



VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing

Hao Fei^{1,2} Shengqiong Wu^{1,2} Hanwang Zhang^{1,3} Tat-Seng Chua² Shuicheng Yan¹
¹Skywork AI, Singapore* ²National University of Singapore ³Nanyang Technological University
 haofei37@nus.edu.sg swu@u.nus.edu hanwangzhang@ntu.edu.sg
 dcscs@nus.edu.sg shuicheng.yan@kunlun-inc.com

Project Homepage: <https://vitron-llm.github.io/>

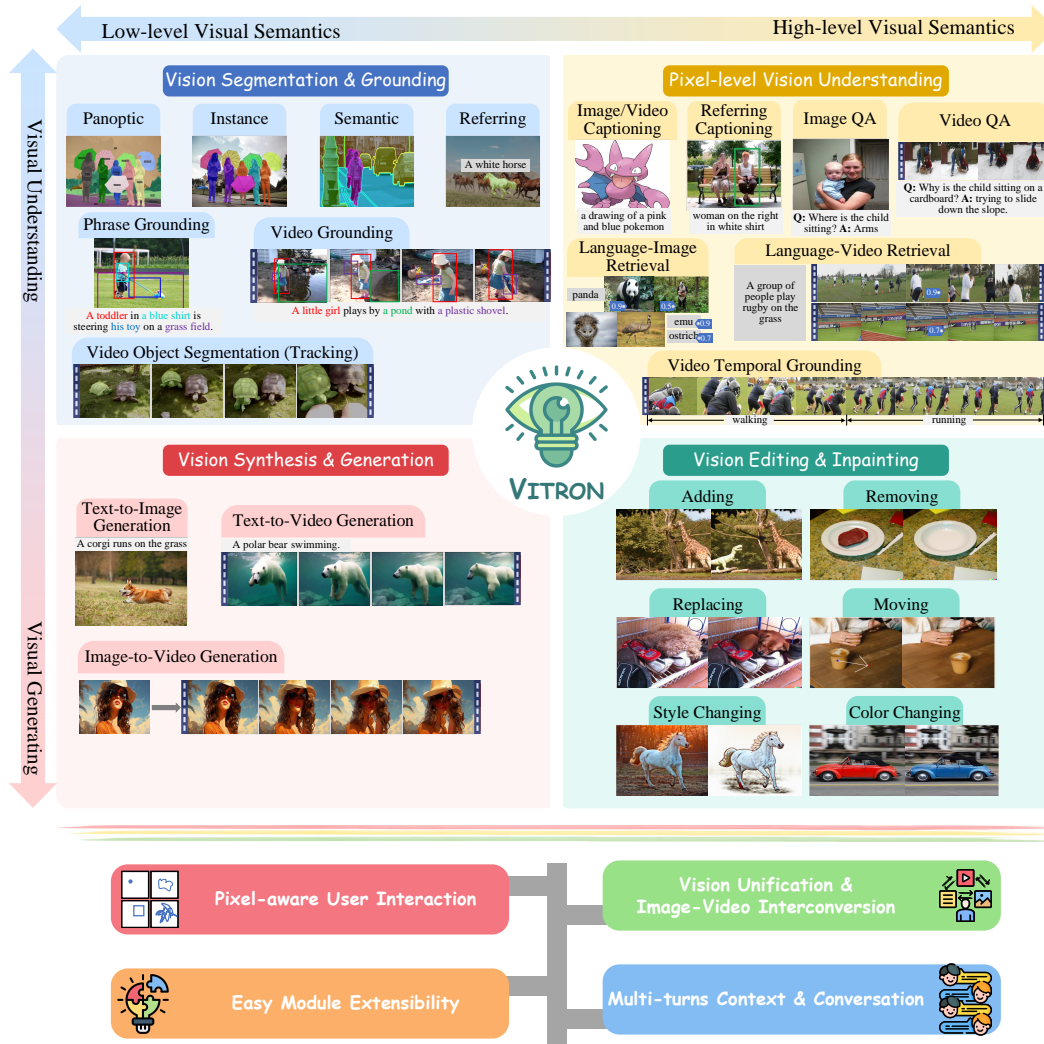


Figure 1: VITRON supports four main task clusters of visions, and also advance in four key features.

*This work was performed when Hao Fei was an Associate Member, and Shengqiong Wu was an Intern at Skywork AI.

Abstract

Recent developments of vision large language models (LLMs) have seen remarkable progress, yet still encounter challenges towards multimodal generalists, such as coarse-grained instance-level understanding, lack of unified support for both images and videos, and insufficient coverage across various vision tasks. To fill the gaps, we present **VITRON**, a universal pixel-level vision LLM designed for comprehensive understanding, generating, segmenting, and editing of both static image and dynamic video content. Utilizing an LLM backbone, VITRON incorporates specialized encoders for images, videos, and pixel-level regional visuals within its frontend architecture, while as its backend, employing a text-centric invocation strategy for integrating diverse state-of-the-art off-the-shelf modules tailored for an array of vision-related end tasks. Via this, VITRON supports a spectrum of vision end tasks, spanning visual understanding to visual generation, from low level to high level. Through joint vision-language alignment and fine-grained region-aware instruction tuning, VITRON achieves precise pixel-level perception. We further enhance its capabilities with invocation-oriented instruction tuning, allowing for flexible and precise module invocation for downstream vision tasks. Demonstrated over 12 visual tasks and evaluated across 22 datasets, VITRON showcases its extensive capabilities in the four main vision task clusters, e.g., segmentation, understanding, content generation, and editing. Various demonstrations also illustrate VITRON’s fortes in visual manipulation and user interactivity. Overall, this work illuminates the great potential of developing a more unified and interactive visual multimodal generalist, setting new frontiers for the next vision research.

1 Introduction

Recently, the field of multimodal large language models (MLLMs) has witnessed rapid and flourishing development across multiple communities. Extensive research efforts have been directed towards augmenting powerful, purely language-based LLMs with modules capable of visual perception, thereby extending their applicability to MLLMs [1, 45, 57, 119, 70]. MLLMs, such as BLIP-2 [45], LLaVA [57], MiniGPT-4 [126] and GPT-4V [112] etc., demonstrate a robust and exceptional capability in image understanding, paralleling the deep semantic comprehension of language. In the realm of vision, the ability to process and comprehend dynamic videos is equally critical. Concurrently, several MLLMs have emerged with a focus on video understanding, e.g., VideoChat [46] and Video-LLaMA [120], demonstrating significant advancements in video comprehension.

Subsequent studies have sought to further expand the capabilities of MLLMs, with efforts bifurcating into two primary dimensions. On one hand, there’s a deepening of MLLMs’ understanding of vision, transitioning from coarse, instance-level comprehension towards a pixel-level, fine-fined understanding of images, thereby achieving visual regional grounding capabilities, as seen in GLaMM [78], PixelLM [79], and MiniGPT-v2 [13], etc., alongside the counterparts in pixel-grounding video LLMs [67]. On the other hand, there’s an expansion in the breadth of functionalities MLLMs can support within the vision field. A portion of the research has already ventured into enabling MLLMs not just to comprehend input vision signals but also to support the generation and output of vision content, with systems like GILL [39], Emu [90], etc., flexibly generating image content, and GPT4Video [99] and NExT-GPT [104] achieving video generation.

