Bridging Feature Spaces: Survey on Multimodal Large Language Model

WashU CSE527A Survey Paper

Jianhong Tu
Washington University in St. Louis
jianhong.t@wustl.edu

Hongzhang Wang
Washington University in St. Louis
hongzhang@wustl.edu

Alexander Wollam
Washington University in St. Louis
atwollam@wustl.edu

Abstract

1 Introduction

In recent years, large language models (LLM) have demonstrated robust accuracy in high-level tasks, such as multi-step reasoning (Rae et al., 2021) and instruction following (Ouyang et al., 2022). However, the input to a LLM has been limited to the linguistic space. To expand the input space, researchers proposed multimodal large language models (MLLM) that leverage the idea of *natural language supervision learning* (Radford et al., 2021) to label features from other modalities with rich and expressive natural language, taking advantage of the *emergent capabilities* (Wei et al., 2022) of LLMs. This survey provides an overview of recent advancements in MLLM and summarises the ideas behind integrating multiple modalities by extending the representation learning method in deep learning.

1.1 Multimodality

"Multimodal" in the context of machine learning typically refers to models or systems that can process and understand information from multiple types of input (Kress, 2009). Although a human's perception of the world is inherently "multimodal," the deep learning models developed in specific domains, including reinforcement learning, computer vision, and natural language processing, have synchronously evolved in parallel (Yin et al., 2023). Though LLMs have exploited rich text at a large scale, further expanding the breadth of the domain is beneficial: the injection of signals from other channels is expected to help LLMs develop a more holistic understanding of real-world entities.

1.2 Unimodel Deficiencies

Pretraining via self-supervision on some proxy objectives to acquire general features about the data is a popular DL technique. However, the learned representations are highly abstract, so researchers often rely on ablation studies to understand the contribution of each component of the data (Zhao et al., 2020). The richness and expressivity of natural language have the potential to instantiate abstract signals from images and audio, allowing an intuitive interpretation through words. In *grounded* tasks that require reasoning with reference to other modalities or involve interaction with the peripherals, especially in robotics, LLMs alone are incapable of taking geo-localization and state estimates in robotic policy into account (Driess et al., 2023). Thus, a major goal of MLLM is to *fuse* multiple modalities into an overarching model and complement LLMs with extra information. From a machine-human interaction perspective, an MLLM is preferable due to its flexibility in accepting and generating information from diverse modalities. Such flexibility renders the model suitable for a wider range of downstream tasks (OpenAI, 2023; Betker et al., 2023).

1.3 Challenges

A few unique challenges naturally arise in building an MLLM system from its goals. First, the abilities of LLM are only possible with abundant textual datasets, but the availability of large datasets for every modality remains a question. Yet, the establishment of correspondence between modalities is an ongoing quest, and a customized dataset is often needed, which is hard due to the ambiguity in labeling. Pairing target and ground truth is indefinite since images or other modalities are often open to interpretation. Equally subjective, choosing an appropriate representation, or an encoder, is also a "dark art" dependent on experience and experiments. Furthermore, the sub-elements, such as a sequence of video frames and video captions, require precise temporal and spatial alignment. From a modeling perspective, the challenge is to devise

an interface that translates multimodal inputs into a *joint representation* without compromising useful signals. Finally, through co-learning, MLLMs aim to acquire an abstract understanding of entities that generalize to out-of-distribution data. A promising *zero-shot* and *few-shot* accuracy better than the unimodel counterparts implies a good *transfer effect* that information from other modalities helps the MLLM acquire a better perception of an entity in a unimodal scenario.

2 Methods

To limit the scope, this section focuses on **Vision-Language Models** (VLM) that accepts textual tokens and images and details the methods to construct a joint feature space for two modalities via step-by-step feature transforming.

2.1 Datasets

Obtaining a large dataset is crucial to the generalization ability of VLM because the quality of unimodal representation learning is positively proportional to the scale of a dataset. Particularly, in learning image representation, the vision transformer, a popular choice in VLM architecture, yields a modest accuracy when trained on a mid-sized dataset but performs better than the SOTA ResNet in a set of common image recognition benchmarks when trained on larger datasets (Dosovitskiy et al., 2021). One method to gather sufficient data points is to integrate benchmark datasets across domains. The PaLM-E group, concerning robotic policy in addition to visuals and texts, gathered the image captioning data "COCO," visual question answer data "VQAv2", text corpus "Wikipedia," and robotic navigation data "SayCan" Driess et al. (2023). Appendix A summarizes the common datasets. However, crowd-labeled datasets are often limited in size or tailored for tasks other than VLM. There is a trend to exploit the richness of information on the internet by compiling a customized dataset (Radford et al., 2021). Sun et al. (2019) utilized YouTube's audio-transcription to gather 176 hours of synchronous video frames and captions through web API, while Radford et al. (2021) collected approximately 400 million pairs of image and descriptions from public sources.

