A Survey of MultiModal Large Language Models

Yahan Yu[1] Duzhen Zhang[2,3] Chenhui Chu[1]
[1]Kyoto University [2]Tencent AI Lab, China [3]Mohamed bin Zayed University of Artificial Intelligence
yahan@nlp.ist.i.kyoto-u.ac.jp, duzhen.zhang@mbzuai.ac.ae, chu@i.kyoto-u.ac.jp

Abstract

In recent years, MultiModal Large Language Models (MM-LLMs) have undergone substantial advancements, augmenting off-the-shelf LLMs to support MM inputs or outputs via cost-effective training strategies. In this paper, we provide a survey aimed at facilitating further research on MM-LLMs. We outline general design formulations for model architecture. Furthermore, we review the performance of selected MM-LLMs on mainstream benchmarks and explore future directions. More latest developments in this field are provided in a real-time tracking website.[1)] We hope that this survey contributes to the ongoing advancement of the MM-LLMs domain.

1 Introduction

MultiModal (MM) pre-training has advanced significantly, improving performance across various tasks [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. However, as models and datasets grow, training from scratch becomes computationally expensive. A promising approach leverages pre-trained foundation models, especially Large Language Models (LLMs) [13], to reduce costs and improve efficiency, giving rise to the emerging field of MM-LLMs.

MM-LLMs utilize LLMs as the core, offering robust language generation, while other foundation models provide high-quality representations. The main challenge lies in effectively connecting LLMs with other modalities. Research focuses on improving modality alignment and human intent alignment through Pre-Training (PT) and Instruction-Tuning (IT).

Figure 1 illustrates the evolution of MM-LLMs. Investigation of MM-LLMs initially focuses on MM comprehension and text generation tasks, such as image-text understanding (e.g., BLIP-2 [14], LLaVA [15], and MiniGPT-4 [16]), video-text understanding (e.g., VideoChat [17], Video-ChatGPT [18], and LLaMA-VID [19]), and audio-text understanding (e.g., Qwen-Audio [20]). Later research extended MM-LLMs to support specific modality outputs, including image-text output (e.g., GILL [21], Kosmos-2 [22], Emu [23], and MiniGPT-5 [24]) and audio-text output (e.g., SpeechGPT [25]). Recent efforts target human-like any-to-any modality conversion (e.g., NExT-GPT [26]) to reduce errors in cascaded systems.

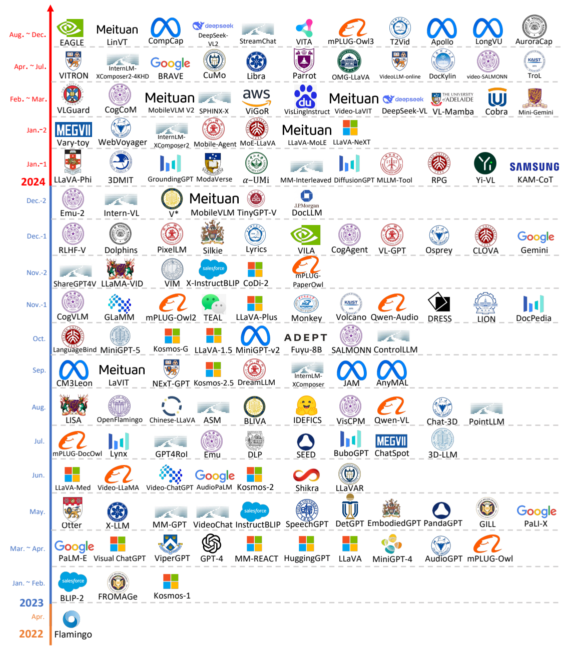

Figure 1 The timeline of MM-LLMs.

In this paper, we present a survey on MM-LLM research. We outline general design principles and the training pipeline. We review benchmark performance of the latest SOTA MM-LLMs, and propose future research directions. We aim to deepen understanding and inspire the development of more effective MM-LLMs.

