

Merlin: Empowering Multimodal LLMs with Foresight Minds

En Yu^{1*}, Liang Zhao^{2*}, Yana Wei³, Jinrong Yang³, Dongming Wu⁴, Lingyu Kong⁵, Haoran Wei², Tiancai Wang², Zheng Ge², Xiangyu Zhang², and Wenbing Tao¹✉

¹ Huazhong University of Science and Technology

² MEGVII Technology

³ ShanghaiTech University

⁴ Beijing Institute of Technology

⁵ University of Chinese Academy of Sciences

{yuen, wenbingtao}@hust.edu.cn



Fig. 1: Demo cases presentation of Merlin. Here we showcase several main capabilities of our built Multimodal Large Language Model (MLLM), *Merlin*. Notably, in the dialogue, the words marked with colors correspond to the trajectory outputs of the targets in the image. To save space, we highlight them using the same colors.

Abstract. Humans can foresee the future based on present observations, a skill we term as *foresight minds*. However, this capability remains under-explored within existing MLLMs, hindering their capacity to understand intentions behind subjects. To address this, we integrate the future modeling into MLLMs. By utilizing the **trajectory**, a highly structured representation, as a learning objective, we aim to equip the model to understand spatiotemporal dynamics. Inspired by the learning

* Equal Contribution, ✉ Corresponding Author

paradigm of LLMs, we first propose *Foresight Pre-Training (FPT)* that jointly learns various tasks centered on trajectories, enabling MLLMs to predict entire trajectories from a given initial observation. Then, we propose *Foresight Instruction-Tuning (FIT)* that requires MLLMs to reason about potential future events based on predicted trajectories. Aided by FPT and FIT, we build an unified MLLM named *Merlin* that supports complex future reasoning. Experiments show Merlin’s foresight minds with impressive performance on both future reasoning and visual comprehension tasks. Project page: <https://ahnsun.github.io/merlin>.

Keywords: Multimodal Large Language Model · Future Reasoning

1 Introduction

Human beings can predict future events or outcomes based on current observations, known in neuroscience theory as *predictive processing* [19]. In this paper, we refer to this ability as *foresight minds*, which involves the use of past experiences, knowledge, sensory information, and probabilistic reasoning to generate expectations about future events. In the artificial intelligence (AI) domain, the capability to predict future events is an important topic towards the realization of artificial general intelligence (AGI).

Recent advancements in Multimodal Large Language Models (MLLMs), such as GPT-4V [48] and Bard [2], have shown significant potential in image understanding and logical reasoning. Despite these achievements, these models struggle to foresee future events based on current image observations. Even provided with additional observations, like sequences of multiple frames, the current MLLM models still struggle to adequately analyze and infer specific target behaviors, such as predicting object movements or interactions (shown in Figure 2). On the contrary, human can reason the future to some extent based on the observed current state [5, 54], which shows powerful foresight minds.

To mitigate this existing deficiency in MLLMs, we start from dividing human’s process of foreseeing the future into a two-stage system [30, 54]: (1) observing the dynamic clues of the subject and then (2) analyzing the behavior pattern and reasoning what might happen according to the observation. For instance, while watching a basketball game, people will first observe the moving players on the court, and then forecast the specific player’s forthcoming actions, e.g., shooting, slam-dunking, or passing, by analyzing the current states and movement patterns of the players. Compare this system to current MLLMs, we find that MLLMs can complete the second stage well, thanks to the powerful logical reasoning ability of LLM [50, 70]. Therefore the key challenge is the first stage. That is, *how to make MLLM acquire correctly spatiotemporal dynamics from the multi-image observation?*

Explicitly modeling next frames (e.g., reconstructing next frames [12, 74]) can be a straightforward way. However, it can be hard to directly extract dynamic clues from the redundant visual information [25], especially from video sequences.

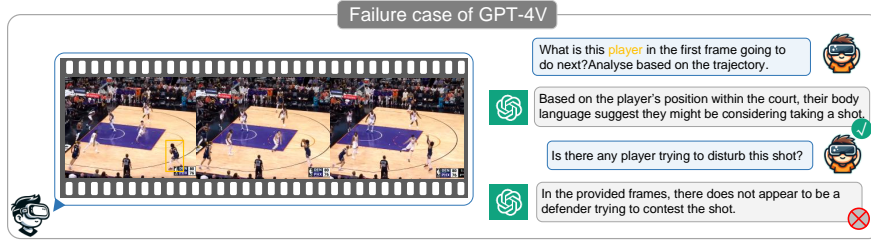


Fig. 2: Failure case of GPT-4V about future reasoning.

It is necessary to construct a suitable learning objective to assist MLLM in obtaining dynamic clues about the specific subjects. To this end, we point out that *trajectory*, as a highly structured representation, is a good learning objective which can link the temporal contexts between the past and the future.

Based on this insight, we propose to model the future to empower existing MLLMs with “foresight minds”. Following the modern learning paradigm of LLMs, our future reasoning learning process includes two stages: (1) *Foresight Pre-Training (FPT)*, a paradigm that causally models the temporal trajectories, which interleave with multi-frame images. The model starts with the initial observation of one or multiple subjects in the first frame as the query and then is required to predict the whole trajectory. Notably, we introduce various tasks containing richly labeled data [18, 26, 31, 58, 63, 80], including object detection, object tracking, etc., to perform multitask learning. And samples from these tasks are properly formatted to ensure coordinated pre-training. (2) *Future Instruction-Tuning (FIT)*, then, considers the trajectory modeling bestowed by FPT as a bridge in the logical chain of future reasoning. Simply put, when querying an MLLM, it must articulate its reasoning in conjunction with the trajectory for each object referenced. This method, as a form of *Trajectory Chain-of-Thought*, effectively narrows the gap between trajectory perception and predictive future reasoning, thereby fully unleashing model’s foresight minds.