We posit that the future trend of vision LLMs necessarily involves the enhancement of their capabilities towards a high degree of unification, i.e., multimodal generalists. However, our observations reveal that despite the diversity of existing vision LLMs developed by the community, there is still a clear lack of unification. **First**, almost all existing vision LLMs treat images and videos as separate entities, either supporting only images or videos [1, 90, 126, 120]. We argue for a unified vision MLLM framework that concurrently supports both images and videos, acknowledging that vision inherently comprises both static images and dynamic videos - both core components of our world and largely interchangeable in most scenarios. **Second**, the current support for vision functionalities in MLLMs is found wanting, with most models only capable of understanding [57, 126], or at most generating images or videos [23, 99]. We contend that future MLLMs should embrace a broader spectrum of vision tasks and operations, enabling unified support for all vision-related tasks and achieving an “one for all” capability, which is vital for real-world applications, especially in vision creation that often involves a series of iterative and interactive operations. For example, users typically start by

Model	Vision Supporting		Pixel/Regional Understanding	Segmenting/ Grounding	Generating	Editing
	Image	Video				
Flamingo [1]	✓	✗	✗	✗	✗	✗
BLIP-2 [45]	✓	✗	✗	✗	✗	✗
MiniGPT-4 [126]	✓	✗	✗	✗	✗	✗
LLaVA [57]	✓	✗	✗	✗	✗	✗
GILL [39]	✓	✗	✗	✗	✓	✗
Emu [90]	✓	✗	✗	✗	✓	✗
MiniGPT-5 [125]	✓	✗	✗	✗	✓	✗
DreamLLM [23]	✓	✗	✗	✗	✓	✗
GPT4RoI [122]	✓	✗	✓	✓	✗	✗
NExT-Chat [118]	✓	✗	✓	✓	✗	✗
MiniGPT-v2 [13]	✓	✗	✓	✓	✗	✗
Shikra [14]	✓	✗	✓	✓	✗	✗
Kosmos-2 [72]	✓	✗	✓	✓	✗	✗
GLaMM [78]	✓	✗	✓	✓	✗	✗
Osprey [117]	✓	✗	✓	✓	✗	✗
PixelLM [79]	✓	✗	✓	✓	✗	✗
LLaVA-Plus [58]	✓	✗	✗	✓	✓	✓
VideoChat [46]	✗	✓	✗	✗	✗	✗
Video-LLaMA [120]	✗	✓	✗	✗	✗	✗
Video-LLaVA [52]	✓	✓	✗	✗	✗	✗
Video-ChatGPT [61]	✗	✓	✗	✗	✗	✗
GPT4Video [99]	✗	✓	✗	✗	✓	✗
PG-Video-LLaVA [67]	✗	✓	✓	✓	✗	✗
NExT-GPT [104]	✓	✓	✗	✗	✓	✗
VITRON (Ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparisons of existing (partially, imperfect coverage) representative vision LLM.

generating images from text, transforming an idea into visual content; and then refining this content through further fine-grained *editing* to add more details; following, proceeding to create dynamic content by *generating* videos from the images; and finally, engaging in several rounds of iterative interaction, such as video *editing*, to enhance and finalize their creation.

To address these gaps, this paper introduces **VITRON**, a pioneering universal pixel-level vision LLM. First, VITRON leverages a backbone LLM for comprehending, reasoning, decision-making, and multi-round user interactions. To perceive both image and video modal signals and support fine-grained user visual inputs, VITRON incorporates encoders for images, videos, and regional box/sketch-specified inputs. On the backend, several powerful off-the-shelf image and video modules are integrated for decoding and executing a wide range of vision tasks, spanning from lower to higher level, such as visual understanding (perceiving and reasoning), generating, segmenting (grounding and tracking), editing (inpainting). By adopting a text-centered invocation approach for module integration, VITRON not only achieves system unification but also ensures alignment efficiency and system scalability. Figure 1 vividly depicts VITRON’s comprehensive functionalities in four major visual-related task groups, and highlights its key strengths.

Our overall training for VITRON aims to equip it with robust and powerful vision understanding and manipulation capabilities. We start by undertaking a joint vision-language alignment learning between the frontend and the center LLM, to imbue basic vision understanding. Further, to achieve pixel-level perception, we introduce fine-grained spatiotemporal vision grounding instruction tuning, training LLM on grounding predictions and pixel-aware perception for images and videos. Lastly, to ensure VITRON can flexibly and accurately invoke downstream modules, we propose invocation-oriented instruction tuning. By constructing a rich number of instruction-tuning examples across various pixel-level task scenarios, VITRON has been taught to faithfully follow instructions.

Extensive experiments covering 12 tasks across 22 datasets are performed. Leveraging its advanced architecture as a multimodal generalist, VITRON demonstrates proficiency in a comprehensive range of vision tasks. Notably, the unified system’s performance is on par with or even surpasses singleton state-of-the-art (SoTA) specialists on specific tasks. For each task, demo visualizations highlight VITRON’s robust multi-round user interaction and visual manipulation skills. To our knowledge, VITRON is the first vision LLM generalist that possesses a grand unification of various pixel-level visual capacities for understanding, generating, segmenting, editing of both images and videos. With VITRON, we aspire to create a powerful open-sourced, interactive vision system that can compete with industry-level vision-language systems like OpenAI’s DALL-E series [7] and the Midjourney [71], thereby aiding the advancement of academic research.

2 Related Work

Achieving a profound understanding and comprehensive operational capabilities in vision represents a significant topic within the field of artificial intelligence (AI). The relevant communities have developed a wide array of tasks in this regard. These tasks range from low-level visual pixel understanding, such as visual object detection [9, 59], image semantic segmentation [81, 111, 42, 48], and video segmentation [113, 96], to high-level comprehension of overall semantics, including image classification [21, 41], image/video captioning [64, 64, 29], and visual question answering (VQA) [110, 105, 49, 114, 3] etc. Moreover, tasks also can be categorized from different perspectives, such as visual understanding [44, 70], and visual generation [86, 31, 95, 26, 56] and editing [68, 6, 11]. Over the last few decades, various neural network models [43] have been developed, demonstrating powerful effects across a spectrum of vision tasks. Increasing research has validated the impact of scaling laws on neural network models: the effectiveness of vision models improves with the scaling of model size and training data volume [22, 4], sometimes even exhibiting emergence capabilities [100]. Consequently, recent years have seen the development of highly potent large-scale vision models, such as ViT [24] and CLIP [77], which have achieved remarkable vision understanding capabilities; models like SAM [38] and SEEM [127] have solved vision segmentation tasks; and diffusion-based models [32, 76, 66, 27, 80] have reached unprecedented performance in vision generation. Yet these models might lack an LLM as a central decision processor, unable to flexibly interpret user intent or execute tasks interactively [91, 104].

The emergence of LLMs has garnered unprecedented attention over the past year, even fostering hope for achieving human-level AI [69, 18, 93]. Extending the success of language understanding in LLMs, researchers have promptly investigated and developed various MLLMs, enabling LLMs to comprehend vision. By integrating high-performance vision encoders of images or videos into language-based LLMs, these models have been made capable of understanding vision signals [70, 1, 45, 75, 57]. Going beyond vision understanding, further research has aimed to enhance MLLMs, for instance, by endowing them with vision generation capabilities [39, 90] or supporting pixel-level understanding and grounding [122, 117, 79]. We present a brief summary of some existing popular vision LLMs in Table 1 in terms of the vision function supporting.

However, we observe that current research on vision LLMs lacks depth in two critical aspects. Firstly, current vision LLMs tend to separate images and videos, supporting either one or the other. The construction of a unified MLLM is crucial, as vision inherently encompasses both static images and dynamic videos, both of which are core components of our visual world. Thus, covering both aspects simultaneously is essential for optimally adapting to practical applications. Although models like NExT-GPT [104] have relatively well-supported unification across various modalities, they fall short in supporting pixel-level in-depth vision understanding and comprehensive support for vision operation tasks. The second issue is the incomplete support for vision tasks by existing MLLMs. Most current MLLMs primarily support understanding images or videos [57, 126], with only a few supporting generation [23, 99] or editing/inpainting [103]. LLaVA-Plus [58], for example, supports a broader range of vision functionalities by invoking external tools, but unfortunately suffers from a lack of pixel-level in-depth vision understanding, and does not support pixel-aware user interaction, further with its capabilities limited to image manipulation only. Building a generalist that can handle (almost) all vision-related tasks and operations will be the next major trend for vision LLMs.

On the other hand, this work aims to build a vision system with exceptional user-friendly interactivity. Our goal is to benchmark against some of the industry-level AI-empowered commercial vision systems, such as OpenAI’s DALL-E series² and the Midjourney³ system. For instance, DALL-E, with its integration of the GPT-4V [112] system, can support mult-round user interaction, enabling image generation and editing. However, it remains questionable whether these systems genuinely support precise pixel-level vision understanding and manipulation capabilities. Moreover, these systems do not simultaneously support a unified understanding and operation of both image and video modalities. The most serious issue might possibly be that these outstanding systems are proprietary and close-sourced, which significantly hinders research in this area. It is with these considerations in mind that our system was proposed. To our knowledge, VITRON is the first vision LLM that not only offers unified pixel-level capabilities for understanding, generating, segmenting, and editing both images and videos but also boasts highly interactive user engagement.

