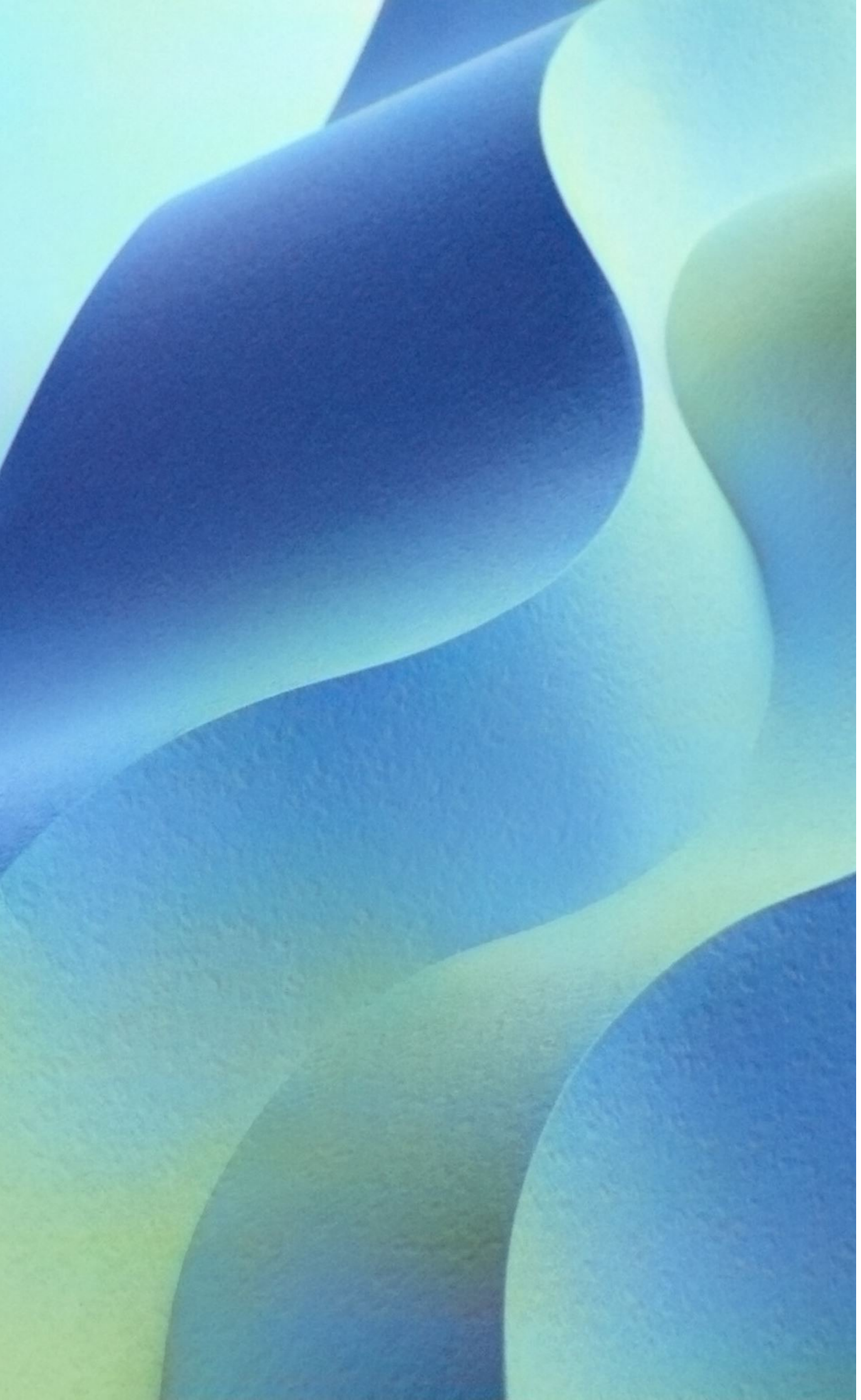
Enhancing Multimodal Retrieval and Generation with Unified Vision-Language Models