Aided by the above future modeling technologies, we provide **Merlin**⁶, a novel and unified MLLM capable of handling inputs and outputs of spatial coordinates or tracklets from single image or multiple frames. Moreover, Merlin is adept at performing inductive reasoning about future events based on current observational results. To demonstrate this, we provide several real dialogues between users and Merlin, as displayed in the Figure 1. Unlike the previous MLLMs [41, 84, 87] which only supported interaction with a single image, Merlin not only provides a richer multi-image interaction, but also on this basis, is capable of executing unique and powerful future reasoning.

We construct a new future reasoning benchmark to evaluate Merlin’s logical reasoning and future prediction abilities. The results, which significantly surpass

⁶ **Merlin** is a legendary character in the tales of King Arthur, renowned as a powerful wizard and a wise counselor in the Arthurian legends. He is depicted as having the power to foresee future events and has a deep understanding of fate and destiny.

previous baselines [10, 39, 41, 75], demonstrate Merlin’s stunning performance in future reasoning. We further reveal Merlin’s exceptional performance in general visual understanding. Through analysis in scenarios such as VQA (Visual Question Answering) [23, 28], comprehensive understanding [42, 79], and hallucination [37], we unexpectedly discovered that our proposed novel paradigm of future learning aids MLLMs in gaining a deeper understanding of images. We believe this brings new insights for the training of future MLLMs.

2 Related Work

2.1 Large Language Models

Large Language Models (LLMs) have gained significant attention due to their capabilities in language generation and logical reasoning. Pioneering models like BERT [15], GPT-2 [52], and T5 [53] laid the groundwork, but GPT-3 [8], the first model with a 175 billion parameter size, made notable strides, demonstrating strong zero-shot performance. An emergent ability, wherein model size scaling results in significant language capability improvements, was also observed in LLMs. This was further facilitated by InstructGPT [49] and ChatGPT [47] using Reinforcement Learning with Human Feedback (RLHF) on GPT-3. These advancements led to what’s called LLMs’ “iPhone moment”. Following GPT’s success, several open-source LLMs, including OPT [83], LLaMA [65], and GLM [81], have been proposed, showing similar performance to GPT-3. Models like Alpaca [64] and Vicuna [11] illustrate the application of these LLMs, using a self-instruct framework to construct excellent dialogue models.

2.2 Multimodal Large Language Models

The advancements in LLMs [47, 65, 66] have projected a promising path towards artificial general intelligence (AGI). This has incited interest in developing multimodal versions of these models. Current Multi-modal Large Language Models (MLLMs) harness the potential of cross-modal transfer technologies. These models consolidate multiple modalities into a unified language semantic space, and then employ autoregressive language models as decoders for language outputs. Models like Flamingo [1] have adopted a gated cross-attention mechanism, trained on billions of image-text pairs, to align visual and linguistic modalities, showing impressive performance on few-shot learning tasks. Similarly, BLIP-2 [36] introduced the Q-Former to align visual features more effectively with language space. The LLaVA series [39, 41, 68] further enhanced this process by using simply a MLP in place of the Q-Former and designing a two-stage instruction-tuning procedure. Apart from creating general MLLMs, techniques have also been developed for visual-interactive multimodal comprehension, involving the precise tuning of referring instructions [10, 82, 85]. Furthermore, another interesting direction in MLLM research involves integrating MLLMs for cross-modal generation [16, 20, 32] by using text-to-image models such as Stable Diffusion.

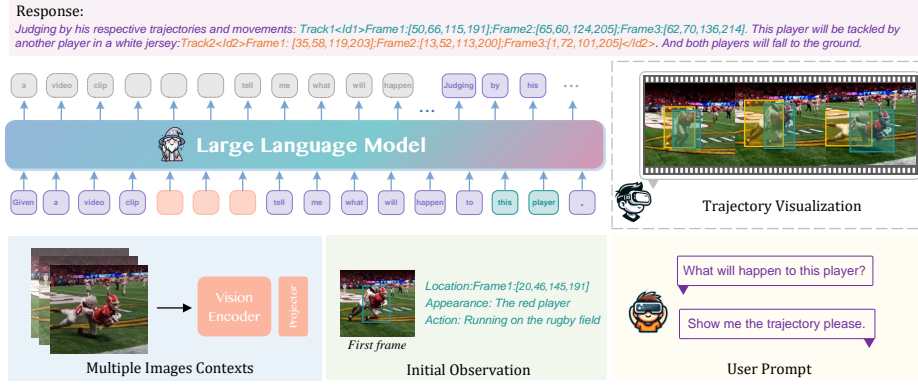


Fig. 3: Overall pipeline of Merlin. The architecture of Merlin consists of three main components: (1) an image encoder, (2) a large language model, and (3) a modality-align projector. **Bottom:** The diverse input format that supports multiple-image contexts, initial observation and the specific user prompt. **Top:** The model response including the predicted trajectory and the future reasoning.