2 Model Architecture

This section details the five components of the general model architecture and their implementation, as shown in

1) https://mm-llms.github.io

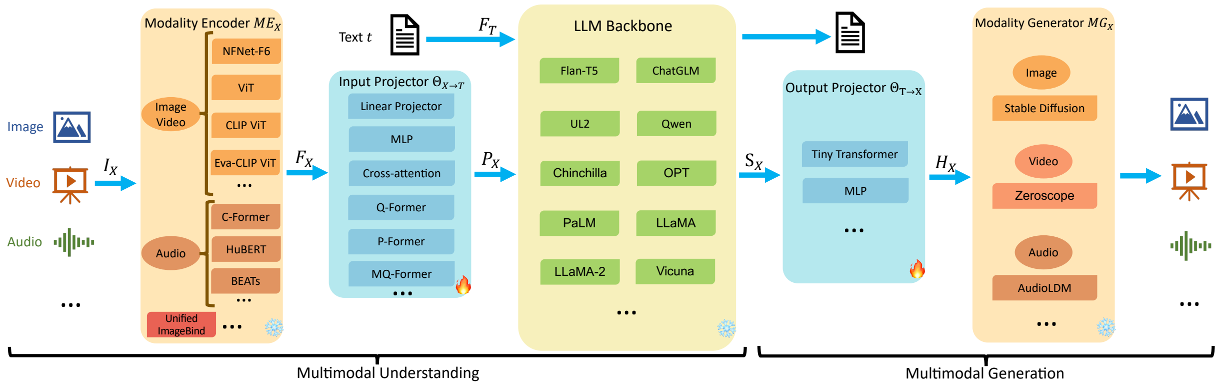

Figure 2 The general model architecture of MM-LLMs and the implementation choices for each component.

Figure 2. During training, the Modality Encoder, LLM Backbone, and Modality Generator are typically frozen, with optimization centered on the lightweight Input and Output Projectors, which constitute around 2% of the total parameters.

2.1 Modality Encoder

The Modality Encoder (ME) is tasked with encoding inputs from diverse modalities I_X to obtain corresponding features $\boldsymbol{F}_X$, formulated as $\boldsymbol{F}_X = \text{ME}_X(I_X)$. Various pre-trained encoder options ME_X exist for handling different modalities, where X can be image, video, audio, 3D, etc.

Visual Modality For images, there are various optional encoders: **NFNet-F6** [27], **ViT** [28], **CLIP ViT** [6], **Eva-CLIP ViT** [29], **BEiT-3** [30], and **OpenCLIP** [31], etc. For videos, they can be uniformly sampled to 5 frames, undergoing the same pre-processing as images.

Audio Modality is typically encoded by **C-Former** [32], **HuBERT** [33], **BEATs** [34], **Whisper** [35], and **CLAP** [36].

3D Point Cloud Modality is typically encoded by **ULIP-2** [37] with a PointBERT [38] backbone.

Moreover, to handle numerous heterogeneous modal encoders, some MM-LLMs, particularly any-to-any ones, use **ImageBind** [39], a unified encoder covering six modalities, including image/video, text, audio, heat map, inertial measurement units, and depth.

2.2 Input Projector

The Input Projector $\boldsymbol{\Theta}_{X\to T}$ is tasked with aligning the encoded features of other modalities $\boldsymbol{F}_X$ with the text feature space T. The aligned features as prompts $\boldsymbol{P}_X$ are then fed into LLM Backbone alongside the textual features $\boldsymbol{F}_T$. Given X-text dataset $\{I_X, t\}$, the goal is to minimize the X-conditioned text generation loss $\mathscr{L}_{\text{txt-gen}}$:

$$\underset{\boldsymbol{\Theta}_{X\to T}}{\arg\min}\, \mathscr{L}_{\text{txt-gen}}(\text{LLM}(\boldsymbol{P}_X, \boldsymbol{F}_T), t), \tag{1}$$

where $\boldsymbol{P}_X = \boldsymbol{\Theta}_{X\to T}(\boldsymbol{F}_X)$.

