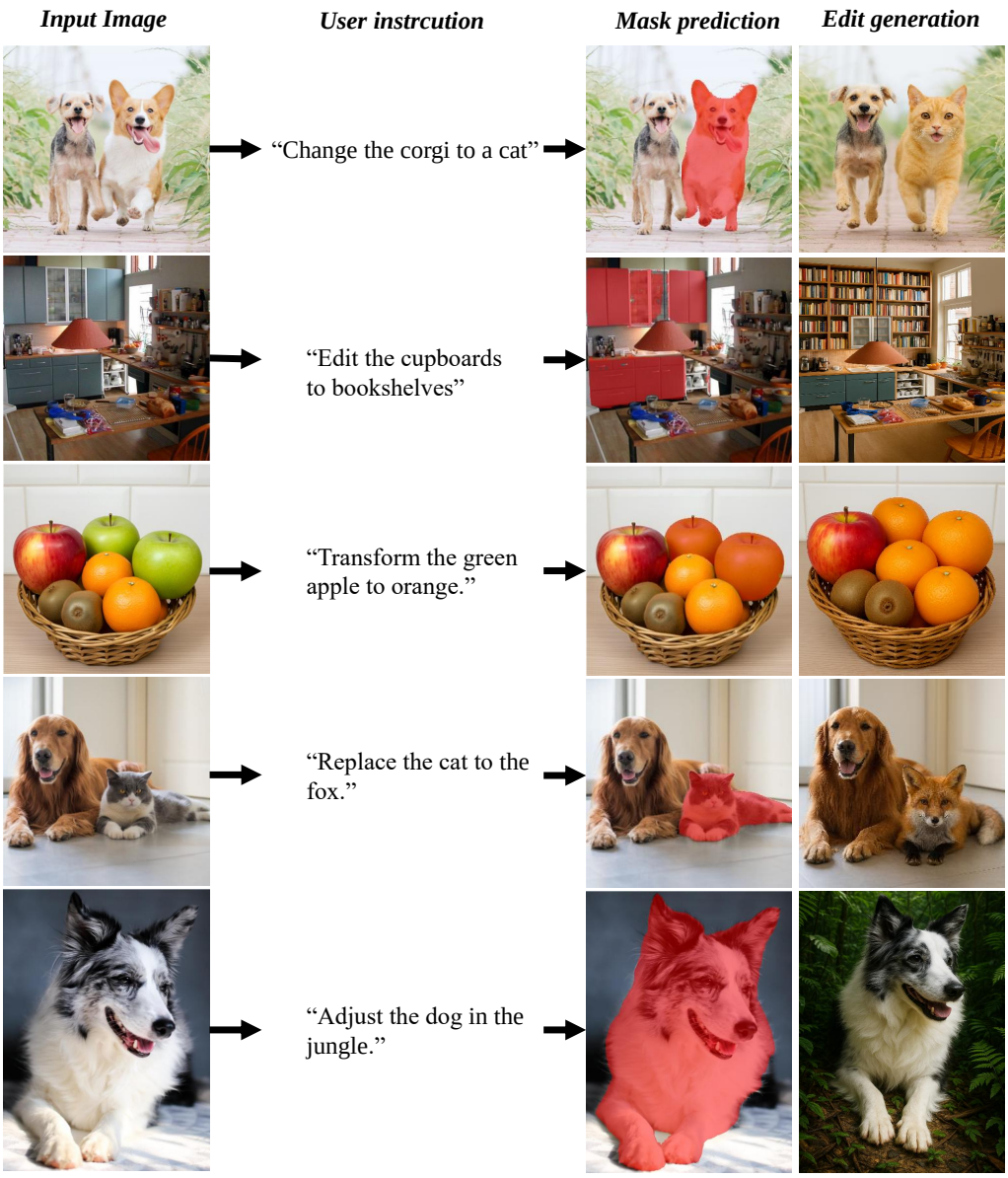

Figure 8: Visualization of controllable image editing results. Given an input image and a user instruction, FOCUS first predicts a spatial mask corresponding to the referential target, then performs localized generation to edit the specified region. The examples demonstrate accurate region identification and high-fidelity edits aligned with the instruction.

https://doi.org/10.52202/085713-1123