3 Methodology

3.1 Overall Architecture

Merlin is designed to unlock the foresight minds based on observations from single images and multi-frame video clips. In order to accomplish this, images and videos are comprehensively represented through a series of visual tokens, which are then integrated into the language sequence that can be comprehended by Large Language Models (LLMs) in a unified framework. Specifically, Merlin consists of an image encoder, a decoder-only LLM, and a modality alignment block as illustrated in Figure 3. Following prevalent practice [10, 39, 41, 87], we opt for the pre-trained CLIP [51] ViT-L/14 [17] as the visual encoder and Vicuna-7B v1.5 [11] as the large language decoder. For more details, please refer to our supplementary materials.

To provide enough visual information and details, the input images are resized to a resolution of 448×448 . At this juncture, the visual encoder iteratively attends to $(448/14)^2$ uniformly divided image patches, yielding 1024 encoded tokens. Considering the limited context length of LLMs and addressing the substantial computational challenges posed by high resolution and multi-frame context modeling, we simply utilize a 2D convolution to achieve both dimension projection and token aggregation [7, 55].

We choose 2D convolution over 1D linear layers [10, 39, 41] or cross-attention layers [4, 36, 87] as connector for the following reasons: (1) 2D convolution clusters local visual tokens on a spatial scale [22], effectively achieving a one-step transformation from spatial to channel information; (2) The good convergence properties [29, 60] of 2D convolution compared with cross-attention lay a solid foundation for foresight learning in a two-step training approach.

3.2 Foresight Pre-Training

Generative Pre-Training (GPT) [8, 47, 48] serves as the cornerstone of this generation’s Language Models (LLMs). Through learning to predict next token, the model efficiently condenses data, thereby yielding emergent forms of intelligence [69]. In this context, a very natural approach to enhance the model’s perception of the dynamic clues across multiple frames is to *explicitly model the next frame* (or image). However, due to the high redundancy in multi-frame visual information, the truly next-frame prediction remains a significant challenge to date. A better approach at this juncture is to *implicitly model high semantic information in the label space* (such as categories, bounding boxes) on a frame-by-frame basis. Temporally, this label information forms a **trajectory**, a highly structured representation. Causally modeling the trajectory in conjunction with each frame of image helps to connect the past and present in time, thus enabling the model to perceive the future.

To this end, we propose the Foresight Pre-Training, a way of *causally modeling the trajectories interleaved with multi-frame images*, to empower the MLLM with the capacity of perceive the dynamic clues, and ultimately achieving future reasoning. Specifically, given a video clip including several frames, we first give the model the observation of the *first frame*, then we require the model to predict the *whole trajectory* of the concerned subject in this video conditioned on the initial observation. Notably, the observation of the first frame can be the description or simple position of the concerned object. Formally,

$$P(Y|X) \sim P(Y|\{X_1, X_2, \dots\}, O_{first}), \quad (1)$$

where X_i denotes the i^{th} frame and O_{first} is the first frame observation, Y refers to the trajectory of the subject in O_{first} within the frame sequence. The observation and the raw frames will be regarded as the condition to prompt MLLM to predict the trajectory.

Data Construction. We first aggregate all valuable multimodal information from diverse data resources and then properly organize them for multi-task foresight pre-training. Specifically, for each sample instance \mathbf{I} , we first collect its multimodal information including consecutive multi-frame images $\{X_1, X_2, \dots\}$, subject observations from the *first* frame O_{first} , and subject trajectory Y constructed from *all* frames. Formally,

$$\mathbf{I} = \{\{X_1, X_2, \dots\}, O_{first}, Y\}. \quad (2)$$

We categorize observations of one subject of the first frame into three main types: **location** description, **appearance** description and **action** description. Then we *randomly* selected one of these observations of a particular subject in the first frame as the query object. (It is also feasible to select observations of multiple attributes as the query according to the characteristics of the dataset)

To better unleash the powerful generative modeling capacity of LLM, we construct this query process as a type of conversation. Here is an example of the constructed data shown in Figure 4. In this case, we want to query the subject




An Example of FPT dialogue	
Interleaved Multiple Images: Given a video clip including: frame1:  \n, frame2:  \n, frame3:  \n	Initial Observation: Randomly choice Location: Frame1:[562, 342, 926, 561] Appearance: A panda on the right side. Action: A panda is lying on the ground.
Dialogue: Question: <i>Interleaved Multiple Images</i> , can you tell me the trajectory according to the <i>initial observation</i> ? To respond correctly, utilize the specified <code><Idi>Frame t:[xmin,ymin,xmax,ymax]</Idi></code> format.	
Answer: Its trajectory is <code><Id1>Frame1:[562,342,926,561]; Frame2: [576,334,960,568];Frame3:[632,366,979,589]</Id1></code> .	

Fig. 4: One example to illustrate the multi-modality pretraining dataset. The top block shows the provided contexts including the multiple images contexts and initial observation (box, appearance and action) about the subject to prompt the LLM. The bottom block shows the dialogue including question and answer.

— *the panda on the right* — with the randomly select observation, and expect the answer with the movement trajectory of this panda across multiple frames. To model this process, we convert the query to question and trajectory to answer with proper natural language for embellishment.

Overall, the aforementioned process of dialogization roughly follows these three principles: (1) **Precise definition of task prompts and answer formats.** In particular, we use a task prompt to tell MLLM *what specific task to do* (detect or track), and also *specified the answer format* with accurate descriptions in each question. In this way, different types of tasks can be flexibly organized together without compromising the general language ability. (2) **Clear indication of multimodal information.** Concretely, for each group of image tokens, we add a *special frame indicator* in front of them, i.e., *frame1:<image> and frame2:<image>*, so as to help MLLM better focus on the corresponding image. (3) **Interleaving of frames and observations.** For the same identity, we interleave the frames in which it appears with its positional observations, and enclose them with two ID tokens (i.e. `<Idi>` and `</Idi>`) to construct a trajectory. We believe that this interleaved organization helps in *generatively training to model causality within the trajectory*, while the ID tokens ensures that the model can distinguish among different identity objects.

