Frontier Review of Multimodal AI

Nan Duan
Microsoft Research Asia
nanduan@microsoft.com

Abstract

Pre-training techniques have enabled foundation models (such as BERT, T5, GPT) to achieve remarkable success in natural language processing (NLP) and multimodal tasks that involve text, audio and visual contents. Some of the latest multimodal generative models, such as DALL·E and Stable Diffusion, can synthesize novel visual content from text or video inputs, which greatly enhances the creativity and productivity of content creators. However, multimodal AI also faces some challenges, such as adding new modalities or handling diverse tasks that require signals beyond their understanding. Therefore, a new trend in multimodal AI is to build a compositional AI system that connects existing foundation models with external modules and tools. This way, the system can perform more varied tasks by leveraging different modalities and signals. In this paper, we will give a brief overview of the state-of-the-art multimodal AI techniques and the direction of building compositional AI systems. We will also discuss the potential future research topics in multimodal AI.

1 Introduction

Large language models (LLMs) have achieved great success in natural language processing (NLP). These models (e.g., BERT (Devlin et al., 2019), T5 (Raffel et al., 2020) and GPT (Brown et al., 2020)) can learn general data representations and commonsense knowledge from large-scale corpora using self-supervised learning tasks (such as masked language modeling or next token prediction). The learned models can be further fine-tuned on downstream tasks and obtain superior performance on them.

The success of LLMs has also been extended to other non-language domains, such as computer vision or speech processing. The convergence of these techniques on different types of data makes "multimodal AI" the hottest direction in the AI community.

This paper aims to briefly summarize the latest trends of multimodal AI research. In short, there are three trends as follows: (1) the underlying architectures of models for different modalities are converging; (2) the focus of multimodal AI research is shifting from multimodal understanding models to multimodal generation models; (3) single multimodal models have shown limitations and they are still far from covering diverse tasks using data with different modalities, and connecting LLMs with external tools and models to complete more tasks is becoming the new AI paradigm. We will introduce these three trends in Sections 2, 3 and 4, respectively. In Section 5, we will discuss the possible future directions of multimodal AI.

2 The Convergence of Model Architecture

The architecture of models for different modalities is becoming more similar in the era of LLMs. Transformers are widely used in text, code, visual and audio scenarios to support understanding and generation tasks. For instance, the latest LLMs like ChatGPT or GPT-4 have integrated text and code in a single model, which can support text-only, code-only, text-to-code and code-to-text generation tasks. Multimodal generation models like DALL·E (Ramesh et al., 2021) and NUWA-Infinity (Wu et al., 2022) are

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 110-118, Harbin, China, August 3 - 5, 2023.
Volume 2: Frotier Forum
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

