Multi-Modality Models for Embodied Visual Question Answering

Haosheng Tan

January 2025

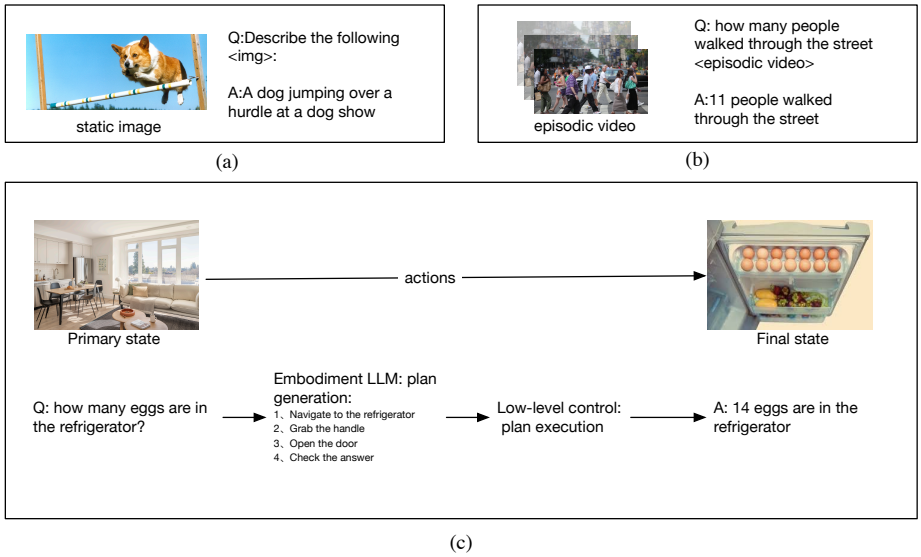

Figure 1: Examples of VQA and E-VQA

3 examples of VQA(a), EM-EQA(b) and A-EQA(c) correspondingly. Subplot (a) is an example of a static VQA that takes in a static image and (b) is an example of EM-EQA where an episodic memory is given. (c) is a instance of A-EQA where action plan is generated to explore the environment and gater enough information for an embodiment to get the solution

1 Introduction

Multi-modality models are blossoming after CLIP[1] and its following 'dual tower' models bring vision and language together, and Vision Transformer(ViT)[2, 3] also offers a unified framework

in vision aligning with language such that Large Language Models(LLMs) bring chain-of-thought and zero-shot learning into multi-modality tasks. Current multi-modality task involves Captioning, Vision Query Answering (VQA), and Multi-modality Sentiment Analysis, all of which require models to understand vision semantics strongly. While lots of multi-modality models demonstrates strong performance on VQA, Embodied Visual Question Answering (E-VQA), which is an application that requires understanding of environment well enough given a video instead of a static image to answer questions in natural language, is an emerging task especially in robotics and embodied intelligence. E-VQA consists of 2 types: episodic memory E-VQA (EM-EQA), and active exploration (A-EQA)[4]. Examples are shown in Figure 1 EM-EQA is mostly applied on mobile intelligent devices, smart glasses for instance, that leverage episodic memory and answer questions, which is similar to video-based VQA. A-EQA, on the other hand, is related to mobile robots that could explore and interact with the environment, gathering enough information to answer the questions. A-EQA could be regarded as a special version of EM-EQA as the episodic memory vision is generated during the exploration of the agent and could be more challenging since it requires not only strong memory and perceptual ability but also the precision of actions plan generated by the multi-modality agent.

2 Challenges of E-VQA

2.1 Poor Spatial Understanding and Semantic Grounding

Several challenges are posed in the modern embodied VQA task. Firstly, current multi-modality models perform poorly on EM-VQA requiring spatial understanding and semantic grounding for video. The experiment carried out by Majumdar et al.[5] also demonstrate that there is no significant difference between a multi-modality agent and a 'blind' LLM, which takes no visual information to make an answer, on spatial relationship understanding and object functional reasoning. The significant reason may be the lack of environment interaction during reasoning, or failure to capture spatial information by the vision encoder[1, 6], such as the relationship of objects. Improper Modality fusion mechanisms and vision encoders may also jeopardize the grounding information injection. Another potential reason is the frozen parameters during training since most of the multi-modality models are trained by freezing pre-trained modules, language models for instance, due to the training efficiency. This also indicates a trade-off between model performance and computational cost.