Training Details. The objective in this stage is to initially endow MLLMs with the capacity of modeling the spatiotemporal dynamics across multi-frame images, while ensuring that its general language capabilities do not diminish. Previous practices [4, 40, 41] typically conducting a separate modality alignment training phase following a multi-task pre-training stage, which however, complicates the training process and data construction. In this paper, we directly incorporate both of them into one stage, and *unfreeze all modules during pre-training*. This is because that we believe the MLLMs are sufficiently powerful to concurrently handle the learning of general multimodal capabilities and multi-task specific abilities *under proper guidance*. Furthermore, we mix a large amount

of image-text pairs and rich-annotated conversation data (formatted according to the above method) from diverse data sources [18,26,31,46,59,61,63,80] to conduct multi-task learning. In doing so, not only endows the model with foresight minds but also ensures its multimodal alignment.

3.3 Foresight Instruction Tuning

Although Foresight Pre-Training equips the model with the ability to observe dynamic clues across multiple frames, it still falls short of true foresight minds. This is because models typically struggle to effectively transform such observations into successful future reasoning [67,86].

Recent work [50,86] has highlighted that Chain-of-Thought (CoT) [70] is crucial in bridge the gap between the observations and actions of MLLMs with theory of mind [56,67]. Meanwhile, several prior studies [10,84] have also demonstrated that prompts indicating position (such as bounding boxes or points) — a principle analogous to CoT — can concentrate an MLLM’s attention on the relevant area, leading to more accurate dialogues and reducing the likelihood of visual hallucination. Drawing inspiration from these findings, we conduct the Foresight Instruction Training (FIT) building upon the foundation of FPT to further enhance the model’s future reasoning capability. In specific, building on the trajectory generating powered by FPT, we further union the trajectories to generatively rationalize the forthcoming events. Mathematically,

$$P(Z|X, Y) \sim P(Z|\{X_1, X_2, \dots\}, O_{first}, Y), \quad (3)$$

where Z refers to the future observation which is deduced from observations in *each* frame. It can be actions, events, trends, or simply likelihoods. In this context, multi-frame images, in conjunction with the first subject observation, and the trajectory of the same subject across all frames, serve as the union condition to prompt MLLM to causally predict the future. This way, akin to ***Trajectory Chain-of-Thought***, effectively bridges the gap between trajectory perception and predictive future reasoning, thereby fully unleashing model’s foresight minds.

Data Construction. The specific data construction method is similar to FPT, but on this basis, we also deduce a future observation Z from the information across multiple frames and append it after the trajectory in the answer. Formally,

$$\mathbf{I} = \{\{X_1, X_2, \dots\}, O_{first}, Y, Z\}. \quad (4)$$

Practically, in this paper, we constitute future observations based on multi-frame, multi-target action descriptions combined with human priors, and further process them with GPT-4 [48] to ultimately form reasonable future inferences. More details are provided in the supplementary materials.

Figure 3 provides an illustrative example of FIT, when a user questions Merlin about the future of a player in red attire, Merlin initially presents the observed trajectory of the concerned player, followed by the trajectory of another player in white. Using these trajectories, Merlin deduces that *the player in white is likely to tackle the one in red, resulting in both players falling to the ground*.

Training Details. We freeze the vision encoder and keep the convolutional projector and the LLM unfrozen in this stage. On this basis, we primarily adopt the open-source instruction tuning datasets, *i.e.*, LLaVA-665K [39], for building the basic ability for multi-round visual-language conversation. For further unleashing the foresight minds of model, we first uniformly sample a certain number of multitask dialogues in FPT, in order to maintain the model’s capacity of modeling the dynamic clues across multi-frame images. In addition, we also sample data from three specific scenario datasets [38, 45, 71] and construct around 30K FIT conversations based on the aforementioned data construction process.

4 Experiment

4.1 Experimental Settings

Datasets. For the foresight pre-training (FPT) stage, we first use 10M image-text pairs sampled from LAION400M [57] to ensure multimodal alignment. On this basis, we gather various open-source datasets with rich annotations to conduct multi-task learning, including (1) **object detection** datasets: Object365 [59] and OpenImage [33]; (2) **tracking** datasets: LaSOT [18], GOT10K [26], MOT17 [46], DanceTrack [63] and SOMPT22 [61]; (3) **grounding** dataset: RefCOCO [31]; (4) **object relation** dataset: VCR [80]. For these data, as described in Section 3.2, we apply strict task definitions and format specifications, and re-organize them in the form of interleaved frames and observations. Ultimately, we obtain approximately 5M question-answer data, which are mixed with 10M paired data for foresight pre-training.

For the foresight instruction-tuning (FIT) stage, we mix approximately 730K conversation data, including (1) open-source instruction-tuning data LLaVA-665K [39], which integrates a series of VQA datasets [62] and multi-round conversation datasets [41]; (2) around 30K FIT multi-frame conversations constructed from three specific scenarios including MultiSports [38], TITAN [45] and STAR [71] based on the data construction method described in Section 3.3; (3) nearly 40K randomly sampled FPT multi-task data. For more details of the datasets, please refer to the supplementary materials.

