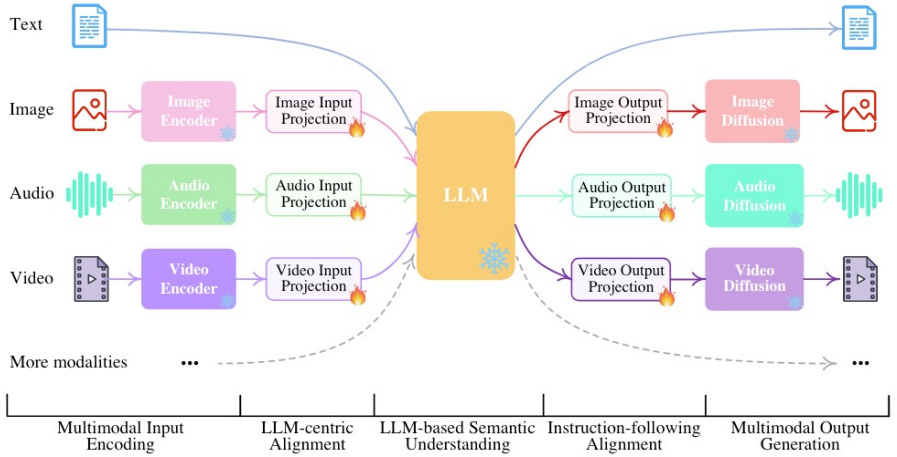
Next-GPT (NUS)

Motivation: 现有视觉语言模型大多受到多模态输入端理解的限制，无法以多种模态生成内容。 现实世界中，人类总是通过各种方式感知世界并交流，因此开发一个可以接受和生成多模态内容的模型十分重要。

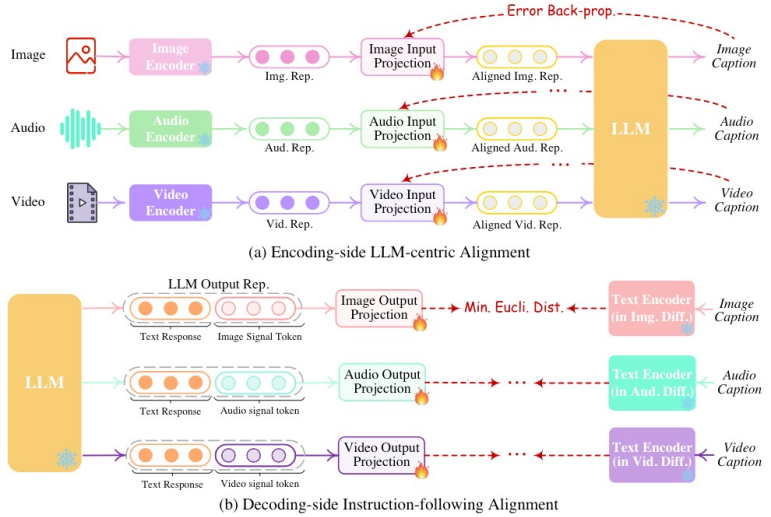
整体思路：将LLM与多模态适配器和不同的扩散解码器进行连接，让模型能够钢制输入并以文本、图像、视频和音频的任意组合生成输出。



整个模型分为三个阶段：

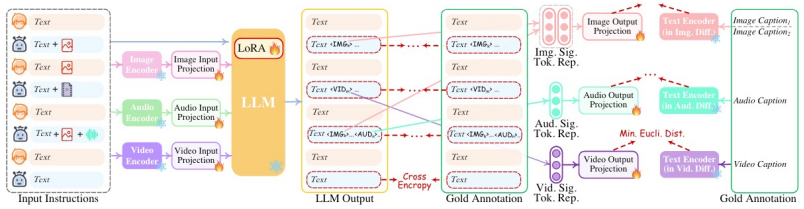
1. **多模态编码阶段**：利用现有比较完善的模型对各种模式的输入进行编码；
2. **LLM理解与推理阶段**：选择一个LLM作为核心Agent，其负责接收不同模态信息并进行理解和推理，输出分为两种(1)文本形式的回复，(2)每种模态的signal tokens用于解码；
3. **多模态生成阶段**：从 LLM（如果有）接收具有特定指令的多模态信号，基于 Transformer 的输出投影层将信号标记表示映射为后续多模态解码器可以理解的表示形式。

	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—			—	—	—	—
Image					Vicuna [12]	7B❄	Transformer	31M🔥	SD [68]	1.3B❄
Audio	ImageBind [25]	1.2B❄	Linear	4M🔥	(LoRA	33M🔥)	Transformer	31M🔥	AudioLDM [51]	975M❄
Video							Transformer	32M🔥	Zeroscope [8]	1.8B❄



为了**弥合不同模态特征空间之间的差距**，并确保不同输入的流畅语义理解，有必要进行对齐学习。

1. **编码侧**：基于现有与语料库和benchmark准备X-caption对，X表示图片、视频、音频，实现其他模态与文本模态的对齐；
2. **解码侧**：为了让LLM输出的signal token中蕴含更有效的模态信息用于扩散模型，最小化signal token与扩散模型的条件文本表示之间的距离；



编码端和解码端与LLM保持一致仍不能保证整个系统可以忠实地遵循和理解用户的指令，因此需要进行**指令微调**。使用 (输入，输出)对进行指令微调，并使用对应的Loss进行优化。