2.2 Transferring Knowledge from Vision Language Model

Secondly, transferring knowledge from VQA to E-VQA is challenging. Apparently, various Vision Language Models (VLM)[7, 8, 9, 10, 11, 12] have achieved astonishing performance on VQA, and adequate benchmarks, as well as public training data, are available in VQA. However, multi-modality models that perform well in VQA are challenging to transfer directly to E-VQA, especially for A-EQA, due to the requirement of exploring and plan generation. In addition, A-EQA requires not only the fusion of vision and language but also the alignment of high-level plan and low-level control[13]. A few works [13, 14, 15] address the plan generation by using chain-of-thought reason-

ing of LLMs that comprehensive 'common' knowledge well, but the modality imbalance also raises an issue of grounding, leading to the unalignment of visual cues and linguistic description. Besides, insufficient public data for E-VQA also lays a barrier in training and evaluation. An effective way of transferring knowledge from VQA to E-VQA is to be developed, and methodologies to mitigate the unbalance of modality are not trivial.

2.3 Catastrophic Forgetting and Undynamic Adaption

Moreover, Catastrophic forgetting and undynamic adaption to the environment is a critical problem in application. Modern Embodied Multi-modality models still suffer short memory, showing a lack of long-horizon plans when persistent memory is required, and this is also a shortcoming of LLM. Embodied agents fail to make proper decisions and flexible adjustments under unexpected environment changes and interruptions, which poses a significant weakness in real-world adoption. In addition, the benefit of scaling law results in inefficient and even impractical solutions in embodied intelligence. While embodied multi-modality models are supposed to work on mobile devices, models are required to be 'lightweight', which brings a constraint on model size. This leads to a trade-off between model's intelligence required by complicated tasks and the feasibility of mobile deployment. Knowledge distillation might be an indispensable technique in application.

3 Related Work

3.1 Embodied Multi-modal Language Model(VLA models)

Large language models encode great semantic knowledge of the world, and chain-of-thought and zero-shot learning capability allow LLMs to achieve 'general' intelligence. Nevertheless, a significant shortcoming is that it lacks real-world interaction, which makes it generate unreasonable decisions given embodiment. SayCan[16] addresses this problem by leveraging temporal-difference-based (TD) reinforcement learning to provide an affordance function to make decisions. However, SayCan only provides linguistic input and ignores the vision encoding as well as other geometric configurations of the scene, leading to a weak grounding of the LLMs. After taking the first step of linking LLMs to robots, Driess et al[17] propose a way of introducing visual modality information. It introduces neural scene representations and entity-labelling multi-modal tokens to inject continuous embodied observations and other modalities into a pre-trained language model, which is similar to the idea of Kosmos[9, 10] using multi-modal sentences. This method seems to push a 'big convergence' in embodied AI by utilizing a unified framework to inject all of the continuous observation of the scene, but it still struggles with complex spatial reasoning and fine-grained object interactions because the ViT and the 3D-aware Object Scene Representation Transformers (OSRTs) it uses for vision encoder is challenging to understand dynamic 3D vision. Although it can generate step-by-step plans for the policy model, it struggles with long-horizon reasoning due to the lack of persistent memory. Following the multi-model sentence, RT-2[14] manages to encode robot actions into tokens injected into multi-modal sentences, bringing an end-to-end framework from high-level tasks to low-level robot actions. Introducing Low-Rank Adaptation (LoRA)[18] and

quantization techniques, OpenVLA[15] provides a highly efficient fine-tuneable model in only 7B size but better spatial and semantic understanding. It greatly improves modern Embodied models in accessibility, efficiency, and adaptability. EmbodiedGPT[13] enhances PaLM-E by proposing a closed-loop embodied control paradigm where Embodied-Former, a cross-attention mechanism, is introduced instead of using multi-modal sentences and integrates low-level control execution, forming a closed-loop system for real-time adjustments. Embodied-Former improves task-relevant feature extraction and better aligns visual features with language-based planning. Extraction of task-relevant features from planning queries also improves the real-time adaption. However, the problem of catastrophic forgetting still remains and it is not fully dynamic to unexpected changes in the environment.