WU, Xiao-Ming

Associate Professor

Department of Data Science & AI

The Hong Kong Polytechnic University

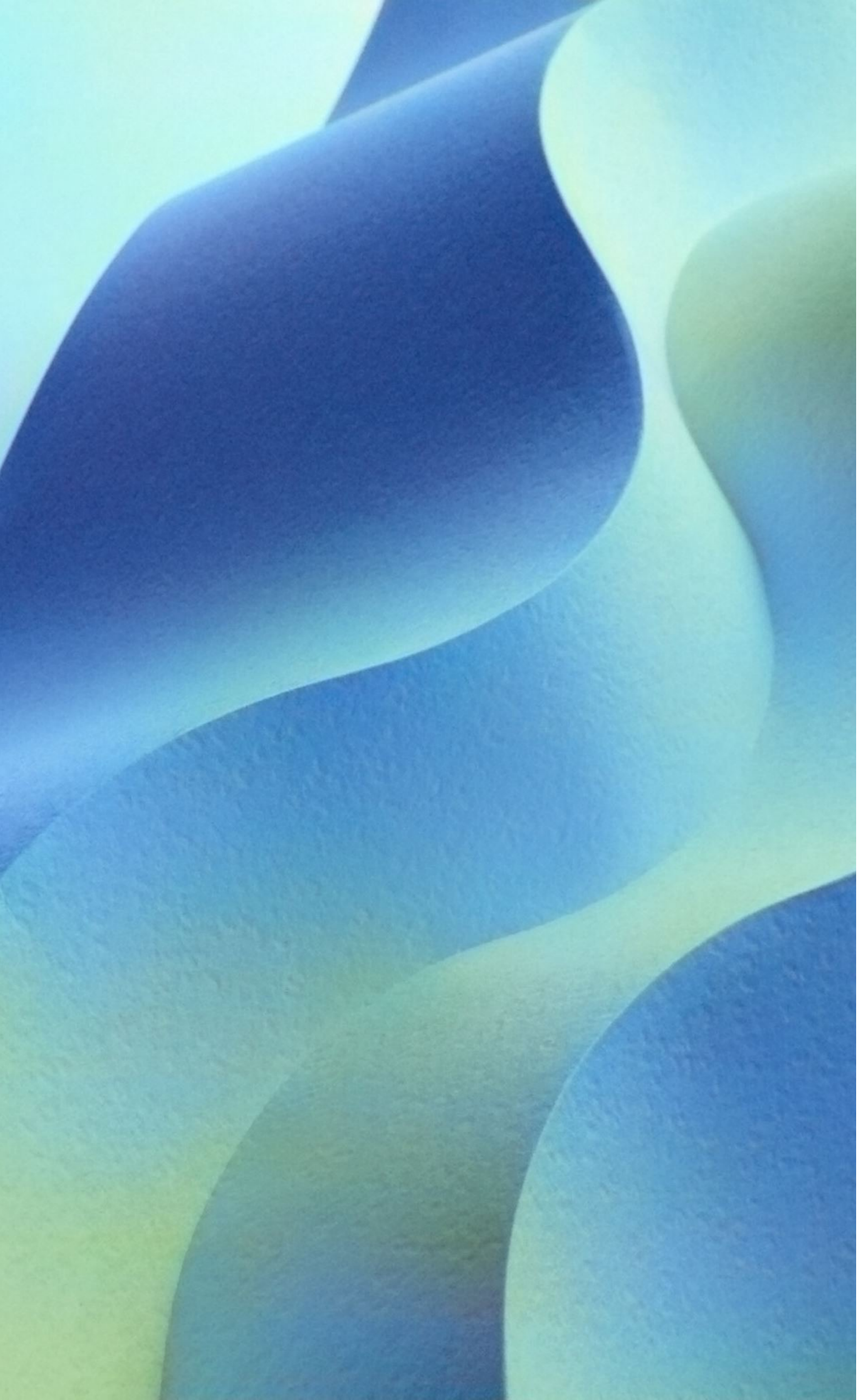


Outline

Multimodal Retrieval & Generation Tasks

Vision-language Models for Multimodal Retrieval & Generation

Unified Vision-language Models for Fashion Retrieval & Generation



Outline

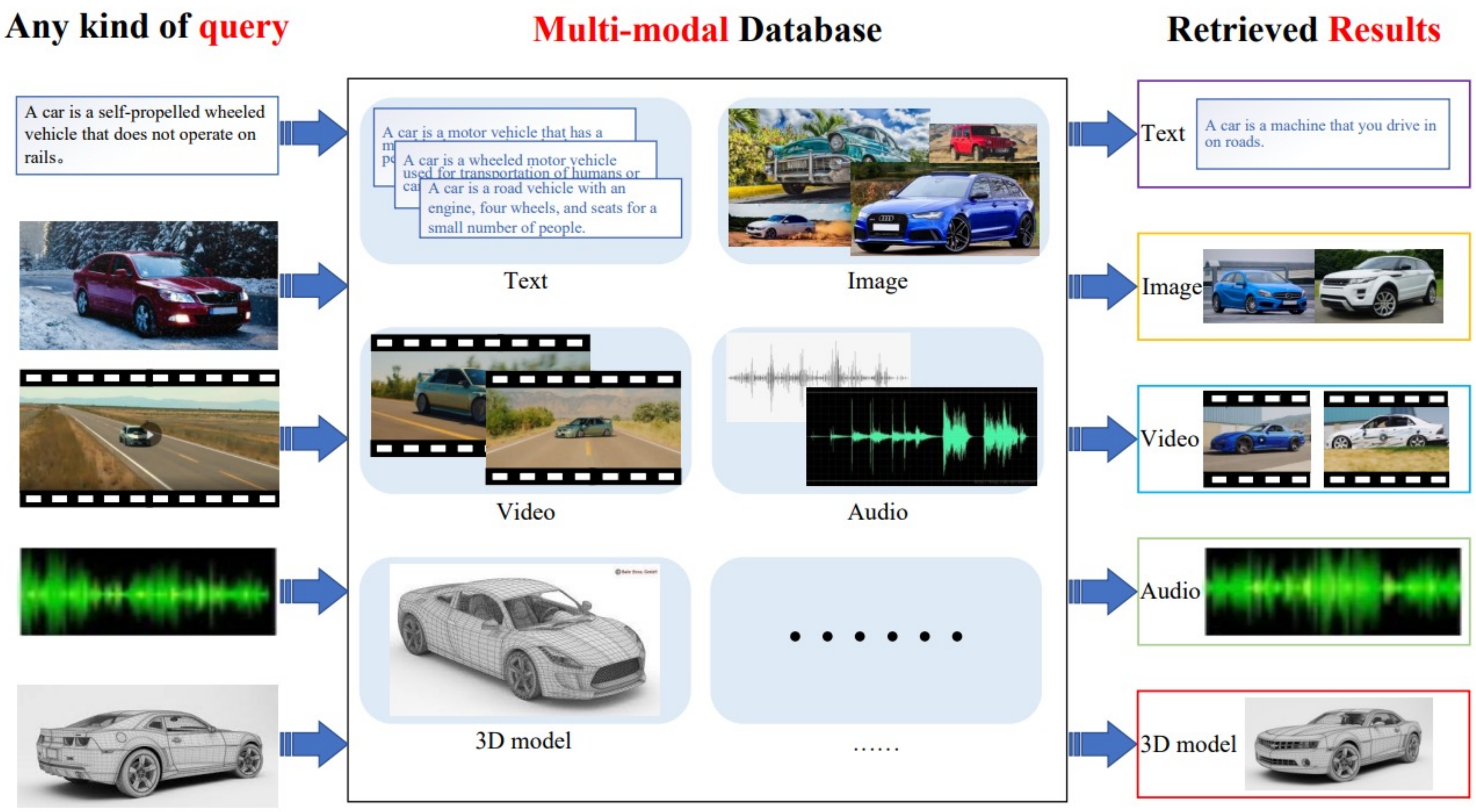
Multimodal Retrieval & Generation Tasks

Vision-language Models for Multimodal Retrieval & Generation

Unified Vision-language Models for Fashion Retrieval & Generation

Cross-Modal Retrieval

- Cross-modal retrieval aims at retrieving relevant items that are of different nature w.r.t. the query format.
- For instance, users might input a text query and retrieve images or videos related to that query.