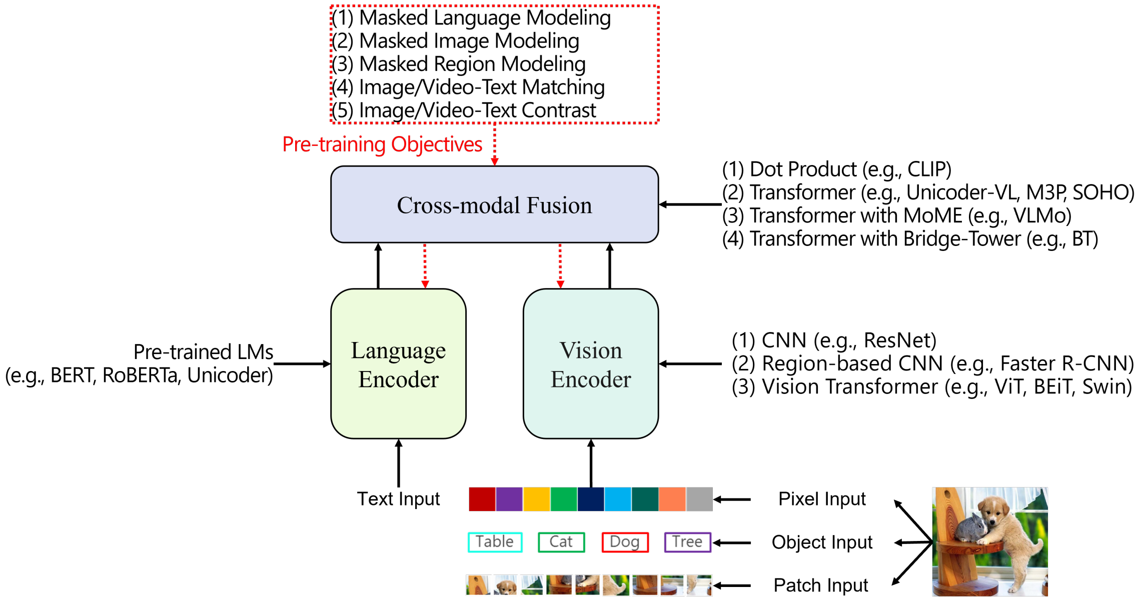

(1) Masked Language Modeling
(2) Masked Image Modeling
(3) Masked Region Modeling
(4) Image/Video-Text Matching
(5) Image/Video-Text Contrast
Pre-training Objectives
Cross-modal Fusion
(1) Dot Product (e.g., CLIP)
(2) Transformer (e.g., Unicoder-VL, M3P, SOHO)
(3) Transformer with MoME (e.g., VLMo)
(4) Transformer with Bridge-Tower (e.g., BT)
Pre-trained LMs
(e.g., BERT, RoBERTa, Unicoder)
Language Encoder
Vision Encoder
(1) CNN (e.g., ResNet)
(2) Region-based CNN (e.g., Faster R-CNN)
(3) Vision Transformer (e.g., ViT, BEiT, Swin)
Text Input
Pixel Input
Table
Cat
Dog
Tree
Object Input
Patch Input

Figure 1: Overview of visual-language models.

also trained based on auto-regressive models like GPT models for image and video generation tasks. VALL-E (Wang et al., 2023) can leverage strong in-context learning capabilities and can be applied for zero-shot cross-lingual text-to-speech synthesis and zero-shot speech-to-speech translations, which is also based on Transformer and GPT-like models. Moreover, we also observed that diffusion models are widely used in content generation tasks as well, such as DALL·E 2 (Ramesh et al., 2022) and Stable Diffusion (Rombach et al., 2022) for visual generation, NaturalSpeech 2 (Shen et al., 2023) for speech generation. But there is also another research thread that aims to unify different types of generation models using diffusion models, which can be also seen as an indication of the model architecture convergence.

Due to the different basic units, data formats, and structures of the contents in different modalities, there is still no universally agreed model architecture for multimodal AIs. However, such convergence is definitely a clear trend in the AI community.

3 From Visual-Language Understanding to Visual Generation from Language

Visual-Text (VL) pre-trained models are the most representative multimodal AIs. The goal of such models is to learn the representations of texts and visuals jointly and support VL tasks such as image retrieval, visual question answering, or text-based image generation. In the past several years, the research focus has shifted from VL understanding tasks to visual generation tasks. Therefore, in this section, we will first review the progress of VL understanding models and then review the latest development of visual generation models from texts.

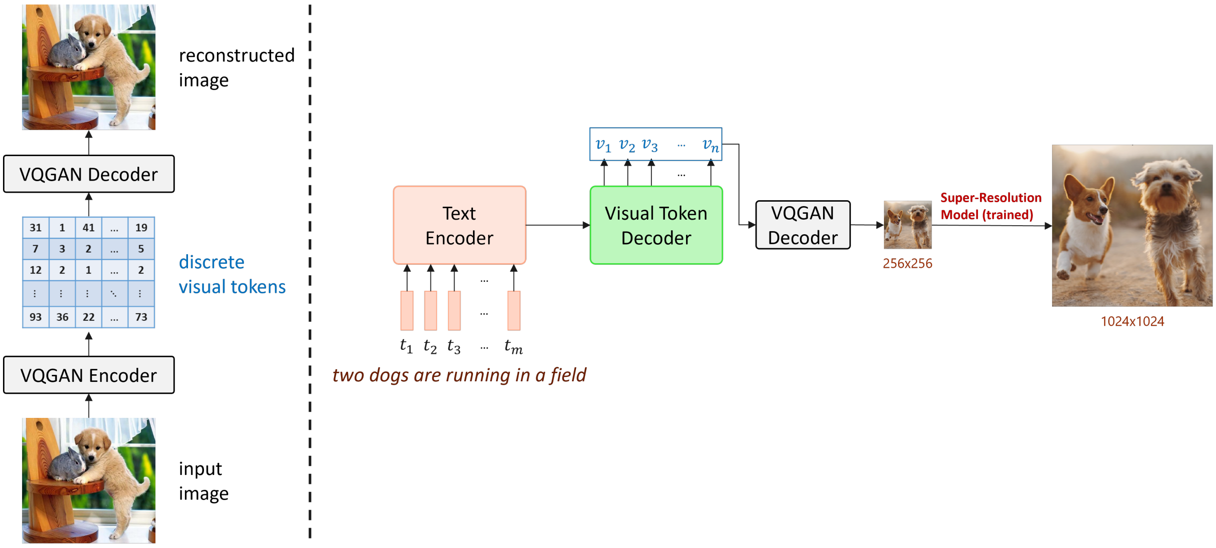
3.1 Visual-Language Understanding

There are 3 key differences between different VL understanding models.