2.2 Joint Representation

LLMs trained on textual data are incompatible with the raw image data as a 3D tensor with height, width, and multiple channels. A common multimodal modeling approach is learning a joint representation of texts, images, and other modalities. Formally, let $X,\ Y$ be the feature spaces, and we intend to find $\phi : X \times Y \to Z$, where Z is the feature space for the VLM. Due to the cost of training end-to-end multimodels (Yin et al., 2023), it is more feasible to concatenate conventional unimodels that are pretrained separately into a multimodel. With the LLM as the backbone, the function ϕ serves as a learnable interface that projects the image space X onto the embedding space $Z = Y$ (Yin et al., 2023), keeping the feature space invariant to pretrained LLMs.

Pretraining refers to the stage where a task-agnostic model is trained to optimize a proxy objective in order to capture the general features of data. Akin to pretrained embeddings which can capture semantic similarities, pretrained image classifiers can encode abstract signals. Consider a trained deep classifier as a composite function $h \odot \theta$ where h is a SoftMax classifier and θ is a feature transformation. The function h may be understood as a mapping from the raw feature space to the transformed feature space where the data is linearly separable. Sun et al. (2019) obtains such feature mapping by pretraining a ConvNet on the Kinetics dataset, then detaching the final SoftMax layer, resulting in 1024-dimensional continuous feature vectors. Continuous features are tokenized using hierarchical K-Means to extract 2^{14} centroids as the final image representations. An alternative model is to apply a vision transformer, which is an extension of the transformer architecture and is able to conveniently map images as a sequence of patches into patch embeddings (Dosovitskiy et al., 2021; Driess et al., 2023; Radford et al., 2021). However, patch embeddings are not equivalent to the word embeddings.

2.3 Multimodal Pretraining

Unimodal pretraining learns a useful image prior, and multimodal pretraining further establishes the connection between image and word embeddings. Most simply, the connection can be learned by treating video embeddings as word embeddings. Sun et al. (2019) appends the new image embeddings to a pretrained BERT's lookup table and optimizes a weighted sum of unimodel and multimodel objectives detailed in Appendix B. More directly, *contrastive objectives* are used to fine-tune the text and image encoders by explicitly training on an

Model	VQAv2 test-dev	test-std	OK-VQA val	COCO Karpathy test
PaLI-3B	81.4	-	52.4	145.4
PaLI-15B	82.9	-	56.5	146.2
PaLI-17B	84.3	84.3	64.5	149.1
PaLM-E-12B	77.7	77.9	60.1	136.0
PaLM-E-66B	-	-	62.9	-
PaLM-E-84B	80.5	-	63.3	138.0

Table 1: Performance of PaLI (Chen et al., 2023) and PaLM-e (Driess et al., 2023) on General Visual-language Benchmarks

image-caption paring task to maximize the cosine-similarity between paired images and text embeddings. (Radford et al., 2021). Another light-weight training scheme, as opposed to the end-to-ending training above, is to learn a linking function that transforms images to "prompts" while "freezing" the encoders and the LLM (Driess et al., 2023). The multimodel is pretrained on next-sentence prediction tasks with sequences of visual and linguistic embeddings, updating the affine transformation function only.

2.4 Fine-tuning

Fine-tuning is a technique to adapt a pretrained model for specific downstream tasks by transferring acquired unimodal knowledge into the final model. The MLLM optimizes some appropriate objective functions used to assess the model's performance on a curated set of benchmarks, such as visual question answering (VQA) (Yin et al., 2023). Fine-tuning enables the application of the model in high-level tasks, which may be evaluated qualitatively. Figure 2 provides 2 examples: VideoBERT model (Sun et al., 2019) captions the given image by describing the scene and actions and predicting successive steps, while the PaLM-e model (Driess et al., 2023) generates a robotic policy based on the given prompt.

3 Results

Evaluation of multimodels is data and task specific. We generally discusses 3 perspectives to assess models' performance on downstream tasks and the quality of join representation learning.