Zhu L, Wang T, Li F, et al. Cross-modal retrieval: a systematic review of methods and future directions[J]. arXiv preprint arXiv:2308.14263, 2023.

Composed Image Retrieval

- Users can refine product searches by providing details and constraints in natural language.
- The system utilizes both visual and textual features to retrieve the desired results.

Zhu L, Wang T, Li F, et al. Cross-modal retrieval: a systematic review of methods and future directions[J]. arXiv preprint arXiv:2308.14263, 2023.

Multimodal Generation

- **Image-to-Text generation** include visual conversation, visual knowledge reasoning, visual commonsense reasoning, storytelling, personalized image-to-text generation, etc.

Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." International conference on machine learning. PMLR, 2023.

Multimodal Generation

- **Conditional Image generation** include text-to-image generation, try-on, spatial control, etc.

Text-to-image generation **Input**: Text

"An oil painting of a space shuttle"

Input Canny edge | Default | "masterpiece of fairy tale, giant deer, golden antlers"

Conditioned text-to-image generation **Input**: Canny edge, text

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." ICCV 2023.

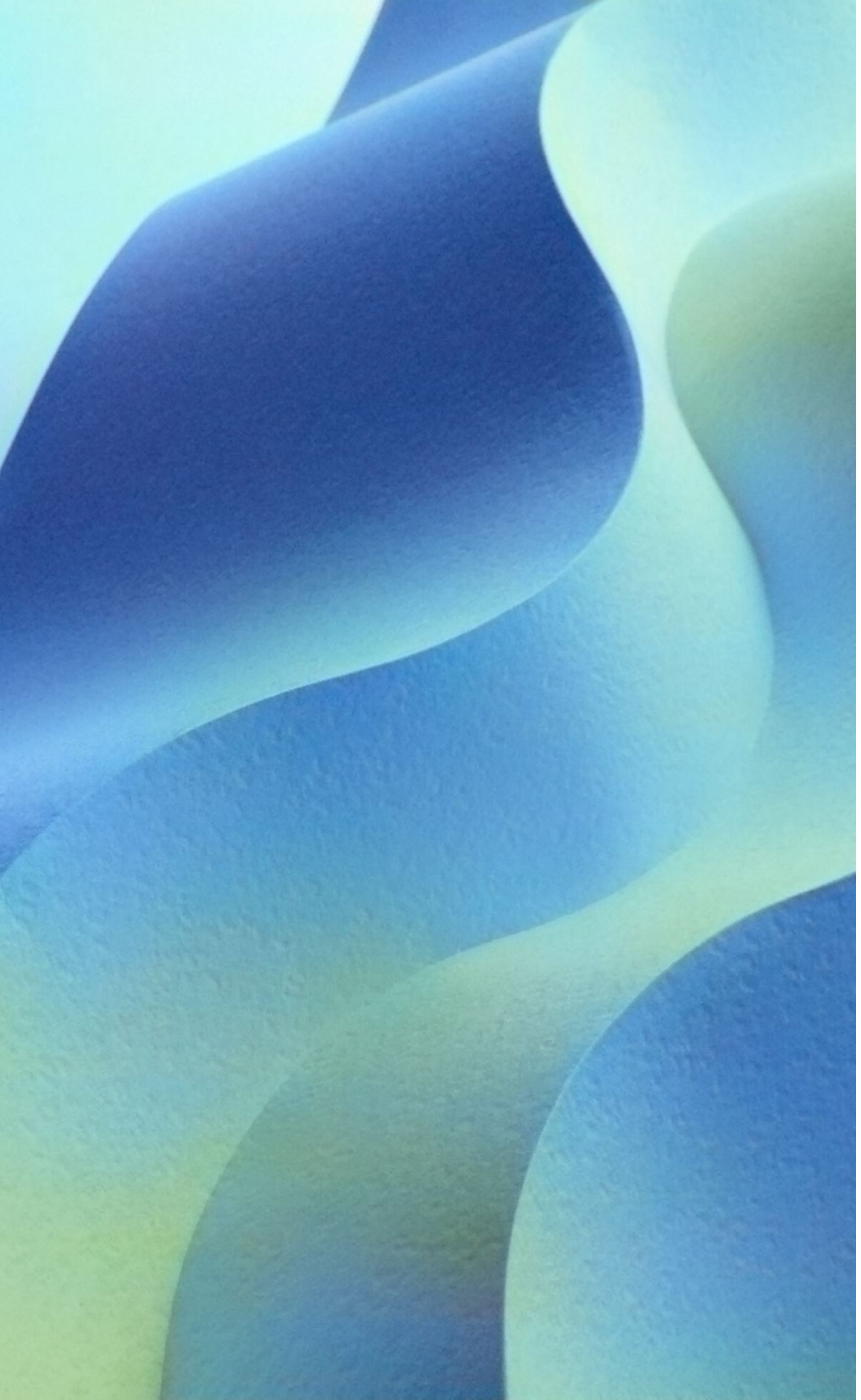
Multimodal Generation

- **Conditional Image generation** include text-to-image generation, try-on, spatial control, etc.

Try-on task in fashion domain **Input**: clothing image, person image

Human-centric fashion images design **Input**: text, human body poses, and garment sketches

Baldrati, Alberto, et al. "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing." ICCV 2023.
Kim, Jeongho, et al. "Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on." CVPR 2024.