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
► *Existing data*							
MiniGPT-4 [109]	CC [10], CC3M [71]	T+I→T	Auto	✗	134M/-/-	1	5K
StableLLaVA [47]	SD [68]	T+I→T	Auto+Manu.	✗	126K/-/-	1	126K
LLaVA [104]	COCO [50]	T+I→T	Auto	✓	81K/-/-	2.29	150K
SVIT [106]	MS-COCO [50], VG [41]	T+I→T	Auto	✓	108K/-/-	5	3.2M
LLaVAR [104]	COCO [50], CC3M [71], LAION [70]	T+I→T	LLaVA+Auto	✓	20K/-/-	2.27	174K
VideoChat [44]	WebVid [5]	T+V→T	Auto	✓	-/8K/-	1.82	11K
Video-ChatGPT [54]	ActivityNet [28]	T+V→T	Inherit	✗	-/100K/-	1	100K
Video-LLaMA [103]	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto	✓	81K/8K/-	2.22	171K
InstructBLIP [15]	Multiple	T+I/V→T	Auto	✗	-	-	~ 1.6M
MIMIC-IT [42]	Multiple	T+I/V→T	Auto	✗	8.1M/502K/-	1	2.8M
PandaGPT [77]	MiniGPT-4, LLaVA	T+I→T	Inherit	✓	81K/-/-	2.29	160K
MGVLID [107]	Multiple	T+I+B→T	Auto+Manu.	✗	108K/-/-	-	108K
M^3IT [45]	Multiple	T+I/V/B→T	Auto+Manu.	✗	-/-/-	1	2.4M
LAMM [97]	Multiple	T+I+PC→T	Auto+Manu.	✓	91K/-/-	3.27	196k
BuboGPT [108]	Clotho [20], VGGSS [11]	T+A/(I+A)→T	Auto	✗	5k/-/9K	-	9K
mPLUG-DocOwl [96]	Multiple	T+I/Tab/Web→T	Inherit	✗	-	-	-
► *In this work*							
T2M	Webvid [5], CC3M [71], AudioCap [38]	T→T+I/A/V	Auto	✗	4.9K/4.9K/4.9K	1	14.7K
MosIT	Youtube, Google, Flickr, Midjourney, etc.	T+I+A+V→T+I+A+V	Auto+Manu.	✓	4K/4K/4K	4.8	5K

Table 2: Summary and comparison of existing datasets for multimodal instruction tuning. T: text, I: image, V: video, A: audio, B: bounding box, PC: point cloud, Tab: table, Web: web page.

数据集分为两种：Text+X->Text 和 Text->Text+X. 现有数据集不满足全模态生成的模式，因此该论文又自己构建了一个数据集**MosIT**(5k对话).

Method	FID (↓)
CogVideo [17]	27.10
GLIDE [58]	12.24
CoDi [78]	11.26
SD [68]	**11.21**
NExT-GPT	11.28

Text-to-Image

Method	FD (↓)	IS (↑)
DiffSound [95]	47.68	4.01
AudioLDM-S [51]	29.48	6.90
AudioLDM-L [51]	23.31	8.13
CoDi [78]	**22.90**	**8.77**
NExT-GPT	23.58	8.35

Text-to-Audio

Method	FID (↓)	CLIPSIM (↑)
CogVideo [30]	23.59	0.2631
MakeVideo [74]	13.17	0.3049
Latent-VDM [68]	14.25	0.2756
Latent-Shift [2]	15.23	0.2773
CoDi [78]	—	0.2890
NExT-GPT	**13.04**	**0.3085**

Text-to-Video

Method	B@4	METEOR	CIDEr
Oscar [46]	36.58	30.4	124.12
BLIP-2 [43]	43.7	—	145.8
OFA [86]	44.9	32.5	154.9
CoDi [78]	40.2	31.0	149.9
NExT-GPT	44.3	**32.9**	**156.7**

Image-to-Text

Method	SPIDEr	CIDEr
AudioCaps [38]	0.369	0.593
BART [26]	0.465	0.753
AL-MixGen [39]	0.466	0.755
CoDi [78]	0.480	0.789
NExT-GPT	**0.521**	**0.802**

Audio-to-Text

Method	B@4	METEOR
ORG-TRL [105]	43.6	28.8
GIT [85]	54.8	33.1
mPLUG-2 [91]	57.8	34.9
CoDi [78]	52.1	32.5
NExT-GPT	**58.4**	**38.5**

Video-to-Text

Method	Object		Background	
	CLIP (↑)	FID (↓)	CLIP (↑)	FID (↓)
PTP [29]	30.33	9.58	31.55	13.92
BLDM [4]	29.95	6.14	30.38	20.44
DiffEdit [14]	29.30	**3.78**	26.92	**1.74**
PFB-Diff [36]	**30.81**	5.93	**32.25**	13.77
NExT-GPT	29.31	6.52	27.29	15.20

Text+Image-to-Image

Method	MCD (↓)
CampNet [87]	0.380
MakeAudio [33]	0.375
AudioLDM-L [51]	0.349
NExT-GPT	**0.302**

Text+Audio-to-Audio

Method	CLIP-T (↑)	CLIP-I (↑)
CogVideo [30]	0.2391	0.9064
TuneVideo [89]	0.2758	0.9240
SDEdit [55]	0.2775	0.8731
Pix2Video [9]	**0.2891**	**0.9767**
NExT-GPT	0.2683	0.9645

Text+Video-to-Video

在不同模态设置下进行了详细的实验验证，通过与各个模态设置的基线方法进行对比可知，该方法取得了较优的实验结果，证明了方案的可行性。实验分析较少。

http://arxiv.org/abs/2309.05519