First, how to represent visual inputs. Different VL understanding models use different granularity to represent visual contents, such as pixels, objects and patches of the images or videos. The most commonly used granularity recently is patches.

Second, how to generate visual representations. Some models use CNN-style models such as ResNet or Faster R-CNN, while other models use Transformers such as ViT (Dosovitskiy et al., 2021), Swin (Liu et al., 2021), etc.

Third, how to fuse the representations from text and visual inputs. There are several ways for this task. For example, CLIP (Radford et al., 2021) model uses a simple dot-product component in the



31	1	41	...	19
7	3	2	...	5
12	2	1	...	2
⋮	⋮	⋮	⋱	⋮
93	36	22	...	73

Figure 2: Overview of auto-regressive model.

fusion, which makes the computation cost very low and the resulting framework very effective in the image-text matching task. Some early VL understanding models (Unicoder-VL (Li et al., 2019), M3P (Ni et al., 2019), Uniter (Chen et al., 2020), etc.) used Transformers to further fuse the text and visual representations. Mixture-of-Experts are used to fuse representations from different input modalities as well, such as VLMo (Bao et al., 2022), which makes the model parameters for different modalities more tunable. Some recent work, such as BridgeTower (Xu et al., 2023) and ManagerTower (Xu et al., 2023), leveraged the text or visual representations from different layers to generate better uni-modal representations for the later VL understanding tasks.

In summary, using patches as the visual representation units and using Transformer to fuse text and visual representations is the current state-of-the-art VL pre-trained model setting. Besides images, video understanding is also very important for the development of many future AI systems. Currently image-based visual models are efficiently used in the video models. However, it is straightforward to leverage the large-scale video corpus directly in the future, which can train more powerful multimodal AI models for video-related tasks.

3.2 Visual Generation from Language

Currently, there are two typical text-based visual generation methodologies.

The 1st generation methodology is based on VQGAN (Yu et al., 2022) and autoregressive model. In VQGAN, an encoder can transform each image into discrete visual tokens. Each visual token is an integer code coming from a codebook and represents the content appeared in the corresponding image region. For example, the image region at the top-left corner is represented by a visual token whose ID is 31. Based on these visual tokens, a decoder can reconstruct the original image. It means if a natural language sentence can be translated into a visual token sequence, the VQGAN decoder can simply use the sequence to generate an image that reflects the meaning of the input sentence. This is exactly what DALL E (Ramesh et al., 2021) and Parti (Yu et al., 2022) do in their text-to-image generation procedures. In such models, a text encoder first encodes each natural language description into text embeddings and then a vision decoder follows an autoregressive formulation to generate visual tokens in a left-to-right order. This is similar to the typical text generation procedure in many NLP tasks such as machine translation or text summarization, where each word is generated based on all previous words already generated. Last, the pre-trained VQGAN decoder will generate an image based on the predicted visual tokens and a super-resolution model can further up-sample the output image to a bigger resolution. This methodology has its own pros and cons: thanks to the autoregressive mechanism, these models can capture the dependencies between generated visual tokens and also support variable-length generation