²<https://openai.com/research/dall-e>

³<https://www.midjourney.com>

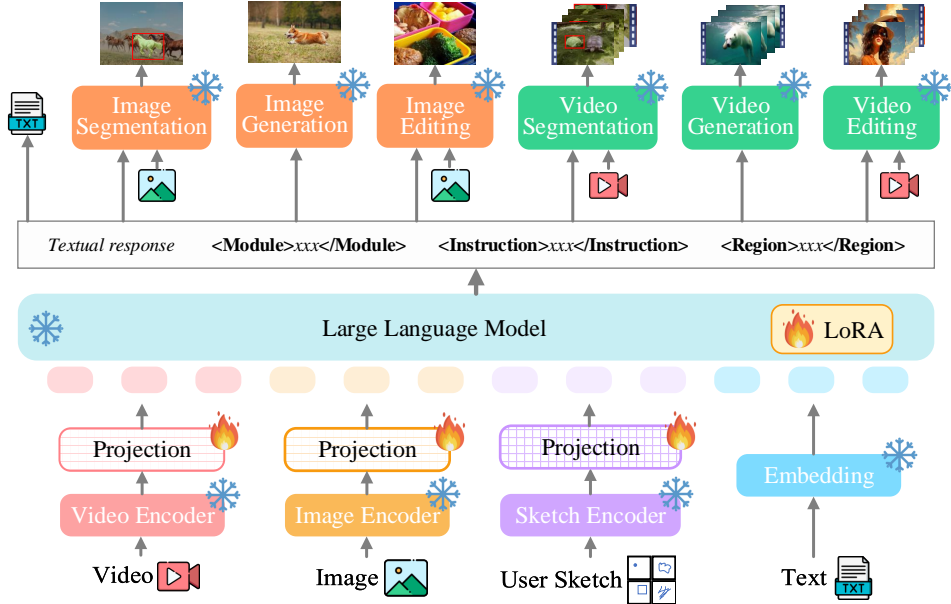


Figure 2: Overview of the VITRON framework.

3 Architecture of VITRON

The overall framework of VITRON is shown in Figure 2. VITRON takes most common architecture as in relevant MLLMs [57, 23, 104]. Three key blocks are included: 1) frontend vision&language encoding, 2) central LLM understanding and text generation, and 3) backend user responding and module invocation for vision manipulation.

3.1 Vision-Language Encoding Block

The text inputs (their embedding) are directly passed to the core LLM without additional preprocessing. For both the encoding of images and videos, we employ the CLIP ViT-L/14@336px [77] as the encoder, respectively. The encoders are with a patch size of 14, and convert all images and video frames into 336px resolutions. The video encoder independently processes each frame, employing average pooling across the temporal dimension to yield temporal representation features. This is essential for the model’s effective capture of spatiotemporal representations within videos. Then, we employ a regional pixel-aware visual extractor as the sketch encoder for user interaction. Typically, the types of interactions can vary, including actions like clicking, drawing boxes or polygons, and making scribbles. We mainly follow [117], and use the object-based representations of mask regions that come from user’s inputs, which not only encode the pixel-level visual features but also gather the spatial position information of each region. The region features are pooled with also the binary mask of spatial geometry of the object region encoded, and the resulting embeddings are used. Then, the multimodal feature representations are passed to the corresponding linear projection layers, which are latter mapped into language-like embeddings that are understandable to the LLM.

3.2 Core LLM Block

In VITRON, an LLM serves as the pivotal agent. Technically, we utilize Vicuna (7B, version 1.5), an open-source, text-based LLM that is extensively adopted across various MLLMs, as frequently referenced in the relevant literature [17, 88, 120]. The LLM processes inputs from both language and visual modalities to perform semantic understanding and reasoning. It generates outputs in two main forms: 1) direct textual responses, and 2) formatted text for module invocation. The latter includes detailed information on how to call specific modules. For example, based on the module invocation text, the system might trigger an image segmentation module to isolate the region of interest around a target object. We regard the foundational LLM as an agent, specifically tasked with producing module invocation text. This approach is highlighted as an optimal trade-off between maintaining effective alignment with the LLM-decoder and ensuring the training process remains efficient. Given our system integrates a vast array of backend modules, linking various decoders and the LLM via

No.	Function	Model	Input	Output
1	Text Generation	-	-	-
2	Image Generation	GLIGEN [50]	Text	Image
3	Image Segmentation	SEEM [127]	Text, Image	Image, Mask BBox
4	Image Editing	GLIGEN [50]	Text, Image [BBox Mask]	Image
5	Video Generation	ZeroScope [10]	Text	Video
6		I2VGen-XL [123]	Image	Video
7	Video Segmentation	SEEM [127]	Text, Video [BBox Mask]	Video, Mask BBox
8	Video Editing	StableVideo [11]	Text, Video	Video

Table 2: Summary of backend modules in VITRON.

feature embedding [104] instead would significantly increase the complexity and learning burden for LLM-decoder alignment [98], which would also hinder the potential expansion of the system with additional decoding modules.⁴ Thus, such a text-centered invocation approach decouples the LLM from various modules, making the backend decoders more easily and flexibly replaceable.

3.3 User Responding & Module Invocation Block

To address the inability of text-based LLMs in handling various vision tasks, we consider integrating off-the-shelf external modules. Once the LLM generates invocation details through understanding the input and recognizing the user’s intent, the corresponding modules are activated to produce non-textual outputs. Technically, we employ a variety of current SoTA expert models for vision processing. For image generation and editing, we integrate the diffusion-based model GLIGEN [50]. For image and video segmentation, we opt for SEEM [127]. For video generation, ZeroScope [10] and I2VGen-XL [123] are utilized for text-to-video and image-to-video tasks, respectively. Lastly, for video editing functionality, we incorporate StableVideo [11]. In Table 2, we provide a detailed look at the functionality of each backend module, along with a specification to the inputs and outputs corresponding to each module.

4 Pixel-aware Vision-Language Understanding Learning

Building on the architecture of VITRON, we proceed to train the model to endow it with robust vision understanding and task execution capabilities. The model training is structured into three distinct phases. Initially, we undertake a joint training of the frontend with the language-based LLM for vision-language alignment learning, enabling the model to acquire basic vision understanding capabilities. Subsequently, we engage in fine-grained vision grounding instruction tuning to further enhance the model’s pixel-level perception abilities. In the final step, we carry out invocation-oriented instruction tuning, allowing the LLM to flexibly and accurately invoke downstream modules.

4.1 Step-I: Overall Vision-Language Alignment Learning

In line with the methodologies employed by current MLLMs, our approach involves mapping the input vision language features to a unified feature space. This space creates representations that the central LLM can interpret, thereby enabling it to process incoming vision signals effectively. For this purpose, we utilize datasets comprising ‘image-caption’ pairs (CC3M [84]), ‘video-caption’ pairs (Webvid [5]), and ‘region-caption’ pairs (RefCOCO [36]) drawn from existing established corpora and benchmarks. When provided with an image, video, or specific visual region, we engage the frozen LLM to generate a text description or caption that aligns with the reference caption.

4.2 Step-II: Fine-grained Spatiotemporal Vision Grounding Instruction Tuning



Our system leverages an invocation approach to execute various pixel-level vision tasks by utilizing external modules, but the LLM itself has not undergone fine-grained vision training any. This limitation prevents the system from achieving true pixel-level vision understanding. Also, possessing pixel-level vision understanding would significantly enhance the system’s ability to accurately issue commands for invoking the correct modules to perform tasks. To address this, we propose a fine-grained spatiotemporal vision grounding instruction tuning for VITRON. The core idea is to enable the LLM to ground the fine-grained spatiality of images and the detailed temporality of videos.

⁴Currently, there would be a debate in the community about which is better to bind LLM with backend modules, discrete language tokens [98], or continual signal embeddings [39, 104].



Technically, we leverage LoRA [33] to enable a small subset of parameters within the LLM to be updated during the tuning.