Implementation Details. As outlined in Section 3.1, Merlin utilizes the CLIP-ViT-L/14 [51] as its vision encoder for image encoding and the open-source Vicuna-7B v1.5 [11] for foresight decoding. Between them, a 3×3 convolution layer with padding set to 1 and a stride of 2 is employed for both dimension projection and token aggregating. During the foresight pre-training, we optimize all parameters of the model, setting the learning rate to $5e-5$ and training for one epoch. In the instruction tuning stage, we freeze the visual encoder and fine-tune the parameters of the projector and LLM. In both stages, we train Merlin using the AdamW [44] optimizer and a cosine annealing scheduler [43] as the learning rate scheduler. The entire training process is conducted on 64 NVIDIA A800 GPUs, with approximately 12 hours required for pre-training and 3 hours for instruction-tuning. Additional implementation details can be found in the supplementary materials.

Table 1: The Effectiveness of Prediction Reasoning. We mainly select 5 metrics from MMBench develop and test set, respectively, including **OL**: Object localization (Prediction), **PPR**: Physical property reasoning, **FR**: Function reasoning, **IR**: Identity reasoning, and **FP**: Future prediction. **Avg.** denotes the average score. The best and second-best performances are shown in bold font and underlined respectively.

Method	LLM Size	Prediction Reasoning (Dev.)						Prediction Reasoning (Test)					
		Avg.	OL	PPR	FR	IR	FP	Avg.	OL	PPR	FR	IR	FP
InstructBLIP [13]	13B	42.0	14.8	30.7	56.8	88.9	19.0	44.4	5.7	24.0	67.3	<u>92.7</u>	32.4
MiniGPT-4 [87]	13B	43.3	28.4	30.7	49.4	86.7	21.4	48.9	21.0	35.0	67.3	90.2	31.1
OpenFlamingo [3]	7B	5.28	2.5	10.7	8.6	2.2	2.4	11.5	2.9	14.0	9.3	11.0	20.3
MMGPT [42]	7B	19.5	1.2	24.0	9.9	60.0	2.4	16.8	3.8	13.0	12.1	52.4	2.7
MiniGPT-4 [87]	7B	26.8	7.4	14.7	19.8	80.0	11.9	27.9	8.6	13.0	29.9	61.0	27.0
InstructBLIP [13]	7B	34.8	6.2	17.3	51.9	84.4	14.3	39.0	2.9	17.0	52.3	78.0	44.6
LLaVA [41]	7B	38.7	8.6	25.3	53.1	77.8	28.6	39.7	13.3	35.0	48.6	82.9	18.9
mPLUG-Owl [75]	7B	41.0	18.5	18.7	66.7	86.7	14.3	45.9	16.2	23.0	59.8	91.5	39.2
Shikra [10]	7B	51.5	32.1	30.7	63.0	88.9	42.9	<u>60.0</u>	27.6	<u>50.0</u>	<u>70.1</u>	<u>92.7</u>	59.5
Kosmos-2 [27]	1.6B	54.4	38.3	33.3	56.8	91.1	<u>52.4</u>	58.2	<u>40.4</u>	30.0	65.4	89.0	66.2
LLaVA-1.5 [39]	7B	<u>59.6</u>	43.2	<u>52.0</u>	<u>71.6</u>	<u>93.3</u>	38.1	-	-	-	-	-	-
Merlin (Ours)	7B	64.4	<u>42.0</u>	54.7	72.8	97.8	54.8	66.5	41.3	51.0	83.0	97.6	<u>59.7</u>

4.2 Properties Evaluation of Foresight Minds

In this section, we mainly verify the foresight minds (future reasoning) of Merlin from two aspects, *i.e.*, prediction reasoning and identity association ability, where the former focuses on forecasting and reasoning location, events or behavior based on image observation, and the latter focuses on the model’s ability to establish subject identity associations across multiple frames to obtain dynamic clues for future reasoning.

Prediction Reasoning. To evaluate this ability, we probe this ability based on the several sub-tasks of MMBench [42]. MMBench provides a comprehensive evaluation system to assess various capabilities of MLLM, with some metrics focusing on the model’s prediction and reasoning capabilities. To this end, we pick out these metrics to establish this new future reasoning benchmark and compare Merlin with the existing SOTA models. As shown in Table 1, Merlin achieves the best overall performance (64.4 average score on the development set and 66.5 average score on the test set). Moreover, it obtains the best in 8/10 indicators and ranks second in all other indicators, which favorably demonstrates Merlin’s strong prediction and reasoning ability.

Identity Association. We examine this ability by evaluating the performance of object-tracking tasks [72, 76–78], which can comprehensively demonstrate object association and prediction capabilities. To this end, we evaluate Merlin in existing mainstream tracking benchmarks, *i.e.*, LaSOT [18] and GOT10K [26]. It is worth noting that Merlin is the *first MLLM that can also carry out tracking tasks*. As shown in Table 2, Merlin achieves comparable performance with expert models and even outperforms on some metrics. Notably, we only *sample a small amount* of tracking data to train Merlin instead of the full amount

Table 2: Comparison on main track- Table 3: Comparison with SOTA ing benchmarks. Notably, the original LLaVA-1.5 [39] model was incapable of performing tracking tasks. Therefore, we utilized the model configuration of LLaVA-1.5 and trained a version of the model with the same dataset as Merlin. [†]Includes using in-house data that is not publicly accessible.