$\boldsymbol{\Theta}_{X\to T}$ can be achieved directly by a **Linear Projector**, or Multi-Layer Perceptron (**MLP**), or more complex implementations like **Cross-attention** and **Q-Former** [14]. **Cross-attention** [40] uses a set of trainable vectors as queries and $\boldsymbol{F}_X$ as keys to compress the feature sequence to a fixed length, and then fed them into the LLM. **Q-Former** extracts relevant features from $\boldsymbol{F}_X$ with learnable queries, and the selected features are then used as prompts $\boldsymbol{P}_X$. Meanwhile,

2.3 LLM Backbone

Taking LLMs [41] as the core agents, MM-LLMs can inherit some notable properties like zero-shot generalization. The LLM Backbone produces direct textual outputs t, and signal tokens $\boldsymbol{S}_X$ from other modalities (if any). These signal tokens act as instructions to guide the generator on whether to produce MM contents and, if affirmative, specify the content to produce t, $\boldsymbol{S}_X = \text{LLM}(\boldsymbol{P}_X, \boldsymbol{F}_T)$, where the aligned representations of other modalities $\boldsymbol{P}_X$ can be considered as soft Prompt-tuning for the LLM. Moreover, some works have introduced Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA [42]. In these cases, the number of additional trainable parameters is exceptionally minimal, even less than 0.1% of the total LLM parameter count.

2.4 Output Projector

The Output Projector $\boldsymbol{\Theta}_{T\to X}$ maps $\boldsymbol{S}_X$ into features $\boldsymbol{H}_X$ understandable to the following Modality Generator MG_X. To facilitate alignment of the mapped $\boldsymbol{H}_X$, the goal is to

Table 1 The summary of mainstream MM-LLMs. I→O: Input to Output Modalities, I: Image, V: Video, A: Audio, and T: Text.

Model	I→O	Modality Encoder	Input Projector	LLM Backbone	Output Projector	Modality Generator
BLIP-2	I+T→T	I: CLIP/Eva-CLIP ViT@224	Q-Former w/ Linear Projector	Flan-T5/OPT	–	–
LLaVA	I+T→T	I: CLIP ViT-L/14	Linear Projector	Vicuna-7B/13B	–	–
MiniGPT-4	I+T→T	I: Eva-CLIP ViT-G/14	Q-Former w/ Linear Projector	Vicuna-13B	–	–
mPLUG-Owl	I+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–	–
InstructBLIP	I+V+T→T	I/V: ViT-G/14@224	Q-Former w/ Linear Projector	Flan-T5/Vicuna	–	–
Video-LLaMA	I+V+A+T→T	I/V: Eva-CLIP ViT-G/14; A: ImageBind	Q-Former w/ Linear Projector	Vicuna/LLaMA	–	–
mPLUG-DocOwl	I_D+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–	–
Qwen-VL-Chat	I+T→T	I: ViT@448	Cross-attention	Qwen-7B	–	–
LaVIT	I+T→I+T	I: ViT	Cross-attention	LLaMA-7B	–	I: Stable Diffusion
MiniGPT-5	I+T→I+T	I: Eva-CLIP ViT-G/14	Q-Former w/ Linear Projector	Vicuna-7B	Tiny Transformer	I: StableDiffusion-2
LLaVA-1.5	I+T→T	I: CLIP ViT-L@336	MLP	Vicuna-v1.5-7B/13B	–	–
MiniGPT-v2	I+T→T	I: Eva-CLIP ViT@448	Linear Projector	LLaMA-2-Chat-7B	–	–
CogVLM	I+T→T	I: Eva-2-CLIP ViT	MLP	Vicuna-v1.5-7B	–	–
Qwen-Audio	A+T→T	A: Whisper-L-v2	Linear Projector	Qwen-7B	–	–
VILA	I+T→T	I: ViT@336	Linear Projector	LLaMA-2-7B/13B	–	–
LongVU	V+T→T	SigLIP + DINOv2	Cross-attention	Llama3.2-3B/Qwen2-7B	–	–

minimize the distance between $\boldsymbol{H}_X$ and the conditional text representations of MG_X: $\arg\min_{\boldsymbol{\Theta}_{T\to X}} \mathscr{L}_{\mathrm{mse}}(\boldsymbol{H}_X, \tau_X(t))$. The optimization only relies on captioning texts, without utilizing any audio or visual resources X, where $\boldsymbol{H}_X = \boldsymbol{\Theta}_{T\to X}(\boldsymbol{S}_X)$ and τ_X is the textual condition encoder in MG_X. The Output Projector is implemented by a **Tiny Transformer** with a learnable decoder feature sequence or **MLP**.