Image Spatial Grounding. Considering that the LLM alone can only output text, we design it to respond with the corresponding bounding box areas. We focus on two types of tasks: grounded image captioning and referring image segmentation. Initially, for grounded image captioning, we input an image and identify a specific object within it, prompting the LLM to describe the identified object. Conversely, for referring image segmentation (where we consider outputting a bounding box, akin to phrase grounding), the task involves inputting a complete image along with a related phrase or sentence description, and the LLM is expected to output the object’s spatial bounding box, represented by coordinate numbers (X_l, Y_t, X_r, Y_b) . The X and Y coordinates are normalized real values within the range [0, 1], where $\langle X_l \rangle$ and $\langle Y_t \rangle$ indicate the top-left corner, and $\langle X_r \rangle$ and $\langle Y_b \rangle$ represent the bottom-right corner of the bounding box.

We primarily consider datasets such as Flickr30K Entities [73], RefCOCO [36], RefCOCO+ [115] and RefCOCOg [62]. Utilizing GPT-4, we preprocess and expand the original data into a multi-turn QA format, thereby generating the most suitable instruction-tuning data.



<p>► Grounded Image Captioning:</p> <p>Input: </p> <p>Please generate a detailed caption for the given image, and clearly link each part of the caption to specific objects or areas in the image which you can denote with a bounding box with “object: (X_l, Y_t, X_r, Y_b)” format.</p> <p>Output:</p> <p><i>A girl wearing a pink dress is sitting on a long bench in the park reading a book.</i></p> <p><i>girl: (X_l, Y_t, X_r, Y_b),</i></p> <p><i>pink dress: (X_l, Y_t, X_r, Y_b), ...</i></p>	<p>► Referring Image Segmentation:</p> <p>Input: </p> <p>Please identify the target object from the given images based on the following text query: “A man in a black shirt holding a water bottle”.</p> <p>Please output the bounding box (X_l, Y_t, X_r, Y_b) of the target object.</p> <p>Output:</p> <p>(X_l, Y_t, X_r, Y_b)</p>
---	--

Video Spatial-Temporal Grounding. For videos, the LLM must not only identify spatial regions but also ground these within the temporal context of the video, essentially achieving video tracking. Similarly, we explore tasks such as grounded video captioning and referring video tracking. While the approach is akin to that used for images, it requires not just the output of object bounding box coordinates in each frame but also the specification of the duration of the frames (F_s, F_e) in which these coordinates are to be outputted. Here $\langle F_s \rangle$ denotes the starting frame number, and $\langle F_e \rangle$ means the ending frame number. We mainly construct the required instruction tuning data from several key video tracking datasets: LaSOT [25], GOT10K [34], DanceTrack [89], and MOT16 [63].

<p>► Grounded Video Captioning:</p> <p>Input: </p> <p>Please generate a caption for the given video, and link each part of the caption to specific objects in the video with its temporal presence duration. You should denote it with a bounding box with the starting and ending frame number in format as “object: $(X_l, Y_t, X_r, Y_b F_s, F_e)$”.</p> <p>Output:</p> <p><i>A dog is running after a ball on the grass.</i></p> <p><i>dog: $(X_l, Y_t, X_r, Y_b F_s, F_e)$,</i></p> <p><i>ball: $(X_l, Y_t, X_r, Y_b F_s, F_e)$, ...</i></p>	<p>► Referring Video Tracking:</p> <p>Input: </p> <p>Please track the specified object throughout the video based on the following given description, and mark its starting and ending position in each frame: “A man is skiing on a thick-snow-covered slope.”</p> <p>Please output in the format of “$(X_l, Y_t, X_r, Y_b F_s, F_e)$”.</p> <p>Output:</p> <p>$(X_l, Y_t, X_r, Y_b F_s, F_e)$</p>
--	--

Grounding-aware Vision QA. The grounding tasks mentioned above only touch upon the low-level aspects of vision perception. However, in many scenarios, it’s essential for the LLM to possess high-level, in-depth vision reasoning capabilities, building upon the foundational low-level pixel grounding. Therefore, we further introduce grounding-aware vision QA, enabling the LLM to undertake semantic-level QA tasks based on the grounded results. Specifically, we consider both image-grounded QA and video-grounded QA. The data sources for these tasks remain the same as mentioned previously. We again rely on GPT-4 once more to use its imaginative capabilities to design



sensible and reasonable questions and answers based on the given grounding content, formatting these into an instruction-tuning format.

<p>► Grounded Image QA:</p> <p>Input: </p> <p>Based on the given image, please select the correct answer among all the candidates: Q: Where is the child sitting? A: 1) arms, 2) ground, 3) chair, 4) bed.</p> <p>Please first identify and ground the target object (in coordinates) mentioned in the question, and then proceed to answer the question.</p> <p>Output:</p> <p>The target object mentioned in the question is "child," with the position given by (X_l, Y_t, X_r, Y_b). From this, it can be determined that the child is being held by a woman, which means the child is sitting on the woman's arms. Therefore, the answer is 1) arms.</p>	<p>► Grounded Video QA:</p> <p>Input: </p> <p>Based on the provided video, answer the following question by choosing the most appropriate answer from the options given. Q: If the rope in the video breaks, what would happen to the person in the video? A: 1) running, 2) jumping up, 3) flying away, 4) fracture.</p> <p>First, analyze the spatial position and temporality of the target object mentioned in the question within the video, and based on this analysis, determine the answer to the question.</p> <p>Output:</p> <p>The objects involved in the question are the rope $(X_l, Y_t, X_r, Y_b F_s, F_e)$ and the person $(X_l, Y_t, X_r, Y_b F_s, F_e)$, with their relationship being that the person is climbing with the assistance of the rope. Based on common sense, if the rope were to break, the person would fall to the ground from a height. The most likely outcome is that the person would break a bone.</p>
--	--

4.3 Step-III: Output-side Invocation-oriented Instruction Tuning

The above second phase of training endows both the LLM and the frontend encoders with the ability to understand vision at the pixel level. This final step, instruction tuning for invocation, aims to equip the system with the precise capability to execute commands, allowing the LLM to generate appropriate and correct invocation text. This text is then used to trigger various backend task execution modules. Different terminal vision tasks might require distinct invocation commands. To unify this, we try to standardize the LLM's response output into a structured text format, which includes: 1) User response output, which directly replies to the user's input. 2) Module name, indicating which function or task is to be executed. 3) Invocation command, a meta-instruction for triggering the task module. 4) Region (optional), specifying a fine-grained vision feature needed for certain tasks, such as in video tracking or vision editing, where backend modules require this information. For the region, based on LLM's pixel-level understanding, a bounding box described by the coordinates (X_l, Y_t, X_r, Y_b) will be output. Following, we exemplify a video tracking example for the structured LLM response for module invocation.

► **Structured LLM response for module invocation (with video tracking example):**

Input:  

Can you help me track the clock I circled in the video?

LLM response:

Sure! You provide a video where there is a female cop talking on the walkie-talkie, responding emergency call, crime prevention. Also in the scene, the clock is green in color. Following I will outline the clock in the video.

<Module> Video Segmentation </Module>

<Instruction> segmentation: clock </Instruction>

<Region> (0.23, 0.35, 0.11, 0.26) </Region>

To teach the LLM to produce the correct invocation responses, we need to create data specifically designed for instruction tuning. A crucial aspect of this is ensuring that the data covers all possible scenarios. We must take into account different ways users might interact with the system for each functionality mentioned in Table 2 (except for text generation). For example, when requesting video creation, a user might describe what they want purely in text, or provide a reference image as the basis for the desired video. Similarly, for editing images or videos, users could express their editing requests either through text, or by using sketches, scribbles and other interactions. Consequently, the language model needs to be versatile in accepting various types of user inputs and generating an

Function	VITRON Input	Module Input	Data Source	Amount
Image Generation		Image-Caption	CC3M [84]	2,000
Image Segmentation	, ,	,	RefCOCO [36]	2,000
	, , ,	, ,	RefCOCO [36]	2,000
	,	, Object-Name	gRefCOCO [54]	2,028
Image Editing	, ,	,	COCO2017 [53]	1,992
	,	, Bounding-Box	MagicBrush [121]	2,000
Video Generation		Video-Caption	WebVid [5]	2,000
	,		LAION-400M [82]	2,000
Video Segmentation	, ,	, ,	WebVid [5], VG [40]	2,982
	,	, , Bounding-Box	WebVid [5]	1,974
Video Editing	, ,	, Editing-Query	WebVid [5]	1,980
	,	, Editing-Query	WebVid [5]	1,952

Table 3: Feature summary of our invocation-oriented instruction-tuning data. in image segmentation means the reference image provided by user. in video segmentation means the intermediate referred video keyframe interpreted within system.

accurate invocation response that matches the requirements of the backend modules. Technically, we make use of the existing annotated datasets for various vision tasks included in this work. For each task under specific different user input scenarios, with the corresponding data, we design various template dialogue-format examples. Based on these examples we then prompt the GPT-4 to generate more samples under various topics and enriched scenarios. Finally, we collect a total of 22,000+ invocation-oriented instruction tuning samples. In Table 3 we provide a summary of these datasets, including the input content of VITRON and modules as well as the data source&amount.