Method	LaSOT		GOT10k			
	Success	P _{norm}	P	AO	SR _{0.5}	SR _{0.75}
<i>Specialist Models</i>						
SiamFC [6]	33.6	42.0	33.9	34.8	35.3	9.8
ATOM [14]	51.5	-	-	55.6	63.4	40.2
SiamRPN++ [35]	49.6	56.9	49.1	51.8	61.8	32.5
SiamFC++ [73]	54.4	62.3	54.7	59.5	69.5	47.9
<i>Generalist Models</i>						
LLaVA-1.5 [39]	19.4	16.5	12.8	23.5	20.2	9.7
Merlin (Ours)	39.8	40.2	38.1	51.4	55.9	42.8

Method	VQA Task		Generalist		
	GQA	VisWiz	MMB _d	MMB _t	MM-Vet
BLIP-2 [36]	41.0	19.6	-	-	22.4
InstructBLIP [13]	49.2	34.5	36.0	33.9	26.2
Shikra [10]	-	-	58.8	60.2	-
IDEFICS-9B [34]	38.4	35.5	48.2	45.3	-
IDEFICS-80B [34]	45.2	36.0	54.5	54.6	-
Qwen-VL [†] [4]	59.3	35.2	38.2	32.2	-
Qwen-VL-Chat [†] [4]	57.5	38.9	60.6	61.8	-
LLaVA-1.5 [39]	62.0	50.0	64.3	59.5	30.5
Merlin (Ours)	60.5	50.4	66.2	65.5	34.9

of data, which means LLM exhibits significant potential in handling temporal tasks, possibly because tracking, as a temporal task, can be viewed as a casually frame-level autoregressive task.

4.3 General Comprehension

In order to showcase the general multi-modal ability, we further benchmark Merlin on various VQA benchmarks and recent benchmarks proposed for evaluating the comprehensive capabilities of MLLMs.

Visual Question Answering (VQA). We first evaluate Merlin on several mainstream VQA benchmarks to reflect the perceptual abilities of MLLMs in understanding image content. As shown in Table 3, Merlin achieves competitive performance compared with existing advanced MLLMs in the selected VQA benchmarks (VQA). The results indicate that Merlin possesses strong image understanding and question-answering capabilities.

Synthetic MLLM Benchmarks. Recently, several benchmarks have been proposed to evaluate the comprehensive performance of MLLMs, encompassing diverse finer-grained scenarios including visual perception, object recognition, optical character recognition (OCR), future reasoning, and so on. In this part, we select several mainstream MLLM benchmarks to evaluate Merlin. As shown in Table 3, We present performance in accuracy on benchmarks including MM-Vet [79] and MMBench [42]. On MMBench, we report results on the both development and test sets. The results show that Merlin significantly outperforms comparative methods, even though many methods utilized a substantial amount of in-house data for pre-training, or employed several times more parameters. This implies that, while introducing foresight minds into MLLMs, we not only preserved their original visual capabilities but even *further enhanced their overall level of visual perception*.

Table 4: Zero-shot object hallucination evaluation on the COCO validation set. “Yes” represents the proportion of positive answers that the model outputs.

Method	LLM Size	Random			Popular			Adversarial		
		Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes
LLaVA [41]	13B	64.12	73.38	83.26	63.90	72.63	81.93	58.91	69.95	86.76
MiniGPT-4 [87]	13B	79.67	80.17	52.53	69.73	73.02	62.20	65.17	70.42	67.77
InstructBLIP [13]	13B	88.57	89.27	56.57	82.77	84.66	62.37	72.10	77.32	73.03
Shikra [10]	13B	86.90	86.19	43.26	83.97	83.16	45.23	83.10	82.49	46.50
MultiModal-GPT [21]	7B	50.10	66.71	99.90	50.00	66.67	100.00	50.00	66.67	100.00
mPLUG-Owl [75]	7B	53.97	68.39	95.63	50.90	66.94	98.57	50.67	66.82	98.67
LLaVA [41]	7B	72.16	78.22	76.29	61.37	71.52	85.63	58.67	70.12	88.33
LLaVA-1.5 [39]	7B	83.29	81.33	-	81.88	80.06	-	78.96	77.57	-
Qwen-VL [4]	7B	84.73	82.67	-	84.13	82.06	-	82.26	80.37	-
Merlin (Ours)	7B	91.58	91.66	49.38	89.53	89.56	50.27	84.10	84.95	55.63

Table 5: Ablation study of the proposed strategies in Merlin. (ITP: Image-text pair data, ITD: instruction-tuning data). We mainly report the AO score of GOT10k and the average score of future reasoning.

Pre-Training		Inst.-Tuning		GOT10K	Prediction Rea.
ITP	FPT-Data	ITD	FIT-Data	AO	Average_{dev}
✓	✗	✓	✗	-	59.5
✓	✗	✓	✓	-	60.7
✗	✓	✓	✓	15.5	52.8
✓	✓	✓	✗	51.4	61.2
✓	✓	✓	✓	51.4	64.4

4.4 Object Hallucination

Hallucination presents a significant challenge in existing MLLMs. This term describes the phenomenon where the generated textual content exhibits inconsistencies when compared to its corresponding image content. In this section, we present the experiments from the Polling-Based Object Probing Evaluation (POPE [37]). As demonstrated in Table 4, Merlin surpasses recent SOTA methods with clear margins. More specifically, Merlin achieves optimal performance in all metrics across three scenarios: **Random**, **Popular** and **Adversarial**, with improvements of up to **5** points compared to the highly competitive baseline Shikra [10]. Surprisingly, in multiple scenarios, the “yes” rate of Merlin is quietly closed to 50%, demonstrating its extraordinary visual perception capabilities.