2.5 Modality Generator

The Modality Generator MG_X is tasked with producing outputs in distinct modalities. Commonly, existing works use off-the-shelf Latent Diffusion Models (LDMs) [43], **i.e.**, **Stable Diffusion** [44] for image synthesis, **Zeroscope** [45] for video synthesis, and **AudioLDM-2** [46, 47] for audio synthesis. $\boldsymbol{H}_X$ mapped by the Output Projector serves as conditional inputs in the denoising process to generate MM content.

3 Training Pipeline

MM-LLMs' training pipeline can be delineated into MM PT stage and MM IT stage. During the PT stage, typically leveraging the X-Text datasets, Input and Output Projectors are trained to achieve alignment among various modalities by optimizing predefined objectives.

MM IT comprises Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), aiming to align with human intents and enhance the interaction capabilities of MM-LLMs. SFT converts part of the PT stage data into an instruction-aware format. Next, it fine-tunes pre-trained MM-LLMs using the same optimization objectives. After SFT, RLHF involves further fine-tuning of the model, relying on feedback regarding the MM-LLMs' responses (**e.g.**, Natural Language Feedback (NLF) labeled manually or automatically) [48]. This process employs a reinforcement learning algorithm to effectively integrate the non-differentiable NLF [49, 50].

4 SOTA MM-LLMs

Based on the previously defined design formulations, we conduct a comprehensive comparison of the architectures and training dataset scales for current SOTA MM-LLMs, as illustrated in Table 1.

Trends in Existing MM-LLMs: (1) Progressing from a dedicated emphasis on MM understanding to the generation of specific modalities and further evolving into any-to-any modality conversion; (2) Adopting a More Efficient Model Architecture, transitioning from complex Q- and P-Former input projector modules in BLIP-2 and DLP to a simpler yet effective linear projector in VILA; (3) From producing foundational multimodal models to leveraging existing models to achieve more challenging goals and focus on more specialized problems (**e.g.**, Video-LLaVA → LongVU).

5 Benchmarks and Performance

To provide a comprehensive comparison, we have compiled a table featuring major MM-LLMs across Vision-Language (VL) benchmarks, as reported in various papers [14, 51, 52, 53]. Results are presented in Table 2. Given the numerous benchmarks available, we focus on evaluating and comparing different MM-LLMs based on OKVQA, IconVQA, VQA^{v2}, and GQA.

OKVQA requires reasoning with a variety of knowledge

Table 2 Comparison of mainstream MM-LLMs on VL benchmarks. The **red** denotes the highest result, and the **blue** denotes the second highest result.

Model	LLM Backbone	OKVQA	IconVQA	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME^{P}	MME^{C}	MMB	MMB^{CN}	$SEED^{I}$	$LLaVA^{W}$	MM-Vet	QBench	HM	VSR
BLIP-2	Flan-$T5_{XXL}$($13B$)	45.9	40.6	65.0	44.7	19.6	61.0	42.5	85.3	1293.8	290.0	–	–	46.4	38.1	22.4	–	53.7	50.9
LLaVA	Vicuna-13B	54.4	43.0	–	41.3	–	–	38.9	–	–	–	–	–	–	–	–	–	–	51.2
MiniGPT-4	Vicuna-13B	37.5	37.6	–	30.8	–	–	19.4	–	–	–	–	–	–	–	–	–	–	41.6
InstructBLIP	Vicuna-7B	–	–	–	49.2	34.5	60.5	50.1	–	–	–	36.0	23.7	53.4	60.9	26.2	56.7	–	–
Qwen-VL	Qwen-7B	–	–	78.8	59.3	35.2	67.1	63.8	–	–	–	38.2	7.4	56.3	–	–	**59.4**	–	–
Qwen-VL-Chat	Qwen-7B	–	–	78.2	57.5	38.9	68.2	61.5	–	1487.5	**360.7**	60.6	56.7	58.2	–	–	–	–	–
LLaVA-1.5	Vicuna-1.5-7B	–	–	78.5	62.0	50.0	66.8	58.2	**85.9**	1510.7	**316.1**	64.3	58.3	58.6	63.4	30.5	58.7	–	–
LLaVA-1.5	Vicuna-1.5-13B	–	–	**80.0**	**63.3**	53.6	71.6	61.3	**85.9**	1531.3	295.4	67.7	**63.6**	61.6	**70.7**	**35.4**	**62.1**	–	–
MiniGPT-v2	LLaMA-2-Chat-7B	**56.9**	**47.7**	–	60.3	30.3	–	51.9	–	–	–	–	–	–	–	–	–	**58.2**	**60.6**
MiniGPT-v2-Chat	LLaMA-2-Chat-7B	**55.9**	**49.4**	–	58.8	42.4	–	52.3	–	–	–	–	–	–	–	–	–	**59.5**	**63.3**
VILA-7B	LLaMA-2-7B	–	–	79.9	62.3	**57.8**	68.2	64.4	**85.5**	1533.0	–	68.9	61.7	61.1	69.7	34.9	–	–	–
VILA-13B	LLaMA-2-13B	–	–	**80.8**	**63.3**	**60.6**	73.7	66.6	84.2	**1570.1**	–	70.3	**64.3**	62.8	**73.0**	**38.8**	–	–	–
StreamChat-7B	Qwen-7B	–	–	–	**62.4**	–	**85.5**	**72.4**	–	1520.0	–	**74.4**	–	**74.3**	–	–	–	–	–
StreamChat-14B	Qwen-14B	–	–	–	**63.3**	–	**85.8**	**74.4**	–	**1617.0**	–	**79.0**	–	**75.5**	–	–	–	–	–