5 Experiments

5.1 Settings and Briefings

In this section, we aim to quantify the performance of VITRON on a variety of standard benchmarks for downstream vision tasks and compare it against some of the currently strong-performing systems. It is noteworthy that, given the countless vision tasks within the community, our experiments focus only on 1-2 of the most representative tasks from each task category for validation. To ensure a fair comparison, all subsequent experiments adopt settings same or similar to those of baseline systems, with evaluations following established practices. Before conducting experiments, we perform targeted pre-training on all of VITRON’s backend modules (such as GLIGEN and SEEM) on their respective datasets. This ensures our system is optimized for the best possible performance during testing. Our approach centers on training the linear projection layers of all encoders and efficiently fine-tuning the language model using LoRA. To train our model, we employ the AdamW optimizer along with a learning rate scheduler. The pre-training of VITRON unfolds in three phases, all conducted on $10 \times$ A100 (80G) GPUs. Initially, we train the model using a global batch size of 128 and a maximum learning rate of $3e-4$, a process that takes approximately 40 hours. In the second tuning phase, we adjust the model with a maximum learning rate of $1e-5$, utilizing a global batch size of 90. This stage of training lasts about 35 hours. The third phase of training employs a global batch size of 128 and maintains the maximum learning rate of $1e-5$, completing in roughly 10 hours.

Method	RefCOCO [36]			RefCOCO+ [115]			RefCOCOg [62]	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
LAVT [111]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [54]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA [42]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
NExT-Chat [118]	74.7	78.9	69.5	65.1	71.9	56.7	67.0	67.0
VITRON	74.4	78.7	71.6	66.3	72.1	57.8	67.2	67.3

Table 4: Results (cIoU) of image referring image segmentation.

5.2 Vision Segmentation

We consider both the image segmentation and video segmentation tasks.

Image Segmentation. Table 4 presents the results of referring image segmentation on three datasets: RefCOCO [36], RefCOCO+ [115] and RefCOCOg [62]. We compare with several significant models, including state-of-the-art non-MLLM approaches and the MLLM baseline, NExT-Chat. It is evident that our VITRON, while slightly underperforming compared to NExT-Chat on the RefCOCO Val&TestA datasets, achieves superior performance on the remaining sets. Figure 3 further demonstrates an example of how our VITRON processes image segmentation tasks in an interactive manner with user. When users sketch or doodle outlines on specific areas of an image, VITRON is capable of accurately identifying the corresponding objects within the image. Following this identification, it precisely generates the bounding box and mask area for the identified objects.

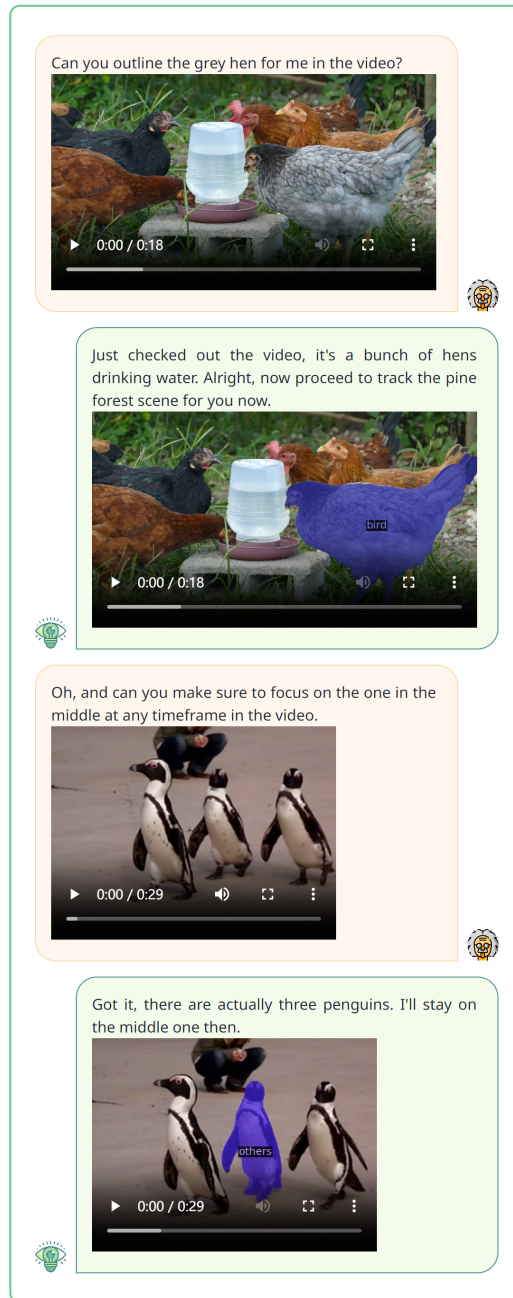
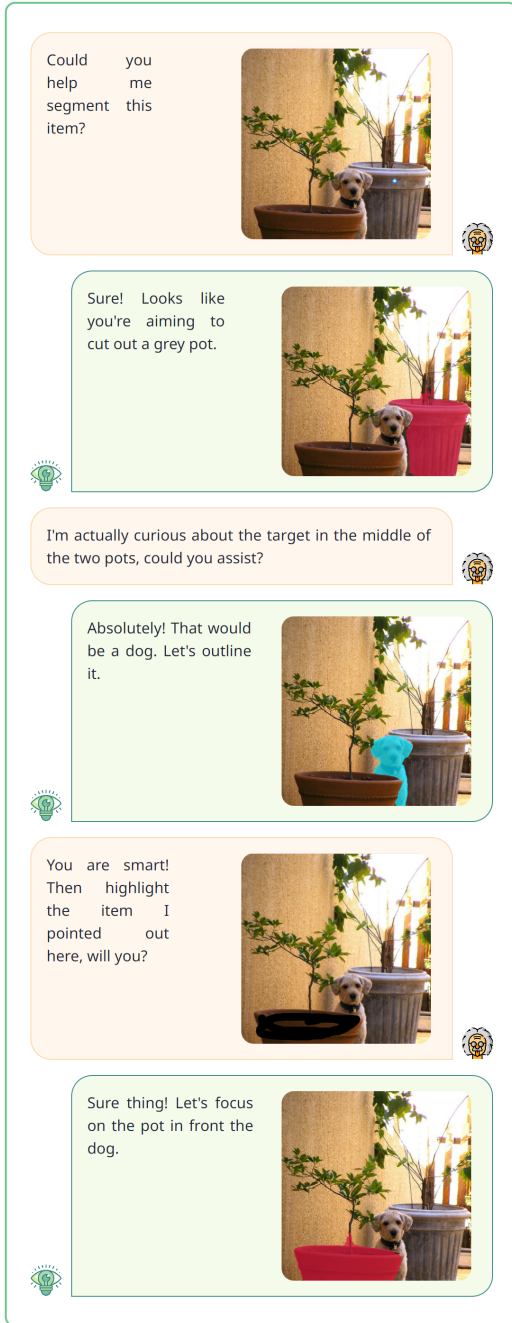


Figure 3: Demonstration of image segmentation.

Figure 4: Example of video object segmentation.