We analyze this success largely owing to the proposed foresight learning (FPT and FIT). By enabling the model to learn the dynamic correspondence between trajectories across multiple images, the model has gained a *more precise ability to attend to relevant object (trajectories) contexts in the image*, which helps to better avoid misidentification and misalignment of irrelevant targets. We believe that this result will provide new thinking about addressing the issue of hallucinations in MLLM.

Table 6: Ablation studies of the model settings including resolution, vision encoder and projector of Merlin.

Exp	Resolution	Projector	Visual Encoder	Tokens Num	Prediction Rea.	Got-10K
❶	448x	Conv2d	unfrozen	256	64.4	51.4
❷	336x	Conv2d	unfrozen	256	59.8	47.3
❸	336x	MLP	unfrozen	576	58.1	23.5
❹	448x	Conv2d	frozen	256	60.8	28.4

4.5 Ablative Analysis of FPT & FIT

As introduced in Section 3.2 and Section 3.3, FPT serves as the pre-training strategy to enable MLLM to encapsulate dynamic information across frames by predicting the trajectory of the next frame. FIT is designed to activate the ability of foresight minds in a way of *Trajectory CoT* during instruction fine-tuning. To further explore the effect of FPT and FIT, we conduct an ablation study based on the established future reasoning benchmark and tracking dataset GOT10K [26]. As shown in Table 5, we mainly report the average overlap (AO) of GOT10K and the average score of future reasoning in the development set.

The results show that both FPT and FIT training strategies contribute to the improvement of the metrics. Combining both FPT and FIT, Merlin achieves the best performance which proves the effectiveness of the proposed strategies. Furthermore, we can also observe that the lack of image-text pair data during the pre-training stage considerably hampers the model’s general ability. This phenomenon supports our perspective that, during the comprehensive pre-training phase, the integration of image-text pair data is essential for maintaining modality alignment and preventing a decline in combined capabilities.

4.6 Ablative Analysis of Model Configuration

The configuration of model architecture for large-scale models is also a focal point of interest for researchers. In this subsection, we specifically investigate the impact of Merlin’s model configuration on performance. As depicted in Table 6, we focus on examining the effects of model input resolution, the visual encoder of the model, and the model’s projector on the ultimate performance of Merlin. From the experimental outcomes, we can draw the following conclusions:

- (i) High-resolution input is more conducive to visual perception and understanding tasks (row ❶ and ❷), particularly for tasks that require precise localization, such as detection and tracking.
- (ii) The primary contribution of Conv2d is the ability to compress the number of tokens efficiently and elegantly, which is crucial for supporting high-resolution images. In contrast, MLPs cannot compress tokens. This high token count hinders the training with multiple images. Moreover, more visual tokens does not improve performance in future reasoning tasks (row ❶ and ❸). We speculate that an increased number of visual tokens may lead to the sparsity of supervision.



Fig. 5: Attention map visualization. To facilitate the observation, we map the attention between the box responses and the visual tokens of each frame for visualization.

(iii) During the pre-training phase, the visual encoder should be unfrozen (row ❶ and ❹), which is beneficial for modal alignment and the expansion of the fine-grained spatial information. Similar conclusion is also claimed in [9].

4.7 Visualization Analysis

In this subsection, we visualize the attention map of Merlin to further substantiate the effectiveness of utilizing the proposed strategies. As shown in Figure 5, we select the output attention map of the middle-level layers of LLM for visualization. We can observe that the word embedding of the output trajectory coordinates can attend to the corresponding object from different frames correctly. This visualization results further prove that the trajectory representation is a good interface to enable MLLM to establish the alignment between the language description and the multi-images dynamic visual contexts. Furthermore, this effectively explains why Merlin possesses a more powerful comprehensive visual capability and a greatly lower level of hallucination compared to previous baselines. Indeed, *the trajectory-driven foresight learning allows the large language model to read images more profoundly!*

5 Limitation and Conclusion

This study highlighted an obvious deficiency in Multimodal Large-Language Models (MLLMs), specifically their ability to predict future events or outcomes based on current observations, referred as “*foresight minds*”. To address this, we serve as the first to point out that trajectory, as a highly structured representation, is a good learning objective to assist MLLM in obtaining dynamic information from the image observations. Based on this insight, we introduced a unique training method including *Foresight Pre-Training (FPT)* and *Foresight Instruction-Tuning (FIT)*. By synergizing FPT and FIT, we created **Merlin**, a unified MLLM that effectively understands and outputs spatial coordinates or tracklets from single images or multiple frames. Merlin excels at a range of traditional vision-language tasks while demonstrating powerful future reasoning capacities. Despite the substantial advancements made by Merlin, there still are some limitations, particularly in processing long sequential videos and more comprehensive future reasoning evaluation. We aspire for Merlin to guide the enhancement of more advanced MLLMs in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62176096 and Grant 61991412. The work was also supported by the Research on AI Terminal Computing Power and End-Cloud Framework (R2411B0R) of China Mobile Group Device Co., Ltd.