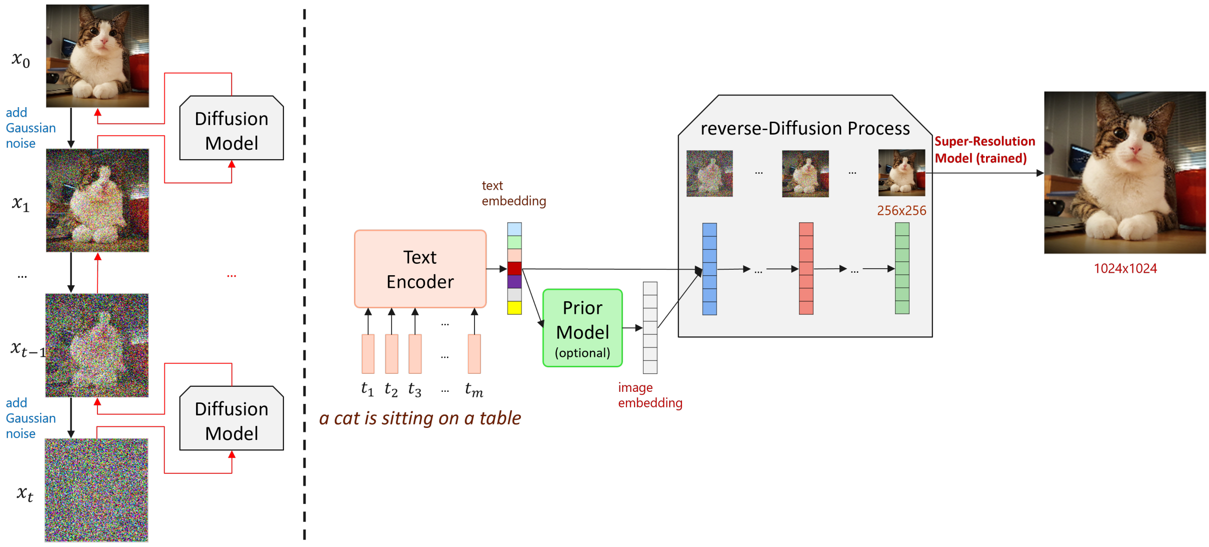


Figure 3: Overview of diffusion model.

tasks. But such generation process is not computationally efficient, as all tokens are generated one after one, instead of concurrently.

The 2nd way to generate images from text is to use diffusion models. Diffusion models work by adding noise to an image and then learning to remove the noise and recover the original image. This is called the forward and reverse diffusion processes. The reverse diffusion process can also use text or image as a condition to guide the image reconstruction. In diffusion model-based methods, a text encoder first turns text into embeddings. Then a reverse diffusion process uses noise and the text embedding to create output images. Some methods, like DALL E 2, also use a prior model to create an image embedding from the text embedding and use it as a condition for the reverse diffusion process to increase the image variety. Finally, a super-resolution model is used to make the output image bigger. Unlike autoregressive models, diffusion models are fast, because they can create the image at each time step at the same time. But they are not good at capturing the relationships between different parts of the image. They also cannot generate images of different sizes, because the image size is fixed beforehand.

To overcome the fixed-size limitation of diffusion models, NWUA-Infinity (Wu et al., 2022) proposed a method that can generate high-quality images and videos with any resolution, by creating them patch by patch. Given a text input and a resolution, a module called Arbitrary Direction Controller (or ADC) first decides the order of patch generation. Based on this order, NUWA-Infinity will create each patch one after another in the patch-level. For example, when it creates patch 13, a module called Nearby Context Pool (or NCP) first collects patch 7, 8, 9 and 12 as the context, because they are close to patch 13 within a certain distance. Then the vision decoder will create visual tokens for patch 13 based on these context patches and VQGAN decoder will create the corresponding image for patch 13. Because the vision decoder uses the nearby patches as its context when it creates each patch, the patches look smooth and natural when they are put together. This is how NUWA-Infinity can make the final image from all the patches. Also, because the number of context patches in NCP is small, as the model will discard those irrelevant patches during the generation process, the computation cost of the local autoregressive model can be greatly reduced, as it doesn't need to consider all patches created before. In this way, NUWA-Infinity can create the remaining patches and get the complete image output. NUWA-Infinity can also generate videos. The main difference from image generation is that the context patches in NCP come from both the current video frame and the previous video frames. For example, when NUWA-Infinity wants to create patch 14 in the second video frame, the context patches in NCP will include patch 1 to patch 9 from the first video frame and patch 10 to patch 13 from the second video frame. After patch 14 is created, it will be added to NCP as a new context and patch 1 will be removed from NCP as it has no impact on the future patches.

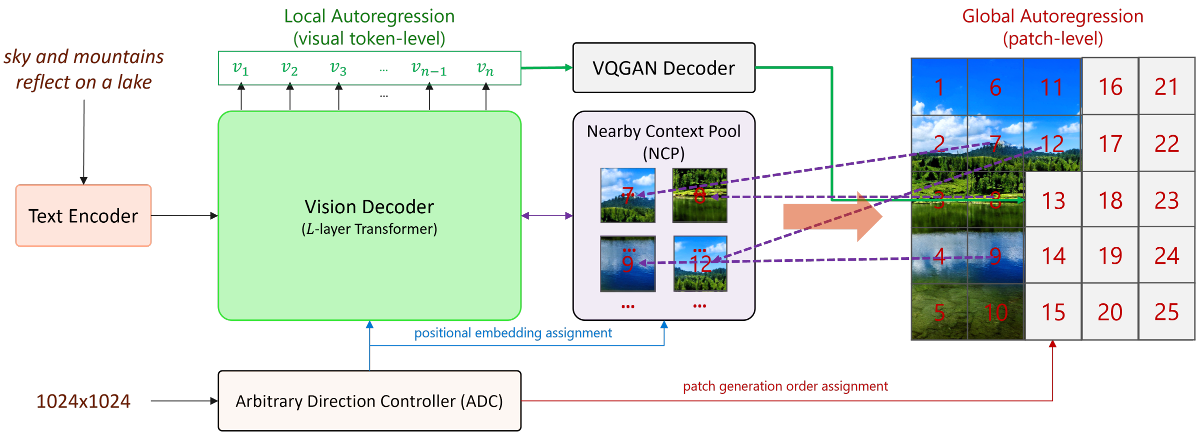


Figure 4: Overview of NUWA-Infinity for text-based image generation.