In summary, previous works manage to link LLMs to embodied agents and borrow the strength of chain-of-thought reasoning and modality fusion mechanisms to allow LLMs to better perceive the real world and generate task-relevant plans. Also, LLM provides a great source of knowledge learned from the Internet, which could improve complicated task understanding. Catastrophic forgetting in long-horizon plans is still a challenge, and real-time adaptive agents are yet to be achieved. Inferior spatial understanding and poor generalization to unseen environments need to be improved. The gap between state-of-the-art embodied models and human performance remains large, which is a sign of a huge challenge on embodied intelligence to achieve human-level performance but also reveals more improvement to be achieved.

3.2 Vision Language Pre-trained models

Modern Embodied multi-modality model relies heavily on Multi-modality LLM (MM-LLM) backbone that provides high-level task-relevant plan and modality perception. Leveraging MM-LLM as a core backbone also reduces the cost of training as the robot manipulation dataset is limited[15, 5]. Transformer and all the multi-modaltiy models based on it have been bringing vision and language all together and multi-modality models, which compress multiple modalities such as language and vision to enhance the overall performance, are undergoing a 'big convergence'. In the field of static VQA, model architecture has been switching from 'dual tower' models[1, 19, 7], which train uni-modality models separately and combine them through multi-modality fusion layers, to training multi-modality simultaneously by introducing a more unified architecture[8, 11, 20, 21]. Moreover, some models[11, 22] demonstrate a superior performance by using a single training loss instead of multiple losses to train the alignment between vision and linguistic description, which brings the unified framework of the multi-modality models and pushes the 'big convergence'. MetaLM[22] propose the way of multi-modal sentence that simply treat other modality as 'foreign language', which improve the efficiency of training and provide flexibility of input ,and Kosmos[10, 9] leverage language as a task interface, showing great adaption to various sub-stream tasks. Multi-modal sentence is also borrowed by some following embodied models[17, 14, 15] for injecting continuous observations to the language input.

The unified framework borrows the strength from scaling and shows a great performance in vision, language, and multimodality tasks, which makes it become an potential component of embodied AI and other modern AI system. Those success indicate a great power of scaling that even with a simple architecture like CLIP[1] could show great performance as massive data and model size

is provided. The ability of fusing the modality is favorable in various machine-intelligent fields requiring multiple sources of input, such as robotics and auto-driving. However, transferring those success from static VQA to E-VQA is challenging due to the grounding issue but it also indicates that a potential identical scaling law could be acquire in the field of embodied intelligence.

4 Methodology

This section briefly summarizes the methodologies that might be investigated in this research project. The proposed project could focus on one of the technologies to address current problems or investigate combination for novel solutions.

4.1 Spatial Feature Encoder

Spatial feature understanding is a critical ability yet to be improved in embodied multi-modality models and the fundamental key is the vision encoder. Following methodologies focus on spatial feature extraction and reasonably be injected in modern embodied multi-modality models.

4.1.1 DINO-v2 for spatial extraction

Self-distillation with no labels, also known as DINO[23], is a self-supervised learning method in computer vision. Unlike conventional ViT-based vision encoders that are trained on Image-Net with labels, DINO adopt self-supervised learning combined with knowledge distillation allowing vision encoder to capture more robust features. It shows great performance in image segmentation and object detection and achieve state-of-the-art performance on video object segmentation task DAVIS 2017, which requires strong spatial understanding. What can be learned from DINO is that self-supervised learning on images allows model to capture fine-grained features while labeled data, like using a caption, provide general view and lose the information of the details. Also, self-supervised learning doesn't strongly rely on language as most of the vision encoders are trained with language guidance that encourage models to bridge visual observation with linguistic description but rely heavily on language. DINO also adopts multi-views of an image during training where the student model learn from locality while the teacher observe the global given same model architecture but different parameters. A cross-entropy loss 1 of student and teacher output distribution force the models to learn connection between image locality and global view, which could stimulate better understanding of object relationship within the image.