References

1. Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)
2. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023)
3. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
5. Bates, C., Battaglia, P.W., Yildirim, I., Tenenbaum, J.B.: Humans predict liquid dynamics using probabilistic simulation. In: CogSci (2015)
6. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCV. pp. 850–865 (2016)
7. Bolya, D., Fu, C., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: ICLR. OpenReview.net (2023)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
9. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., Florence, P., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. arXiv preprint arXiv:2401.12168 (2024)
10. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm’s referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. <https://lmsys.org/blog/2023-03-30-vicuna/> (2023)
12. Cholakov, R., Kolev, T.: Transformers predicting the future. applying attention in next-frame and time series forecasting. arXiv preprint arXiv:2108.08224 (2021)
13. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)

14. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: CVPR. pp. 4660–4669 (2019)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
16. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)
18. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019)
19. Friston, K.: The free-energy principle: a unified brain theory? *Nature reviews neuroscience* **11**(2), 127–138 (2010)
20. Ge, Y., Ge, Y., Zeng, Z., Wang, X., Shan, Y.: Planting a seed of vision in large language model. arXiv preprint arXiv:2307.08041 (2023)
21. Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790 (2023)
22. Goyal, A., Bengio, Y.: Inductive biases for deep learning of higher-level cognition. *CoRR* **abs/2011.15091** (2020)
23. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)
24. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)
25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
26. Huang, L., Zhao, X., Huang, K.: 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)
27. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al.: Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045 (2023)
28. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
29. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. JMLR Workshop and Conference Proceedings, vol. 37, pp. 448–456. JMLR.org (2015)
30. Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L.: Identifying natural images from human brain activity. *Nature* **452**(7185), 352–355 (2008)

31. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)
32. Koh, J.Y., Fried, D., Salakhutdinov, R.: Generating images with multimodal language models. arXiv preprint arXiv:2305.17216 (2023)
33. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* **128**(7), 1956–1981 (2020)
34. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., et al.: Obelics: An open web-scale filtered dataset of interleaved image-text documents. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
35. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: SiamRPN++: Evolution of siamese visual tracking with very deep networks. In: CVPR. pp. 4282–4291 (2019)
36. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
37. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
38. Li, Y., Chen, L., He, R., Wang, Z., Wu, G., Wang, L.: Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13536–13545 (2021)
39. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
40. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
41. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
42. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
43. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
44. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
45. Malla, S., Dariush, B., Choi, C.: Titan: Future forecast using action priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11186–11196 (2020)
46. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
47. OpenAI: Chatgpt. <https://openai.com/blog/chatgpt/> (2023)
48. OpenAI: Gpt-4 technical report (2023)
49. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
50. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Zhang, L.K.T.: Detgpt: Detect what you need via reasoning. arXiv preprint arXiv:2305.14167 (2023)
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

- natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
52. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
 53. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
 54. Ramnani, N., Miall, R.C.: A system in the human brain for predicting the actions of others. Nature neuroscience **7**(1), 85–90 (2004)
 55. Ryoo, M.S., Piergiovanni, A.J., Arnab, A., Dehghani, M., Angelova, A.: Tokenlearner: What can 8 learned tokens do for images and videos? CoRR **abs/2106.11297** (2021)
 56. Sap, M., Bras, R.L., Fried, D., Choi, Y.: Neural theory-of-mind? on the limits of social intelligence in large lms. In: EMNLP. pp. 3762–3780. Association for Computational Linguistics (2022)
 57. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
 58. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)
 59. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)
 60. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
 61. Simsek, F.E., Cigla, C., Kayabol, K.: Sompt22: A surveillance oriented multi-pedestrian tracking dataset. In: European Conference on Computer Vision. pp. 659–675. Springer (2022)
 62. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
 63. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20993–21002 (2022)
 64. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)
 65. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
 66. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
 67. Ullman, T.D.: Large language models fail on trivial alterations to theory-of-mind tasks. CoRR **abs/2302.08399** (2023)

68. Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yu, E., Sun, J., Han, C., Zhang, X.: Small language model meets with reinforced vision vocabulary. arXiv preprint arXiv:2401.12503 (2024)
69. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. *Trans. Mach. Learn. Res.* **2022** (2022)
70. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **35**, 24824–24837 (2022)
71. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021)
72. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring multi-object tracking. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14633–14642 (2023)
73. Xu, Y., et al.: Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In: *AAAI*. pp. 140–148 (2020)
74. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
75. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
76. Yu, E., Li, Z., Han, S.: Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8834–8843 (2022)
77. Yu, E., Liu, S., Li, Z., Yang, J., Li, Z., Han, S., Tao, W.: Generalizing multiple object tracking to unseen domains by introducing natural language representation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 37, pp. 3304–3312 (2023)
78. Yu, E., Wang, T., Li, Z., Zhang, Y., Zhang, X., Tao, W.: Motrv3: Release-fetch supervision for end-to-end multi-object tracking. arXiv preprint arXiv:2305.14298 (2023)
79. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
80. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6720–6731 (2019)
81. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022)
82. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023)
83. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)

- 84. Zhao, L., Yu, E., Ge, Z., Yang, J., Wei, H., Zhou, H., Sun, J., Peng, Y., Dong, R., Han, C., et al.: Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. arXiv preprint arXiv:2307.09474 (2023)
- 85. Zhao, L., Yu, E., Ge, Z., Yang, J., Wei, H., Zhou, H., Sun, J., Peng, Y., Dong, R., Han, C., et al.: Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. arXiv preprint arXiv:2307.09474 (2023)
- 86. Zhou, P., Madaan, A., Potharaju, S.P., Gupta, A., McKee, K.R., Holtzman, A., Pujara, J., Ren, X., Mishra, S., Nematzadeh, A., Upadhyay, S., Faruqui, M.: How far are large language models from agents with theory-of-mind? CoRR **abs/2310.03051** (2023)
- 87. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)