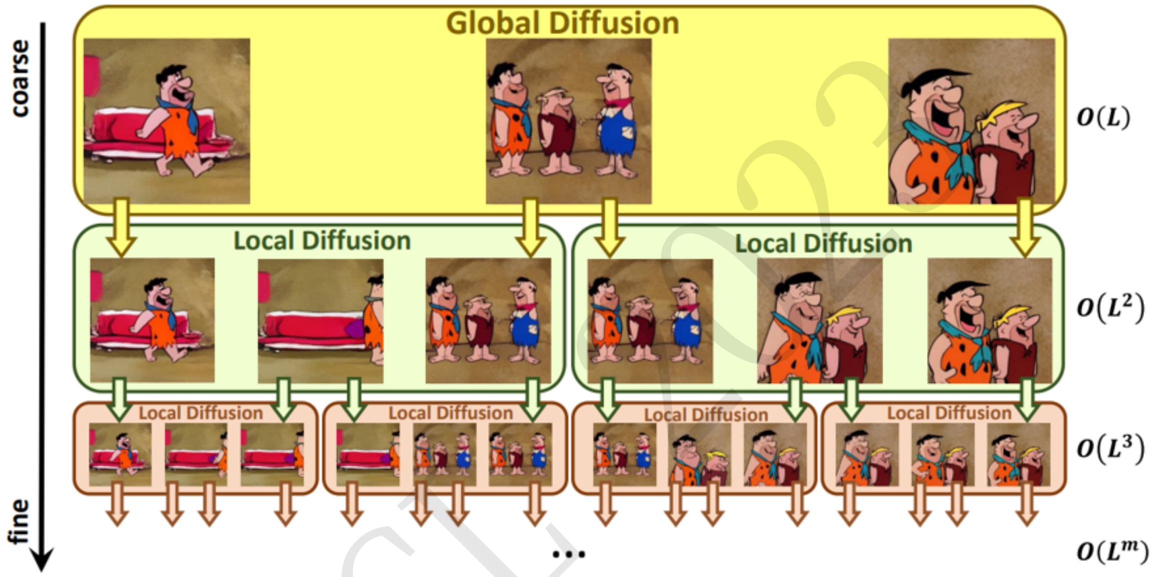
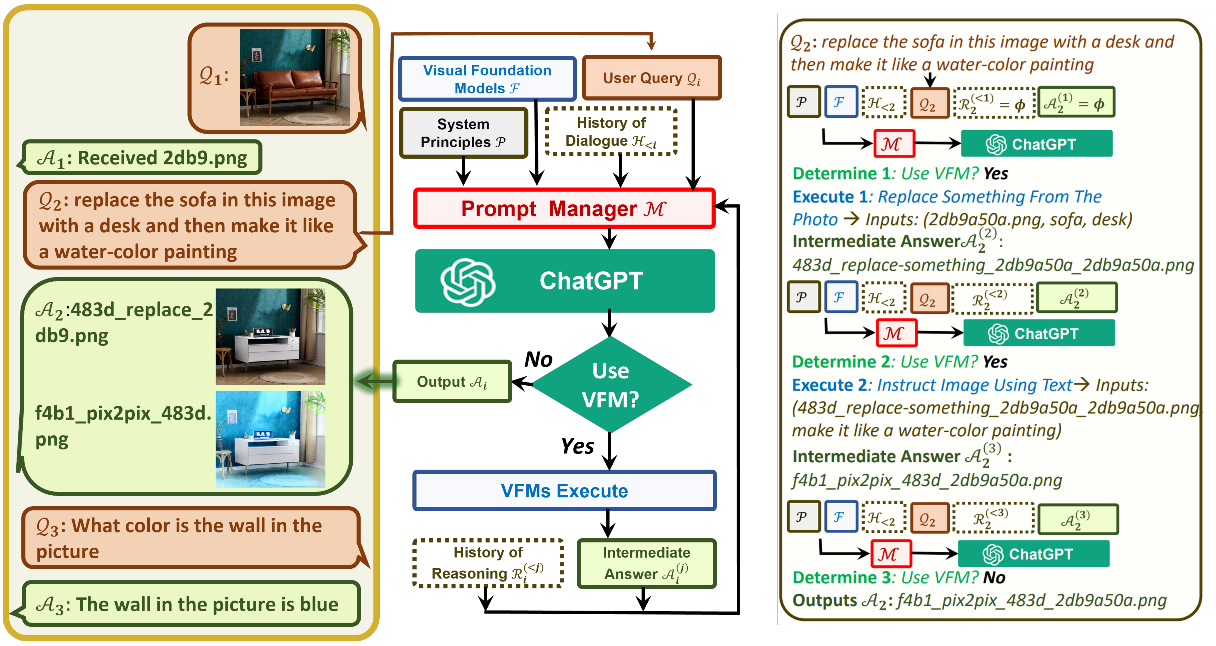