$$\min_{\theta_s} \sum_{x \in \left\{x_1^g, x_2^g\right\}} \sum_{\substack{x' \in V \\ x' \neq x}} \mathcal{L}_{cr}(P_t(x), P_s(x')) \tag{1}$$

where V is a subset of multiple views generated by the original images and x_1^g and x_2^g denote two global views. $\mathcal{L}_{cr}(P_t(x), P_s(x'))$ is a cross-entropy loss of the output distribution of teacher $P_t(x)$ and student $P_s(x')$. For the ViT structure, DINO-v2[24] propose using separate head: DINO head and iBOT[25] head, which extract image-level feature and patch-level feature respectively. This method enhance the robustness by introducing multi-scale representations and the extracted feature is general for various downstream tasks, which is suitable for embodied models.

Based on the robust feature extraction of DINO-v2, the further step could be training vision encoder using DINO-v2 method on embodied video data. Most of vision encoders that DINO is applied to are trained on static image instead of embodied video, which may fail to capture motions dynamics. Therefore, training an encoder with DINO-v2 on video data or introducing masked video modeling task help improve its ability to track objects and infer motion. Moreover, vision encoder trained with DINO-v2 fails to align with language and other modalities since it is self-supervised trained solely on image data. Combining with another vision encoders, such as CLIP[1] and SigLIP[6], from a pre-trained VLM may supplement the alignment of other modalities.

4.1.2 Stable Diffusion for Future Video Anticipation

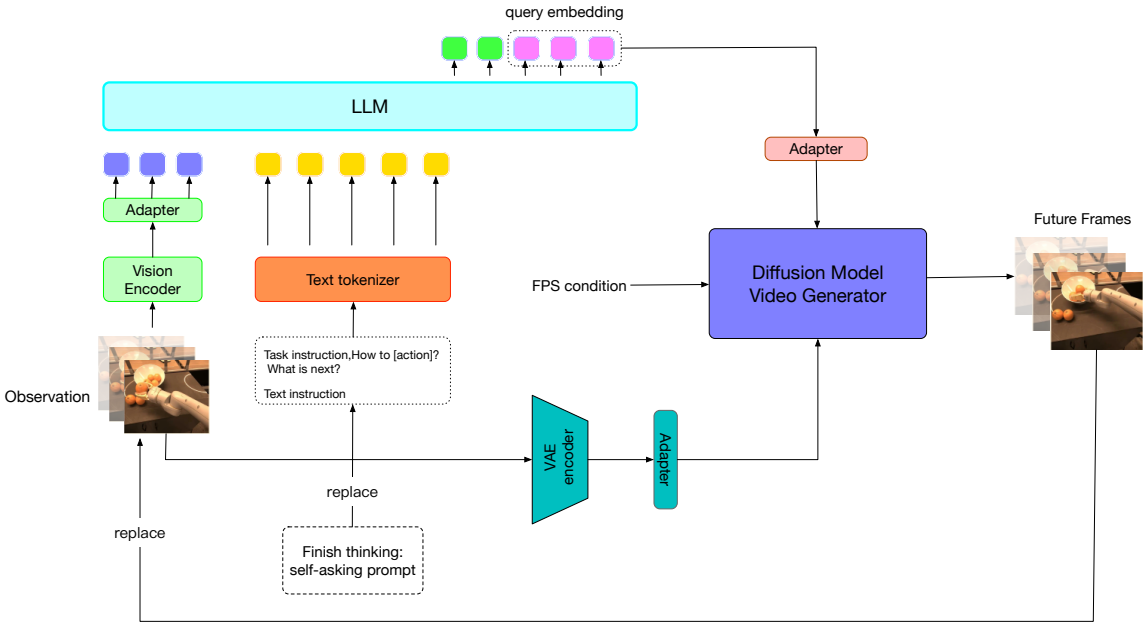

Figure 2: Framework of embodied-video prediction

Embodied Video Anticipation (EVA) is an emerging technique to improve action plan generation's grounding and success rate. It simulates the imagination process generated by human brain before taking actions, and thus, it is relevant to understand the interaction of the physical world. This task allows embodied models to predict future events, improving their ability to answer questions about actions, sequences, and tasks in a dynamic environment and it also bridges the gap between