3.1 Benchmark scores

The usage of common benchmarks enables cross-evaluation of different models. Table 1 provides the reference scores of 2 SOTA MLLM on multimodal tasks. VQA (Goyal et al., 2019) provides an image and prompt as the inputs and assesses

Model	ImageNet	ImageNet-v2	ImageNet-R	ImageNet-A
PaLI-3B	70.06	62.55	80.15	37.92
PaLI-15B	70.27	62.81	81.21	41.16
PaLI-17B	72.11	64.46	81.97	44.7
CLIP	76.2	70.1	88.9	77.1
ResNet	76.2	64.3	37.7	2.7

Table 2: Performance of various models on ImageNet benchmarks.

the model's ability to retrieve information from the images as the answer to the prompt, while COCO caption (Chen et al., 2015) lets the model generate a context-free description of the image. Both tasks directly assess the understanding of images. Text-only tasks like natural language generation are also used to assess whether fine-tuned MLLM retains language ability. Expectedly, frozen MLLM shows no compromise in language ability, though is not as performant in multimodal tasks. One takeaway is that model scale positively impacts both models, which is intuitive for LLMs since their amazing abilities are only possible with a large dataset and complex architecture with billions of parameters. In addition, *zero-shot classification* is of great interest to understand the generalization of abstract knowledge into unseen examples. Driess et al. (2023), Radford et al. (2021) and Sun et al. (2019) report a better zero-shot image classification accuracy than the baseline unimodel and more efficient learning, requiring fewer data to achieve the same level of performance.

3.2 Customized Task Performance

Models fine-tuned to more restricted or less common domains need to be assessed using customized benchmarks as their training does. VideoBERT (Sun et al., 2019) is evaluated in video captioning tasks but using cooking instructional videos only. A successful prediction is when the model predicts key words that is relevant to the ground truth caption. MLLM is also applied to the robotics field with the model as the "brain." PaLM-e (Driess et al., 2023) is able to interact with the environment, which is evaluated on the success rate to carry out human instruction, i.e. pick up the sponge and then bring it to the user. In a VQA fashion, researchers can assess the model's understanding of the affordance, whether an *action* can be performed on an *entity*. Expectedly, given the same experiment setting, multimodels are reported to outperform the unimodel baseline, meaning a successful co-learning. In general, evaluation on customized benchmarks is challenging and the exact metric is

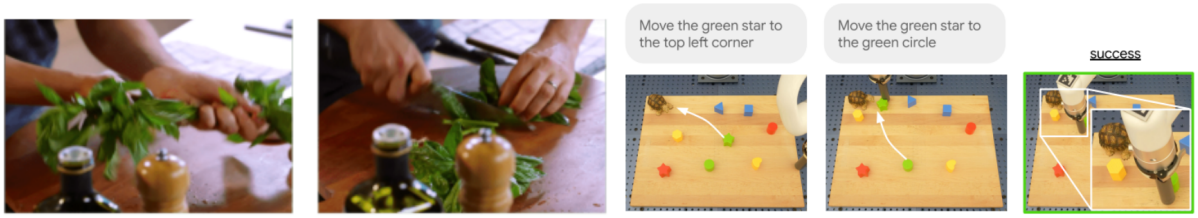

Figure 1: Examples of High Level Tasks

subjective to bias.

3.3 Representation Effectiveness

Recall that a representation mapping can be obtained by detaching the output layer. The quality of the learned representation can be thus accessed by the accuracy of a linear classifier on some classification tasks. Compared against ResNet (He et al., 2015) and BiT-M (Kolesnikov et al., 2020), the features outputted by CLIP (Sun et al., 2019) achieves a >70% accuracy on the 16-shot classification tasks versus 63% and 53% accuracy for BiT-M and ResNet respectively, meaning a more effective representation learning result.

4 Discussion

In the results above we demonstrate how performant MLLM models are and their ability to generalize with competitive zero- to few-shot performance on related tasks such as is seen in (Driess et al., 2023; Sun et al., 2019). Such performance alludes to these models' increasing ability to generalize well; this is an important feature of these models as such generalization entails better understanding of the different modalities, thus better performance and more capability overall. While these results are very promising and notably better than previous models, there is still much room for improvement in regards to topics such as modality reliability, LLM language capability retention, and efficiency.

Regarding modality reliability, looking particularly at vision, (Fu et al., 2023) shows that MLLMs can be inconsistent and misidentify features in images. (Yin et al., 2023) attributes this issue to information loss due to token size capacity and computation burden; they say that increasing image token representation size would improve performance at the cost of computation burden, and as such they advocate that more in-depth image feature compression may be a useful path to improve results without introducing such additional burden. Additionally, (Fu et al., 2023) reports that MLLMs have a tendency to hallucinate non-existing features as being present in input images. (Yin et al., 2023) reasons that this can be ascribed to insufficient modality alignment pretraining and may be resolved with more fine-grained alignment methods. (Ji et al., 2023) describes this behaviour as likely resulting from LLMs existing tendencies to hallucinate incorrect information—particularly relating to its uncertainty. For both perception and hallucination then, the primary issues together can be seen to stem from both model uncertainty regarding fine-grained features of input images, and LLMs general tendency to hallucinate and respond confidently even when uncertain.