Figure 5: Overview of NUWA-XL for text-based extreme-long video generation.

NUWA-Infinity can create images and videos of different lengths with the auto-regressive over auto-regressive generation method. But auto-regressive models have some drawbacks, such as (1) they are very expensive to train and use; (2) they have error propagation problems that affect the generation quality; (3) they are not good at creating different scenes between images, which is important for video generation as scenes change often in video contents.

To solve these problems, NUWA-XL (Yin et al., 2023) proposed a diffusion over diffusion framework, which uses diffusion models in different levels to create long-videos in a fine-to-coarse way. In the first level, a diffusion model creates the key frames, which have enough scene changes and also keep the visual consistency between different video frames. In the second level, another diffusion model creates in-between video frames between any two adjacent video frames created in the first level. In the third level, a third diffusion model creates more in-between video frames between the adjacent video frames. By doing this, NUWA-XL can create very long videos efficiently and reduce the error propagation issue. Of course, the total scene length is still determined by the first level diffusion model, but creating a good key frame scene is much easier than creating the whole video at once.



$\mathcal{Q}_1$:
$\mathcal{A}_1$: Received 2db9.png
$\mathcal{Q}_2$: replace the sofa in this image with a desk and then make it like a water-color painting
$\mathcal{A}_2$:483d_replace_2db9.png
f4b1_pix2pix_483d.png
$\mathcal{Q}_3$: What color is the wall in the picture
$\mathcal{A}_3$: The wall in the picture is blue
Visual Foundation Models $\mathcal{F}$
User Query $\mathcal{Q}_i$
System Principles $\mathcal{P}$
History of Dialogue $\mathcal{H}_{<i}$
Prompt Manager $\mathcal{M}$
ChatGPT
Output $\mathcal{A}_i$
No
Use VFM?
Yes
VFMs Execute
History of Reasoning $\mathcal{R}_i^{(<j)}$
Intermediate Answer $\mathcal{A}_i^{(j)}$
$\mathcal{Q}_2$: replace the sofa in this image with a desk and then make it like a water-color painting
$\mathcal{P}$ $\mathcal{F}$ $\mathcal{H}_{<2}$ $\mathcal{Q}_2$ $\mathcal{R}_2^{(<1)} = \phi$ $\mathcal{A}_2^{(1)} = \phi$
$\mathcal{M}$ ChatGPT
Determine 1: Use VFM? Yes
Execute 1: Replace Something From The Photo → Inputs: (2db9a50a.png, sofa, desk)
Intermediate Answer $\mathcal{A}_2^{(2)}$: 483d_replace-something_2db9a50a_2db9a50a.png
$\mathcal{P}$ $\mathcal{F}$ $\mathcal{H}_{<2}$ $\mathcal{Q}_2$ $\mathcal{R}_2^{(<2)}$ $\mathcal{A}_2^{(2)}$
$\mathcal{M}$ ChatGPT
Determine 2: Use VFM? Yes
Execute 2: Instruct Image Using Text→ Inputs: (483d_replace-something_2db9a50a_2db9a50a.png make it like a water-color painting)
Intermediate Answer $\mathcal{A}_2^{(3)}$: f4b1_pix2pix_483d_2db9a50a.png
$\mathcal{P}$ $\mathcal{F}$ $\mathcal{H}_{<2}$ $\mathcal{Q}_2$ $\mathcal{R}_2^{(<3)}$ $\mathcal{A}_2^{(3)}$
$\mathcal{M}$ ChatGPT
Determine 3: Use VFM? No
Outputs $\mathcal{A}_2$: f4b1_pix2pix_483d_2db9a50a.png

Figure 6: Overview of Visual ChatGPT v1.