video generation and video understanding in embodiment. Chi et al.[26] propose an embodied video anticipator by combining a VLM and a diffusion-based video generator for future video generations. A new framework (shown in 2) is also introduced consisting of 4 meta tasks: Action-Description, Finish-Thinking, How-To, and Next-Step. Action Description is to align task requirement with observation through a VLM while Finish-Thinking, How-To, and Next-Step produce a for loop where the last frame of the predicted video would be self-ask if the task reaches the complete status or not. If it doesn't satisfy the task completion requirement, then the generator is prompted to generate the extension of the video based on the previous prediction until the task is finished. In the core of this framework, a video generator that could be conditioned on the text query and the previous prediction to ensure reasonability and consistency is crucial. Therefore, stable diffusion models[27, 28, 29, 30] are favoured as they are based on denoising a conditional random distribution and could be guided by language in image generation. In specific, diffusion process starts from the video latent z_0 encoded by a Variational Auto Encoder(VAE) and adding a random noise in certain distribution on it iteratively:

$$q(z_{1:T}|z_0) := \prod_{t=1}^{T} q(z_t|z_{t-1}) \tag{2}$$

where T is the total number of steps. The latent video z_0 typically has lower dimension than the original video x in the need of reducing computational cost. The noise could be shown in:

$$q(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}) \tag{3}$$

where β_t controls the noise level at step t. The generation process, also known as the denoising process, reverses the above procedure by predicting the noise added at step $t-1$ given z_t and 3D UNet used for video reconstruction is suitable for this task. Specifically, the objective function can be the mean squared error between the true noise and predicted noise, shown as:

$$\mathcal{L} = \mathbb{E}_{\varepsilon(x), \epsilon \in \mathcal{N}(0,1), c, t} \left[\| \epsilon - \epsilon_\theta \left(z_t, t, c \right) \|_2^2 \right] \tag{4}$$

where $\epsilon_\theta \left(z_t, t, c \right)$ denotes the UNet given latent z_t at step t and the condition c with parameters θ. Latent video diffusion model(LVDM) could incorporate spatial and temporal transformers to ensure consistency between frames and the conventional injection is the cross-attention[29, 27, 31] of which the output is concatenated to the feature representations after Convolution layers and pooling layers. Another strength of the diffusion models is the diversity and adaptability compared to GAN[32], and it provides great generalization in various embodied scenes. The limitation of latent video diffusion is the inferior reasoning ability and weak motion understanding, which are common problems for modern video generation models. Moreover, it struggles with generating long-horizon video generation (Dynamicrafter[29] could only generate 16 frames for 2-second video), and it is computationally demanding in inference as it requires an iterative process to denoise images. Injecting embodied information into VLM for training diffusion generator and accelerating generation are the promising future directions for EVA.

On this basis, several improvements of stable diffusion models for EVA are listed as follow. Firstly, modality fusion mechanism that could efficiently incorporate sensor observation should be introduced to enhance condition of grounding observation. Object-centric representation should also be considered to improve state consistency throughout the video sequence. Secondly, enhancement of temporal modeling for latent diffusion model is also a direction to extend generated video length. Thirdly, domain adaptation techniques like Ensemble-LoRA[26] to fine-tune models for various environments without extensive retraining is a promising future step as embodied VQA requires great generalization. Moreover, optimizing the sampling strategies during inference and distillation-based method, such as consistency model[33], is considered an essential improvement to accelerate the denoising process, reducing the computational overhead of diffusion models.

4.2 Knowledge Distillation

In the application of embodied AI, models are required to be 'lightweight' and could be run efficiently on mobile devices. As the scaling law persists, current multi-modality models tend to increase the model size to gain more intelligence, posing a challenge in running large models on mobile devices. Therefore, knowledge distillation[34] that keeps a relatively same ability while largely reducing the size of models is an indispensable technique in embodied AI. Knowledge distillation introduces a 'student' model $s_\theta(x)$, which is trained to achieve a competitive behavior of a larger, more complex 'teacher' model $t = t_\theta(x)$. If the last softmax layer returns an output of probabilities among all classes, the probability that the model fits to can be denotes as :