Outline

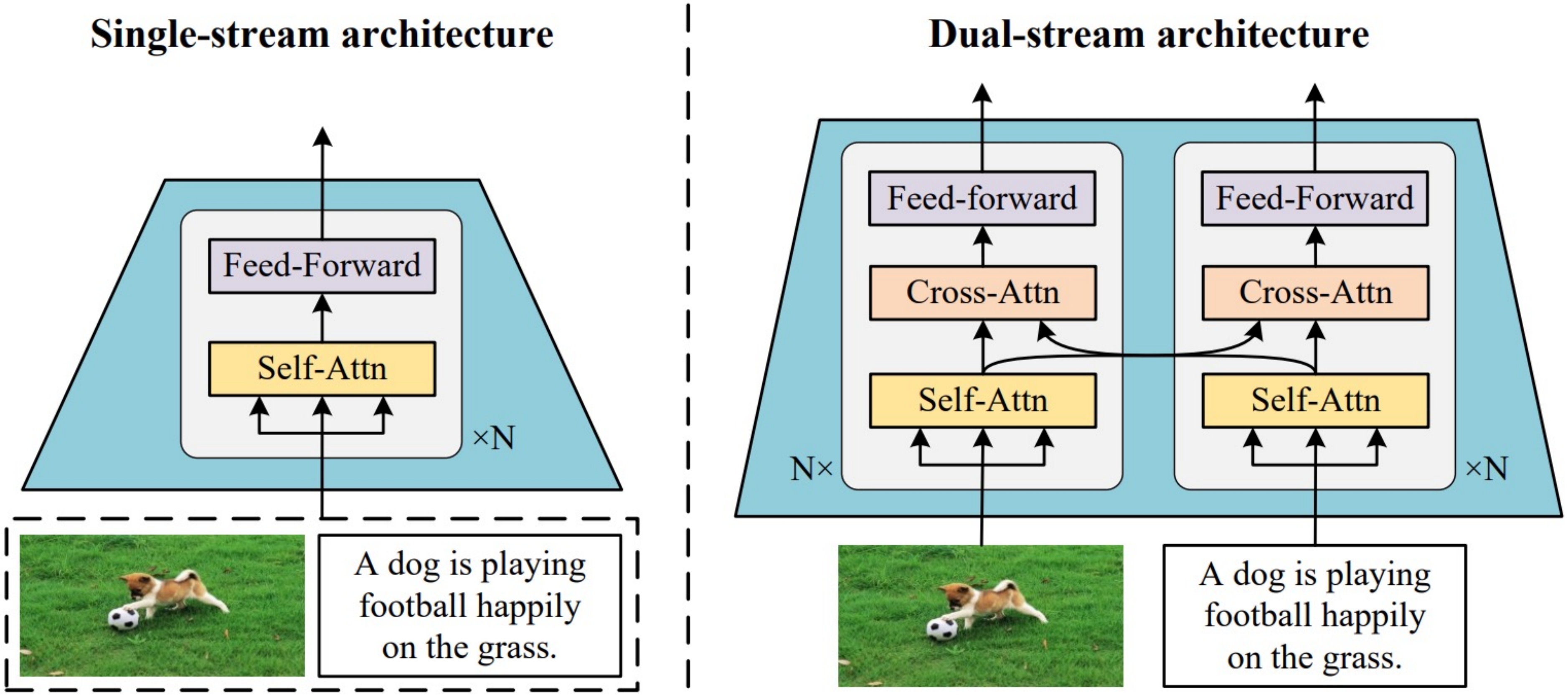
Multimodal Retrieval & Generation Tasks

Vision-language Models for Multimodal Retrieval & Generation

Unified Vision-language Models for Fashion Retrieval & Generation

Cross-modal Retrieval Models