types such as commonsense. MiniGPT-v2 and MiniGPT-v2-chat perform best in this benchmark, showcasing their outstanding reasoning abilities. IconVQA emphasizes the importance of holistic cognitive reasoning in real-world diagram-based word problems, requiring both perceptual acumen and versatile cognitive reasoning. MiniGPT-v2 and MiniGPT-v2-chat also perform best, highlighting their exceptional perception and cognitive reasoning capabilities. VQA^{v2} is a more balanced VQA dataset. VILA-13B performs best, demonstrating its resistance to language biases in the knowledge it acquires. GQA focuses on image scene graphs, offering impartial compositional questions derived from real-world images. Each question is associated with a structured representation of its meaning and the detailed logical steps required to answer it. StreamChat performs best in this benchmark, illustrating their excellent reasoning abilities.

These findings can inspire training recipes. Firstly, higher image resolution can incorporate more visual details for the model, benefiting tasks that require fine-grained details. For example, LLaVA-1.5 and VILA employ a resolution of 336×336, while Qwen-VL and MiniGPT-v2 utilize 448×448. Moreover, StreamChat and VILA reveal several key findings: (1) A dense instruction dataset is crucial to facilitate the training of MM-LLMs; (2) Re-blending text-only instruction data (**e.g.**, unnatural instruction [54]) with image-text data during SFT not only addresses the degradation of text-only tasks but also enhances VL task accuracy.

6 Future Directions

We can enhance the MM-LLMs' strength from the following four key avenues: **(1) Expanding Modalities**: Current MM-LLMs mainly support the following modalities: image, video, audio, 3D, and text. However, the real world involves a broader range of modalities. Extending MM-LLMs to accommodate additional modalities (**e.g.**, web pages, heat maps, and figures&tables) will increase the model's versatility, making it more universally applicable; **(2) Diversifying LLMs**: Incorporating various types and sizes of LLMs provides practitioners with the flexibility to select the most appropriate one based on their specific requirements; **(3) Improving MM IT Dataset Quality**: Current MM IT datasets have ample room for improvement and expansion. Diversifying the range of instructions can enhance the effectiveness of MM-LLMs in understanding and executing user commands; **(4) Strengthening MM Generation Capabilities**: Most current MM-LLMs are predominantly oriented towards MM understanding. Although some models have incorporated MM generation capabilities, the quality of generated responses may be constrained by the capacities of the LDMs. Exploring the integration of retrieval-based approaches [55, 56, 57] holds significant promise in complementing the generative process, enhancing the overall performance of the model.

7 Conclusion

In this paper, we presented a survey of MM-LLMs focusing on recent advancements. Initially, we categorize the model architecture into five components, providing a detailed overview of general design formulations and training pipelines. Subsequently, we introduced various SOTA MM-LLMs, shed light on their capabilities across diverse MM benchmarks, and envision future developments in this rapidly evolving field. Although MM-LLMs have made many breakthroughs, there is still room for improvement. We hope this survey can provide insights and contribute to the ongoing advancements in the MM-LLMs domain.