$$p_i = \frac{exp(s_i(x)/\tau)}{\sum_{j=1}^{n} exp(s_j(x)/\tau)} \quad q_i = \frac{exp(t_i(x)/\tau)}{\sum_{j=1}^{n} exp(t_j(x)/\tau)} \tag{5}$$

where p_i denotes the probability distribution of student output and q_i for teacher's. τ is a softmax temperature parameter to control sharpness of the distribution. The objective function could be the KL divergence:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{ce}(s(x), y) + \alpha D_{KL}(q, p) \tag{6}$$

$$D_{KL}(q, p) = \sum_{i=1}^{n} q_i log(\frac{q_i}{p_i}) \tag{7}$$

$\mathcal{L}_{ce}(s(x), y)$ is the cross-entropy loss between ground truth and logits from student. Knowledge Distillation has been widely used on SOTA models, showing great power in a specific task but fewer parameters. Through distillation, knowledge is transferred from the teacher to the student, which is suitable for embodied application as E-VQA is highly task relevant. Therefore, applying distillation on modern pre-trained general embodied models as initialization, or adopting knowledge distillation before deploying trained model on the robot is an significant step to acquire a more efficient model suitable for mobile devices

4.3 Dynamic Long-horizon planning system

To make up for the lack of long-horizon planning and real-time adaption, a Dynamic long-horizon planning system is required. Instruction Augmented Long-Horizon Planning (IALP)[35] system propose a framework that leverage Planning Domain Description Language (PDDL) that manage to incorporate problem's states, actions, preconditions, and effects. This system works in a closed loop that after each action, it updates the PDDL problem with new sensor observations. This allows the LLM planner to adjust the plan based on the latest environment state and therefore, each action's feasibility is checked using the grounding predicates, ensuring that the robot can actually perform the action before committing to it. This iterative process helps the robot adapt to changes or failures, leading to more reliable task completion and feasibility estimation also help generate a long-horizon plan. This framework emphasis grounding mechanism in plan generation and the future system could release the limit of manual actions library, unfrozen visual encoder or even LLM backbone to increase generalization as well as efficiency.

Based on the methodology used for IALP, a dynamic framework combining Embodied Video Anticipator and PDDL is the next step. EVA provides an effective framework for Video anticipation before taking action, increasing the success rate while incorporating PDDL in the self-asking mechanism to improve real-time adaption. Combining these two frameworks equips embodied intelligence to make future predictions as well as deal with unprecedented situations in a timely, which shows a potential to greatly improve embodied VQA in real scenarios.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *International conference on machine learning*, pp. 8748–8763, 2 2021. 1, 2.1, 3.2, 4.1.1

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 10 2020. 1

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 3 2021. 1

[4] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A Survey of Embodied AI: From Simulators to Research Tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, pp. 230–244, 3 2021. 1

[5] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud, K. Yadav, Q. Li, B. Newman, M. Sharma, V. Berges, S. Zhang, P. Agrawal, Y. Bisk, D. Batra, M. Kalakrishnan, F. Meier, C. Paxton, A. Sax, and A. Rajeswaran, "OpenEQA: Embodied Question Answering in the Era of Foundation Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 2024. 2.1, 3.2

[6] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, and G. Deepmind, "Sigmoid Loss for Language Image Pre-Training," in *CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023. 2.1, 4.1.1

[7] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," *International conference on machine learning*, pp. 19730–19742, 1 2023. 2.2, 3.2

[8] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, and F. Wei, "VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32897–32912, 11 2021. 2.2, 3.2

[9] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding Multimodal Large Language Models to the World," *arXiv preprint arXiv:2306.14824*, 6 2023. 2.2, 3.1, 3.2

[10] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. Khan Mohammed, B. Patra, Q. Liu, K. Aggarwal Zewen Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei, "Language Is Not All You Need: Aligning Perception with Language Models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72096–72109, 2023. 2.2, 3.1, 3.2

[11] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks," *arXiv preprint arXiv:2208.10442*, 8 2022. 2.2, 3.2

[12] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Advances in neural information processing systems*, vol. 36, 2024. 2.2

[13] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "EmbodiedGPT: Vision-Language Pre-Training via Embodied Chain of Thought," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 2.2, 3.1