Regarding LLM language capability retention (Fu et al., 2023) demonstrates the MLLMs can fail in instruction following and reasoning. For instruction following, they find that these models often don't precisely answer as they are prompted. For reasoning, they show how even when MLLMs seem to know key parts of an answer, they often fail to put them together and answer correctly. Both of these shortcomings are perhaps surprising since they are accomplished well by LLMs in general. (Driess et al., 2023) provides some insight into these issues with respect to their own model by directly comparing NLG task performance between their base PaLM LLM and subsequently derived PaLM-E MLLM across different model sizes. They show that for the smaller models, after being trained on the multimodal objectives, they face catastrophic forgetting and perform significantly worse at the same NLG tasks. This is only alleviated for their largest model, where performance only drops a few percent. From this comparison, we can see that the lackluster language capability retention relates to MLLMs forgetting much of their base LLMs' original training when optimized end-to-end. This hints at how important scaling is to MLLMs. Additionally, since the primary issue

pertains to forgetting, this suggests that simultaneous training on multimodal objectives and unimodal NLG objectives may also reduce the gap without requiring additional model size at the cost of training complexity.

For both modality reliability and language capability retention, we see that much of the issues in both can be alleviated through sheer model scale—allowing for more fine-grained representations of images and helping to remember base LLM training relations. This is standard of LLMs in general, as ever increasing scale is a prime contributor to their performance, however it appears to be an even greater factor for MLLMs likely due to their needing to capture more complex multimodal relationships. This unfortunately is at the cost of efficiency due to model size, but also with respect to training, as (Driess et al., 2023) also shows that to get the full performance out of their models, training is required to leverage a full-suite of pretraining and end-to-end optimization. As such, advances in efficiency is especially important in improving MLLMs.

5 Conclusion

In this survey, we provide an introduction to MLLMs. We first motivate these models by contrasting them to unimodal LLMs and describe some of the challenges faced when developing them. Then, we describe the current approaches in creating and training vision MLLMs with respect to some of these challenges and present illustrative results. Finally, we conclude by commenting on some of the weaknesses that these models currently have and highlight some explanations to motivate future work.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. 2023. Improving image generation with better captions.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 491–507. Springer.

G. Kress. 2009. *Multimodality: A Social Semiotic Approach to Contemporary Communication*, 1 edition. Routledge.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *CoRR*, abs/2306.13549.

Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. 2020. Exploring self-attention for image recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10073–10082. Computer Vision Foundation / IEEE.

A Example MLLM Datasets

Dataset	Task	Size	
TAMP	Robotic manipulation planning with VQA	96000 scenes	Source
Language Table	Robotic manipulation planning	600000 sequences	Driess et al. (2023)
Mobile Manipulation	Robotic navigation and manipulation planning with VQA	2912 sequences of action	Driess et al. (2023)
VQA2v	Question answering with images and prompts	1.1M questions with images	Goyal et al. (2019)
COCO	Image captioning	123287 images with captions	Chen et al. (2015)
OK-VQA	Visual question answering requiring external knowledge	14031 questions with images	Marino et al. (2019)
YouCook	Instructional video captioning	2000 videos with caption	Sun et al. (2019)
CLIP benchmark	Image captioning	400M image-text pairs	Radford et al. (2021)

Table 3: Summary of MLLP datasets

B Multimodal BERT Objectives

Pretraining of a multimodel is commonly conducted in a step-by-step fashion: first, the encoders of various modalities are trained, then the pretrained models are concatenated into a multimodel and pretrained on visual-linguistic tasks to acquire a general embedding space of visuals and texts. Here, we explain the pretraining objectives of VideoBERT (Sun et al., 2019) as a paradigm of multimodel pretraining.

VideoBERT is a masked multimodal large language model. With the images summarised into 2^14 discrete tokens, both word embeddings and image tokens can be accepted as input to the BERT model. For each modality, the sequences of words or video frames are randomly masked so the model is trained to restore the maksed embeddings. The video-only training objective is analogous to the text-only or the Valina BERT training objective. It helps the model to acquire long-term state dependencies in video frames. Formally, training each modality separately ensures the model's capability to capture the marginal distribution of image $P(x)$ or text $P(y)$.

The third training objective to establish the joint distribution $P(x, y)$ so that the model understands the interplay of video and texts. To encode image information into sequences of word embeddings, word-based sentences, and visual tokens are combined into visual sentences joined by a special token "[>]."

```
[CLS] orange chicken with [MASK] sauce
 [>] v01 [MASK] v08 v72 [SEP]
```

Researchers proposed a *linguistic-visual alignment* task, where the state of the "[cls]" token is used to determine whether the linguistic sequence and the visual sequence are aligned, which is, to some degree, similar to the *next-sentence prediction* task in the vanilla BERT model. This objective essentially establishes the connection between the two domains.

After pretraining, the model would learn a joint representation of multi-modalities. Further finetuning allows it to be used in various downstream tasks.