Video Segmentation. For video segmentation, we explore two tasks: video spatial grounding (with bounding box) and video object segmentation (aka., video tracking; with mask). Table 5 showcases the comparisons between VITRON and current state-of-the-art (SoTA) video MLLMs in video spatial grounding. It is clear that VITRON significantly outperforms PG-Video-LLaVA. Table 6 presents a comparison of VITRON with some SoTA systems in video tracking, where our system continues to demonstrate superior performance (only with the exception of the \mathcal{F} metric on the DAVIS 17 [74] Test-Dev).

Method	VidSTG [124]	HC-STVG [92]
G-DINO [59]	25.3	19.5
Video-LLaMA [120]	28.6	26.1
Video-ChatGPT [61]	32.8	20.8
PG-Video-LLaVA [67]	34.2	28.3
VITRON	38.4	30.1

Table 5: Results (mIoU) of video spatial grounding on two datasets.

Table 5: Results (mIoU) of video spatial grounding on two datasets.

Method	DAVIS 17 [74] Test-Dev			Youtube-VOS 2019 [109] Val			
	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{J}_s	\mathcal{F}_s	\mathcal{J}_u	\mathcal{F}_u
RDE [47]	77.4	73.6	81.2	81.1	85.5	76.2	84.8
XMem [16]	81.0	77.4	84.5	84.3	89.6	80.3	88.6
DeAOT [113]	80.7	76.9	84.5	84.6	89.4	80.8	88.9
ISVOS [96]	82.8	79.3	86.2	85.2	89.7	80.7	88.9
VITRON	83.0	79.9	85.8	85.4	89.8	81.1	89.6

Table 6: Results of video object segmentation.

Figure 4 also demonstrates the process of video segmentation. VITRON successfully identifies, localizes and tracks the target in the video based on the provided queries. Our system demonstrates a remarkable ability to accurately and flexibly capture the user’s intent, even when the desired object to be segmented is described in a highly implicit manner.

5.3 Fine-grained Vision Understanding

Next, we evaluate VITRON’s capability in achieving fine-grained vision understanding, focusing mainly on region-level tasks for both images and videos.

Region-level Image Understanding. We test VITRON on tasks including image referring expression comprehension and image regional captioning. The comparisons and results shown in Tables 7 and 8 illustrate that VITRON surpasses the best baseline across various datasets and metrics, proving its strong and accurate fine-grained semantic understanding of images. Figure 5 further highlights its fine-grained visual understanding capabilities through regional image captioning, to accurately understand and appropriately describe image regions.

Method	METEOR	CIRer
GRIT [102]	15.2	71.6
Kosmos-2 [72]	14.1	62.3
NExT-Chat [118]	12.0	79.6
MiniGPT-v2 [13]	15.0	86.4
GLaMM [78]	16.2	106.0
Osprey [117]	16.6	108.3
VITRON	16.8	110.8


Table 7: Performance of image regional captioning on RefCOCOg [62].

Method	RefCOCO [36]			RefCOCO+ [115]			RefCOCOg [62]	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
OFA [97]	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
Shikra [14]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
MiniGPT-v2 [13]	88.7	91.6	85.3	79.9	85.1	74.4	84.4	84.6
VITRON	90.3	92.5	88.7	83.4	88.6	76.1	85.8	86.4


Table 8: Results (accuracy) of image referring expression comprehension.

The above two tasks focus solely on the model’s ability to recognize at the region level. Taking a step further, we delve deeper into assessing the capability for image semantics understanding, particularly through image-based Visual Question Answering (VQA) tasks. These tasks effectively reflect the model’s proficiency in comprehending the deeper semantic content of images. Table 9 displays the results across a series of six datasets for image-based VQA. We primarily compare two groups of models: those with and without pixel-wise vision grounding capabilities. The findings indicate that models equipped with fine-grained grounding abilities indeed show stronger task performance, suggesting that fine-grained grounding contributes to a more profound understanding of semantics. Notably, our VITRON achieves the highest performance among the models evaluated.

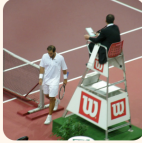
Could you tell me what's happening in this photo?



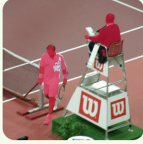
The picture features a group of people riding in a small electric propulsion boat on a body of water. There are three people visible in the boat, with one person sitting in the front, another in the middle, and the third person in the back. The boat appears to be a motorboat with an electric motor, providing a more environmentally friendly mode of transportation. The people seem to be enjoying their time on the water, possibly engaging in recreational activities or simply taking a leisurely ride.



So what is the possible relationship between the two persons according to the picture?



In the image, there is a man holding a tennis racket and a chair umpire on a tennis court. The relationship between the two persons is that the chair umpire is responsible for overseeing the match and making decisions related to the game, such as calling balls and strikes, while the man holding the tennis racket is a tennis player participating in the match.



Then, why the person in black sitting on a such high chair?

The person in black is sitting on a high chair, also known as a chair umpire's chair, to have a better vantage point to observe the tennis match and make accurate judgments. The chair's elevated position allows the umpire to have a clear view of the entire court, including the lines and the players' movements. This helps the umpire to make fair and impartial decisions during the match, such as calling balls and strikes, and ensuring that the game is played according to the rules.

Figure 5: Demonstration of fine-grained visual understanding via image grounding.

Can you tell me what's happening with the girl in the video?



Looks like the girl is getting a mini golf lesson there. She's gripping a golf stick, and there's a man bend down, to teach the girl how to swing.



And how'd she manage to hit the ball on the floor?

She hits the ball on the floor with a golf club, which is a piece of equipment used in the sport of golf.

I would like to see a video where a dogs run past a car on the street in the snow.

Absolutely, I've got just the video for you.



Hmm, it's cool but not quite what I had in mind. Let me paint a picture for you: how about starting with adding a white dog next to the car?



Got it, I've placed a white dog right next to the car, on its right side, to be precise.



Awesome, now could you convert that image into a dynamic video where the dog is seen running past the car?

Sure thing, here's your video showcasing the dog making a swift move past the car.



Figure 7: Demonstration of vision generation across different modalities.

Method	Grounding	OKVQA [83]	GQA [35]	VSR [55]	IconVQA [60]	VizWiz [30]	HM [37]
Flamingo [1]	✗	44.7	-	31.8	-	28.8	57.0
BLIP-2 [45]	✗	45.9	41.0	50.9	40.6	19.6	53.7
InstructBLIP [19]	✗	-	49.5	52.1	44.8	33.4	57.5
MiniGPT-4 [126]	✗	37.5	30.8	41.6	37.6	-	-
LLaVA [57]	✗	54.4	41.3	51.2	43.0	-	-
Shikra [14]	✓	47.2	-	-	-	-	-
MiniGPT-v2 [13]	✓	57.8	60.1	62.9	51.5	53.6	58.8
VITRON	✓	58.3	61.7	63.2	51.8	54.1	59.4

Table 9: Results (accuracy) on image-based VQA.

Region-level Video Understanding. Similarly, for videos, we evaluate the Region-level Video Understanding capability. Building on observations from images, we now directly engage in video QA tasks. Table 10 presents the results on video QA across four representative datasets. Interestingly, while PG-Video-LLaVA has video grounding capabilities, it does not show better results than Video-LLaVA, which lacks grounding. However, our VITRON achieves superior performance. This indirectly proves that our system possesses more accurate video grounding capabilities (as previously demonstrated in Table 4), aiding in better video semantics understanding. Additionally, we showcase a visualization process illustrating how VITRON accomplishes fine-grained visual understanding through video grounding tasks in Figure 6. VITRON’s precise perception of video details ensures a more accurate understanding of content semantics, leading to stronger video understanding capabilities.

Method	Grounding	MSVD-QA [107]		MSRVTT-QA [108]		TGIF-QA [51]		ActivityNet-QA [116]	
		Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
VideoChat [46]	✗	56.3	2.8	45.0	2.5	34.4	2.3	-	2.2
LLaMA-Adapter [28]	✗	54.9	3.1	43.8	2.7	54.3	3.3	34.2	2.7
Video-LLaMA [120]	✗	51.6	2.5	29.6	1.8	51.4	3.4	12.4	1.1
Video-ChatGPT [61]	✗	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7
Video-LLaVA [52]	✗	70.7	3.9	59.2	3.5	70.0	4.0	45.3	3.3
PG-Video-LLaVA [67]	✓	64.1	3.7	51.6	3.3	66.8	3.8	39.9	3.3
VITRON	✓	74.3	3.9	60.8	3.6	71.5	4.0	49.4	3.6

Table 10: Results (accuracy and confidence Score) on video QA.