[14] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *arXiv preprint arXiv:2307.15818*, 7 2023. 2.2, 3.1, 3.2

[15] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An Open-Source Vision-Language-Action Model," *arXiv preprint arXiv:2406.09246*, 6 2024. 2.2, 3.1, 3.2

[16] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," *arXiv preprint arXiv:2204.01691*, 4 2022. 3.1

[17] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "PaLM-E: An Embodied Multimodal Language Model," *arXiv preprint arXiv:2303.03378*, 3 2023. 3.1, 3.2

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 6 2021. 3.1

[19] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 7 2021. 3.2

[20] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut, "PaLI: A Jointly-Scaled Multilingual Language-Image Model," *arXiv preprint arXiv:2209.06794*, 9 2022. 3.2

[21] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, S. Shakeri, M. Dehghani, D. Salz, M. Lucic, M. Tschannen, A. Nagrani, H. Hu, M. Joshi, B. Pang, C. Montgomery, P. Pietrzyk, M. Ritter, A. Piergiovanni, M. Minderer, F. Pavetic, A. Waters, G. Li, I. Alabdulmohsin, L. Beyer, J. Amelot, K. Lee, A. P. Steiner, Y. Li, D. Keysers, A. Arnab, Y. Xu, K. Rong, A. Kolesnikov, M. Seyedhosseini, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut, "PaLI-X: On Scaling up a Multilingual Vision and Language Model," *arXiv preprint arXiv:2305.18565*, 5 2023. 3.2

[22] Y. Hao, H. Song, L. Dong, S. Huang, Z. Chi, W. Wang, S. Ma, and F. Wei, "Language Models are General-Purpose Interfaces," *arXiv preprint arXiv:2206.06336*, 6 2022. 3.2

[23] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 4 2021. 4.1.1

[24] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," *arXiv preprint arXiv:2304.07193*, 4 2023. 4.1.1

[25] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "iBOT: Image BERT Pre-Training with Online Tokenizer," *arXiv preprint arXiv:2111.07832*, 11 2021. 4.1.1

[26] X. Chi, H. Zhang, C.-K. Fan, X. Qi, R. Zhang, A. Chen, C.-m. Chan, W. Xue, W. Luo, S. Zhang, and Y. Guo, "EVA: An Embodied World Model for Future Video Anticipation," *arXiv preprint arXiv:2410.15461*, 10 2024. 4.1.2, 4.1.2

[27] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C. Weng, and Y. Shan, "VideoCrafter1: Open Diffusion Models for High-Quality Video Generation," *arXiv preprint arXiv:2310.19512*, 10 2023. 4.1.2, 4.1.2

[28] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, G. Schindler, R. Hornung, V. Birodkar, J. Yan, M.-C. Chiu, K. Somandepalli, H. Akbari, Y. Alon, Y. Cheng, J. Dillon, A. Gupta, M. Hahn, A. Hauth, D. Hendon, A. Martinez, D. Minnen, M. Sirotenko, K. Sohn, X. Yang, H. Adam, M.-H. Yang, I. Essa, H. Wang, D. A. Ross, B. Seybold, and L. Jiang, "VideoPoet: A Large Language Model for Zero-Shot Video Generation," *arXiv preprint arXiv:2312*, 12 2023. 4.1.2

[29] J. Xing, M. Xia, Y. Zhang, H. Chen, W. Yu, H. Liu, X. Wang, T.-T. Wong, and Y. Shan, "DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors," in *European Conference on Computer Vision*, 10 2023. 4.1.2, 4.1.2

[30] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," *arXiv preprint arXiv:2311.15127*, 11 2023. 4.1.2

[31] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "VideoComposer: Compositional Video Synthesis with Motion Controllability," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7594–7611, 2023. 4.1.2

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139–144, 10 2020. 4.1.2

[33] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency Models," *arXiv preprint arXiv:2303.01469*, 3 2023. 4.1.2

[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 3 2015. 4.2

[35] F. Wang and D. Navarro-Alarcon, "Instruction-Augmented Long-Horizon Planning: Embedding Grounding Mechanisms in Embodied Mobile Manipulation," tech. rep., 2025. 4.3