Acknowledgements

This work was supported by JST BOOST Grant Number JPMJBS2407 and JSPS KAKENHI Grant Number JP23K28144.

References

[1] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In **Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16**, pp. 121–137. Springer, 2020.

[2] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. **Advances in Neural Information Processing Systems**, Vol. 34, pp. 24206–24221, 2021.

[3] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. **arXiv preprint arXiv:2106.11097**, 2021.

[4] Rui Yan, Mike Zheng Shou, Yixiao Ge, Alex Jinpeng Wang, Xudong Lin, Guanyu Cai, and Jinhui Tang. Video-text pre-training with learned regions. **arXiv preprint arXiv:2112.01194**, 2021.

[5] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. **Advances in neural information processing systems**, Vol. 34, pp. 9694–9705, 2021.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PMLR, 2021.

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In **International Conference on Machine Learning**, pp. 12888–12900. PMLR, 2022.

[8] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 16375–16387, 2022.

[9] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In **International Conference on Machine Learning**, pp. 25994–26009. PMLR, 2022.

[10] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 15671–15680, 2022.

[11] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In **International Conference on Machine Learning**, pp. 23318–23340. PMLR, 2022.

[12] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. **arXiv preprint arXiv:2208.10442**, 2022.

[13] OpenAI. OpenAI: Introducing ChatGPT. 2022.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In **International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA**, pp. 19730–19742, 2023.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[16] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. **arXiv preprint arXiv:2304.10592**, 2023.

[17] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. **arXiv preprint arXiv:2305.06355**, 2023.

[18] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. **arXiv preprint arXiv:2306.05424**, 2023.

[19] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. **arXiv preprint arXiv:2311.17043**, 2023.

[20] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. **arXiv preprint arXiv:2311.07919**, 2023.

[21] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[22] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding Multimodal Large Language Models to the World. **arXiv preprint arXiv:2306.14824**, 2023.

[23] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. In **The Twelfth International Conference on Learning Representations**, 2024.

[24] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. **arXiv preprint arXiv:2310.02239**, 2023.

[25] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In **Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023**, pp. 15757–15773, 2023.

[26] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. **arXiv preprint arXiv:2309.05519**, 2023.

[27] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In **International Conference on Machine Learning**, pp. 1059–1071. PMLR, 2021.

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In **International Conference on Learning Representations**, 2020.

[29] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 19358–19369, 2023.

[30] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 19175–19186, 2023.

[31] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 2818–2829, 2023.

[32] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. **arXiv preprint arXiv:2305.04160**, 2023.

[33] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 29, pp. 3451–3460, 2021.

[34] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio Pre-Training with Acoustic Tokenizers. In **International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA**, pp. 5178–5193, 2023.

[35] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In **International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA**, pp. 28492–28518, 2023.

[36] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In **ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 1–5. IEEE, 2023.

[37] Salesforce. Ulip. 2022.

[38] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 19313–19322, 2022.

[39] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 15180–15190, 2023.

[40] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 23716–23736, 2022.

[41] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023.

[42] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In **International Conference on Learning Representations**, 2021.

[43] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In **International Conference on Learning Representations**, 2021.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 10684–10695, 2022.

[45] Cerspense. Zeroscope: Diffusion-based text-to-video synthesis. 2023.

[46] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In **International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA**, pp. 21450–21474, 2023.

[47] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining. **CoRR**, Vol. abs/2308.05734, , 2023.

[48] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. **arXiv preprint arXiv:2309.14525**, 2023.

[49] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. **arXiv preprint arXiv:2311.10081**, 2023.

[50] Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs. **arXiv preprint arXiv:2305.08844**, 2023.

[51] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. **arXiv preprint arXiv:2310.09478**, 2023.

[52] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. **arXiv preprint arXiv:2311.12793**, 2023.

[53] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models. **arXiv preprint arXiv:2312.07533**, 2023.

[54] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. **arXiv preprint arXiv:2212.09689**, 2022.

[55] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Retrieval-based language models and applications. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)**, pp. 41–46, 2023.

[56] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. **arXiv preprint arXiv:2312.10997**, 2023.

[57] Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models. **arXiv preprint arXiv:2402.03181**, 2024.