5.4 Vision Generation

Method	FID (↓)
GLIDE [68]	12.24
SD [80]	11.21
NEX-T-GPT [104]	11.28
Emu [90]	11.66
GILL [39]	12.20
DreamLLM [23]	8.46
VITRON	8.34

Table 11: Text-to-Image generation on COCO-caption data [53].

Method	FID (↓)	CLIPSIM (↑)
CogVideo [31]	23.59	0.2631
MakeVideo [86]	13.17	0.3049
Latent-VDM [80]	14.25	0.2756
Latent-Shift [2]	15.23	0.2773
CoDi [91]	—	0.2890
NEX-T-GPT [104]	13.04	0.3085
VITRON	11.42	0.3260

Table 12: Text-to-Video generation on MSR-VTT [108].

Method	FVD (↓)	IS (↑)
AnimateAny [20]	642.64	63.87
DynamiCrafter [106]	404.50	41.97
SEINE [15]	306.49	54.02
VideoCrafter1 [12]	297.62	50.88
VITRON	190.74	55.65

Table 13: Image-to-Video generation on UCF101 [87].

Next, we assess our system’s capabilities in vision generation, focusing on three of the most representative types of generation tasks: text-to-image generation, text-to-video generation, and image-to-video generation. These tasks broadly cover the spectrum of image generation requirements. Tables 11, 12, and 13 showcase how our VITRON performs in comparison to other SoTA systems, including both MLLM and non-MLLM synthesizers. The results clearly demonstrate that VITRON outperforms on all three tasks. For instance, in both text-to-image and text-to-video generation tasks, VITRON shows more advanced performance compared to NEX-T-GPT. Similarly, in the image-to-video generation task, VITRON still outshines the SoTA baseline, VideoCrafter1, showcasing superior results.

Figure 7 illustrates the process of vision generation across different modalities, including text, image, and video. Initially, users start with a basic text command, and VITRON is capable of transforming a simple idea into a detailed video. However, if users are not satisfied with the video generated directly from text, they can first generate an image from text, then fine-tune or edit this image, and finally

create a satisfying video based on the adjusted image. Our VITRON, thanks to its robust interactive capability via multi-turn dialogue, enables users to perform a series of consecutive operations, ultimately facilitating smooth content creation. This fully helps meet the demands of real-world application scenarios.

5.5 Vision Editing

Finally, we test the vision editing, examining both the image and video editing capabilities.

Image Editing. We use the MagicBrush dataset [121], which challenges models with an editing query that demands a series of complex edits to an image. These edits include removing, changing, inpainting, and adding elements. Since there are currently no MLLM systems that support image editing, our comparison is limited to non-LLM expert systems. In Table 14, we present the performance of different models across various metrics. Except for the L1 metric, VITRON demonstrates stronger performance on CLIP metrics, indicating its stable image editing capabilities.

Method	CLIP _{dir} (↑)	CLIP _{img} (↑)	CLIP _{out} (↑)	L1 (↓)
InstructPix2Pix [8]	0.115	0.837	0.245	0.093
MagicBrush [121]	0.123	0.883	0.261	0.058
PnP [94]	0.025	0.568	0.101	0.280
NT-Inv [65]	0.121	0.752	0.263	0.077
Emu-Edit [85]	0.135	0.897	0.261	0.052
VITRON	0.138	0.904	0.265	0.063

Table 14: Image editing results on MagicBrush [121].

Method	Target-Editing	NonTarget-Unediting
Text2LIVE [6]	4.5	1.3
Tune-A-Video [101]	7.8	4.6
VITRON	8.7	7.9

Table 15: Human evaluation on video editing.

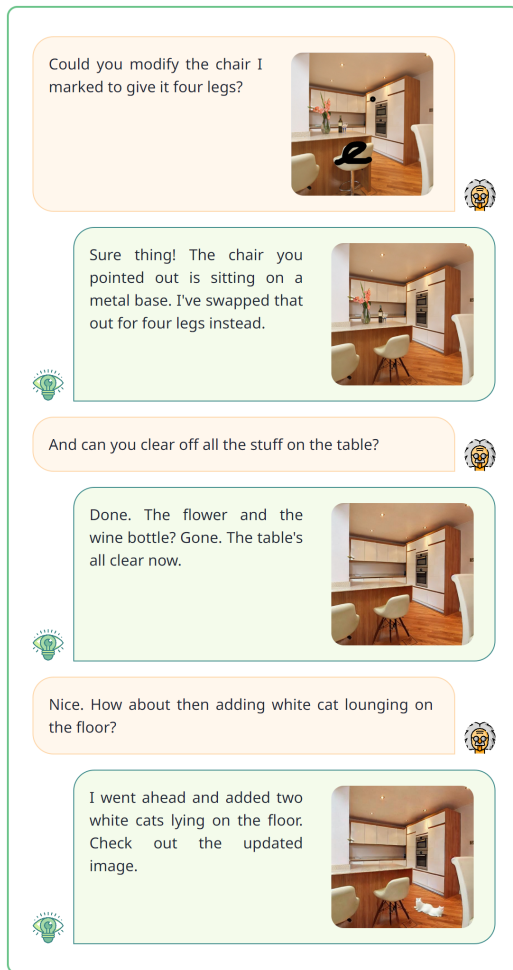


Figure 8: Demonstration of image editing.

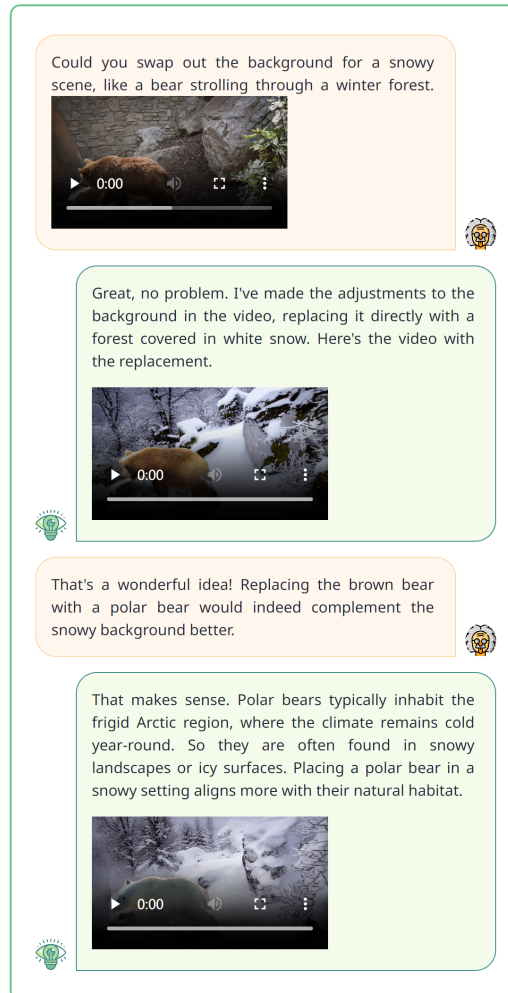


Figure 9: Demonstration of video editing.

We then showcase the specific process of this image editing, as illustrated in Figure 8. VITRON is capable of accepting different forms of user inputs (textual instruction or sketch) for precise image edits. It maintains contextual consistency throughout a series of sequential editing operations, ultimately achieving satisfactory results that meet the user’s expectations.

Video Editing. For video editing, the community currently lacks a standardized benchmark and evaluation method akin to those for image editing. Therefore, we opted for a manual evaluation approach. We asked different video editing systems to edit the same video based on the same query, after which five individuals were asked to score the edited videos. The evaluation focused on 1) the success of target content modifications and 2) the faithfulness/fidelity of non-target content. Table 15 presents the manual evaluation results for video editing. It is clear that VITRON outperforms the two baseline systems in both respects, showcasing superior video editing capabilities. Following this, we visualized the process of video editing by VITRON. Figure 9 illustrates this process. VITRON competently handles video editing tasks, including modifications to the content’s subject, and changes to the video’s style, etc.