4 From Single AI to Compositional AI

Single AI models have obvious limitations. First, it is difficult to include a new modality in an existing multimodal model. This is because adding a new modality needs not only new data with this modality, but also training the model from scratch. So it requires a lot of work on data quality and computing resources, especially GPUs. Second, it is difficult to make a single AI model handle different tasks, even the most advanced LLMs like GPT-4 are not capable of this. This is because a single model is constrained by the current abilities and the predefined modalities.

Therefore, the community is starting to investigate compositional AI (Liang et al., 2023) as a possible new AI paradigm. This involves using and coordinating multiple AI modules with different functions to solve complex problems. Such systems can show new abilities that are beyond what any single module can do. We have seen some examples of this direction in the recent developments of LLMs, from single LLMs, to LLMs with expert sub-modules and the latest trend of combining LLMs with other tools and models to achieve more difficult tasks that are out of the scope of the original LLMs.

The benefits of compositional AI are quite obvious. First, it allows more control over the system's abilities by composing modules with specific functions. Second, it improves the system's interpretability and lowers the chance of hallucination by having clear definitions of modules. Third, it improves the system's continual learning ability and avoids the problem of catastrophic forgetting by not needing to update all modules in each new training stage. Fourth, it makes data collection and training easier for modules with simple skills. Fifth, it lowers data annotation and training costs by not needing to update all modules.

There are two ways to create compositional AI systems from multimodal tasks. First, LLMs can be integrated with external tools using fixed prompts, which can make the LLMs show new abilities on doing different multimodal tasks. Second, LLMs can be connected with new AI modules with specific functions using learnable parameters instead of fixed prompts. This soft connection can transfer information in different modalities in a smooth way and enable better multimodal abilities. Also, as only the learnable connectors are optimized during training with the original LLMs and other AI modules fixed, such system can quickly include new modalities in the system with low computation cost.

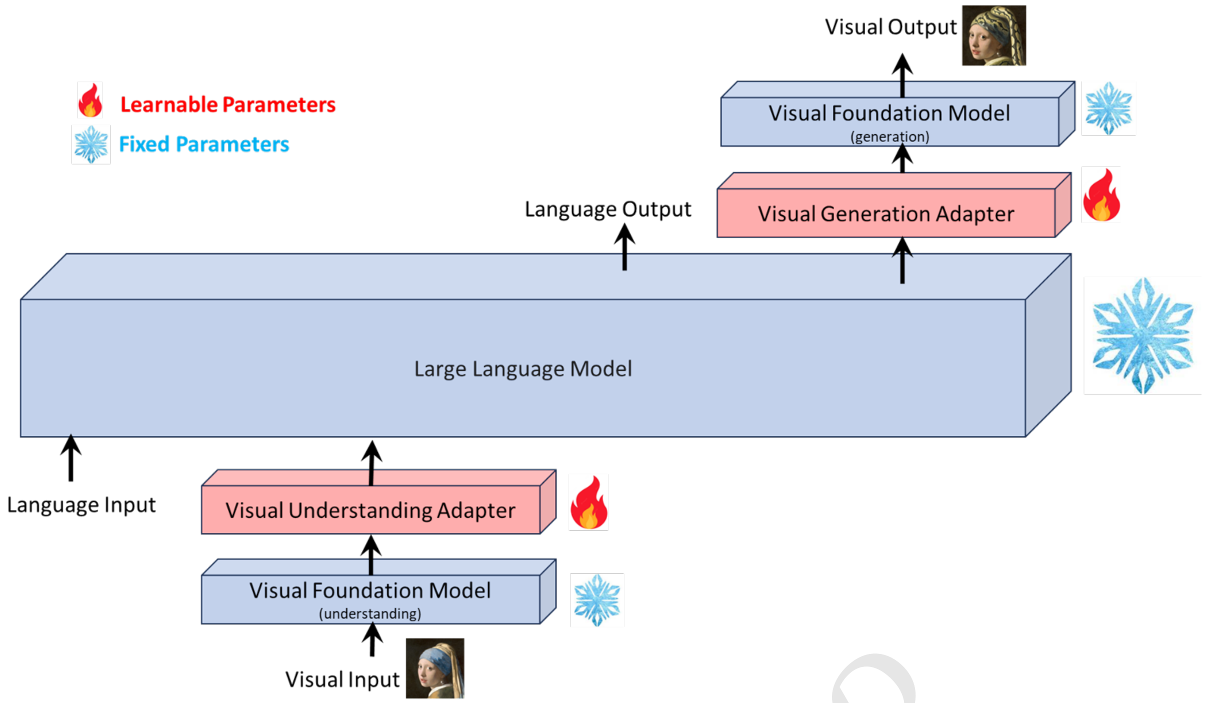


Figure 7: Overview of Visual ChatGPT v2.