6 Conclusion

In this work, we present VITRON, a first unified pixel-level vision LLM for seamlessly understanding (perceiving and reasoning), generating, segmenting (grounding and tracking), and editing (inpainting) both images and videos. Based on an LLM backbone, VITRON integrates image, video and regional visual encoders in its frontend modules, and also adopts a text-centered invocation approach for the integration of various SoTA modules for supporting a spectrum of vision end tasks. Then, the joint vision-language alignment learning and fine-grained instruction tuning enable VITRON precise pixel-aware perception. The invocation-oriented instruction tuning is further proposed for flexible and accurate downstream visual module invocation. On 12 visual tasks across 22 datasets, VITRON exhibits extensive capabilities in visual segmentation, fine-grained vision understanding, generation, and editing. Also various demo visualizations reflect VITRON’s superior user-friendly interactivity and visual manipulation abilities. Overall, this research showcases the great potential to build a vision-language generalist that can advance toward a more unified and interactive AI, paving the way for future vision research.

7 Potential Limitations and Future Work

System Architecture. This paper introduces a unified vision LLM capable of supporting nearly all types of vision operations. However, our current system still employs a semi-joint, semi-agent approach for external tool invocation. While this invocation-based approach facilitates the expansion and replacement of potential modules, it also means that backend modules do not participate in learning within a pipeline structure. This limitation is not conducive to the full-scale improvement of the system, meaning the performance ceiling for different vision tasks will be constrained by the backend modules. Future work should aim to integrate various vision task modules into a cohesive unit. Achieving a unified understanding and output for both images and videos, alongside supporting generation and editing capabilities through a single generative paradigm, remains a challenge.

User Interactivity. Unlike previous models specialized in singular vision tasks (e.g., Stable Diffusion and SEEM), our system aims to facilitate deep interactions between the LLM and users, akin to the proprietary DALL-E system in the industry. Achieving optimal interactivity is a core goal of this work. VITRON leverages an existing language-based LLM, combined with appropriate instruction tuning, to enable some degree of interaction. For instance, the system can flexibly respond to user inputs with any intended message, producing corresponding vision operation results without requiring user inputs to match backend module conditions precisely. However, our work still has significant room to enhance interactivity. For example, drawing inspiration from the closed-source Midjourney system, regardless of the decisions made by the LLM at each step, the system should actively provide feedback to users to ensure that its actions and decisions align with user intentions.

Modal Capabilities. Currently, VITRON incorporates a 7B vicuna backbone LLM, which imposes certain limits on its capacity to comprehend language, images, and videos. Building on the first point regarding system architecture, future efforts could develop a comprehensive end-to-end system and expand the model’s size to achieve a more thorough and comprehensive understanding of vision. Additionally, efforts should be made to enable the LLM to fully unify the understanding of image and video modalities.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the NeurIPS*, 2022.
- [2] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *CoRR*, abs/2304.08477, 2023.
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the CVPR*, pages 6077–6086, 2018.
- [4] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023.
- [5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the ICCV*, pages 1708–1718, 2021.
- [6] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022.
- [7] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, page 8, 2023.
- [8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.
- [9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the ECCV*, 2020.
- [10] Cersense. Zeroscope: Diffusion-based text-to-video synthesis. 2023. URL <https://huggingface.co/cersense>.
- [11] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023.
- [12] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.
- [13] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.
- [14] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- [15] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.
- [16] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.
- [17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90/2023.

- [18] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022.
- [19] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023.
- [20] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Fine-grained open domain image animation with motion guidance. *arXiv preprint arXiv:2311.12886*, 2023.
- [21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [23] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- [24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the ICLR*, 2021.
- [25] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019.
- [26] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023.
- [27] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *CoRR*, abs/2212.05032, 2022.
- [28] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- [29] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the CVPR*, pages 18941–18951, 2023.
- [30] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- [31] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *CoRR*, abs/2205.15868, 2022.
- [32] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *CoRR*, 2021.
- [33] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the ICLR*, 2022.
- [34] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1562–1577, 2019.
- [35] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

- [36] 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)*, pages 787–798, 2014.
- [37] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.
- [38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
- [39] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2023.
- [40] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [42] 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.
- [43] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [44] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the ICML*, pages 12888–12900, 2022.
- [45] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML*, pages 19730–19742, 2023.
- [46] Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023.
- [47] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1332–1341, 2022.
- [48] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? *arXiv preprint arXiv:2401.10229*, 2024.
- [49] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the CVPR*, pages 2918–2927, 2022.
- [50] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.
- [51] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.
- [52] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.
- [53] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proceedings of the ECCV*, pages 740–755, 2014.
- [54] 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.

- [55] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.
- [56] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *Proceedings of the ICML*, pages 21450–21474, 2023.
- [57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023.
- [58] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023.
- [59] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [60] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.
- [61] Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424, 2023.
- [62] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.
- [63] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.
- [64] Victor Siemen Janusz Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? In *Proceedings of the ACL*, pages 504–515, 2020.
- [65] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.
- [66] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.
- [67] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023.
- [68] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the ICML*, pages 16784–16804, 2022.
- [69] OpenAI. Introducing chatgpt. 2022.
- [70] OpenAI. Gpt-4 technical report. 2022.
- [71] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, pages 192–202, 2022.
- [72] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.
- [73] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [74] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

- [75] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024.
- [76] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from LLM for text-to-image generation. *CoRR*, abs/2308.05095, 2023.
- [77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the ICML*, pages 8748–8763, 2021.
- [78] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023.
- [79] 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.
- [80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the CVPR*, pages 10674–10685, 2022.
- [81] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the MICCAI*, pages 234–241, 2015.
- [82] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [83] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.
- [84] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the ACL*, pages 2556–2565, 2018.
- [85] 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. *arXiv preprint arXiv:2311.10089*, 2023.
- [86] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *CoRR*, abs/2209.14792, 2022.
- [87] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [88] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023.
- [89] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022.
- [90] Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.
- [91] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *CoRR*, abs/2305.11846, 2023.
- [92] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.
- [93] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

- [94] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.
- [95] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: extended textual conditioning in text-to-image generation. *CoRR*, abs/2303.09522, 2023.
- [96] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Chuanxin Tang, Xiyang Dai, Yucheng Zhao, Yujia Xie, Lu Yuan, and Yu-Gang Jiang. Look before you match: Instance understanding matters in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2268–2278, 2023.
- [97] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
- [98] Xinyu Wang, Bohan Zhuang, and Qi Wu. Modaverse: Efficiently transforming modalities with llms. *arXiv preprint arXiv:2401.06395*, 2024.
- [99] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for instruction-followed understanding and safety-aware generation. *arXiv preprint arXiv:2311.16511*, 2023.
- [100] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [101] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *CoRR*, abs/2212.11565, 2022.
- [102] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- [103] Jianzong Wu, Xiangtai Li, Chenyang Si, Shangchen Zhou, Jingkang Yang, Jiangning Zhang, Yining Li, Kai Chen, Yunhai Tong, Ziwei Liu, et al. Towards language-driven video inpainting via multimodal large language models. *arXiv preprint arXiv:2401.10226*, 2024.
- [104] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.
- [105] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI*, pages 2804–2812, 2022.
- [106] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.
- [107] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.
- [108] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the CVPR*, pages 5288–5296, 2016.
- [109] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.
- [110] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the ICCV*, pages 1666–1677, 2021.
- [111] 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.

- [112] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of Imms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1, 2023.
- [113] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022.
- [114] Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. Unified language representation for question answering over text, tables, and images. *CoRR*, abs/2306.16762, 2023.
- [115] 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.
- [116] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.
- [117] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. *arXiv preprint arXiv:2312.10032*, 2023.
- [118] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.
- [119] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Vptrans: Transfer visual prompt generator across llms. *Advances in Neural Information Processing Systems*, 36, 2024.
- [120] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858, 2023.
- [121] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.
- [122] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.
- [123] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.
- [124] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.
- [125] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.
- [126] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023.
- [127] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.