4.1 Connecting LLMs with External Tools with Fixed Prompts

For the first type of work, we use Visual ChatGPT (Wu et al., 2023) as an example to illustrate how such work operates.

Visual ChatGPT is one of the first work that aims to combine visual tools with ChatGPT to perform different kinds of visual tasks. As ChatGPT is an LLM, which can only handle textual tasks, the first thing Visual ChatGPT needs to do is inform ChatGPT that it can try to use external visual tools to accomplish visual tasks. This work uses prompts as the system principles to let ChatGPT understand its new capabilities.

After adding system principles, Visual ChatGPT should also let ChatGPT know which tools it can use, when it can use them and how it can use them. For example, for Visual QA, the name and usage fields of this tool will briefly explain the function of the tool and when ChatGPT can use it, and the inputs/outputs field tells ChatGPT what kind of inputs and outputs are needed by this tool.

As the tasks require multiple steps to be completed, Visual ChatGPT also adds the string “do I need to use a tool” as another prompt after each user query, to let ChatGPT decide whether it needs to invoke a tool at the current step. If the answer is NO, then ChatGPT will return the current results to the user. Otherwise, ChatGPT will continue to call new tools and use all intermediate results as the context prompt in the next step.

By adding the above mentioned mechanisms, Visual ChatGPT can achieve many visual understanding, generation and editing tasks that the original ChatGPT model cannot do. This shows the biggest advantage of compositional AI models, new abilities will emerge by composing multiple tools with specific functions.

4.2 Connecting LLMs with External Modules with Learnable Parameters

Systems like Visual ChatGPT are easy to implement and build, as they do not require any weights to be learned. However, such systems also have obvious limitations. First, the fixed prompts are not stable and robust enough to link LLMs and tools. Second, in such systems, non-text information will be turned into text descriptions before sending them to LLMs. And such conversion will lose a lot of information of the original contents.

Therefore, there are recent related work that use learnable parameters to connect LLMs with other tool modules instead of using prompts. Such work will not change the parameters of the LLM and the external tool modules, as they are already well trained and further fine-tuning requires a lot of computing resources. Instead, they only train the adapters between LLM and tool modules using a small amount of annotations. By doing this, such system can do better message passing between LLM and other tool modules and avoid the catastrophic forgetting problem. For example, instead of converting the input image into a natural language description, Visual ChatGPT v2 gets the image representation based on a visual foundation model first, and then projects the image representation into the LLM input, by a visual understanding adapter. Similarly, another output adapter can be used to pass the LLM's output to the visual generation module, to create output images.

5 Future Directions

This paper briefly reviews the recent developments of multimodal AI research, including (1) the model architectures are becoming more similar, (2) the research focus is moving from multimodal understanding models to multimodal generation models; (3) combining LLMs with external tools and models to accomplish diverse tasks is emerging as the new AI paradigm.

There are several directions that can be further explored in the future. First, concentrating more on video generation, which could trigger the next ChatGPT breakthrough in the AI community. Second, concentrating more on compositional AI for multimodal systems with more modalities covered and less computation costs needed. Third, concentrating more on the autonomous robotics, which can complete more tasks in the physical world, to further enhance human's creativity and productivity.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, NAACL.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, JMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language Models are Few-Shot Learners*, arXiv.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. *Zero-Shot Text-to-Image Generation*, arXiv

Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. 2022. *NUWA-Infinity: Autoregressive over Autoregressive Generation for Infinite Visual Synthesis*, NeurIPS.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, Furu Wei. 2023. *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*, arXiv.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. *Hierarchical Text-Conditional Image Generation with CLIP Latents*, arXiv.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. *High-Resolution Image Synthesis with Latent Diffusion Models*, CVPR.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, Jiang Bian. 2023. *NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers*, arXiv.

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*, arXiv.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*, ICCV.

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*, arXiv.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, Ming Zhou. 2019. *Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training*, AAAI.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, Nan Duan. 2019. *M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training*, CVPR.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. *UNITER: UNiversal Image-TExt Representation Learning*, ECCV.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2022. *VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts*, arXiv.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. *BridgeTower: Building Bridges Between Encoders in Vision-Language Representation Learning*, AAAI.

Xiao Xu, Bei Li, Chenfei Wu, Shao-Yen Tseng, Anahita Bhiwandiwalla, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. *ManagerTower: Aggregating the Insights of Uni-Modal Experts for Vision-Language Representation Learning*, ACL.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2022. *Vector-quantized Image Modeling with Improved VQGAN*, arXiv.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation*, arXiv.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Gong Ming, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. 2023. *NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation*, ACL.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. *Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models*, arXiv.

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. 2023. *TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs*, arXiv.