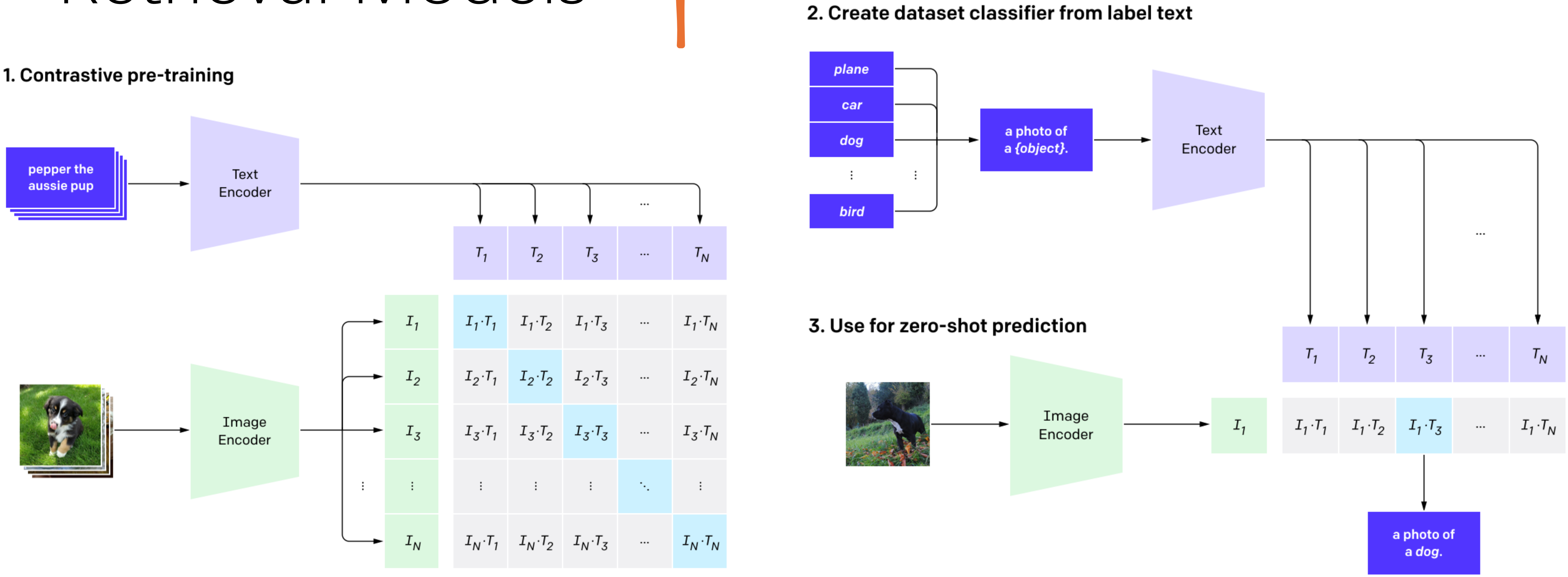
- Recently, researchers have leveraged the powerful representational capabilities of VLP models to significantly enhance cross-modal retrieval performance. VLP models include both single-stream and dual-stream architectures.



Zhu, Lei, et al. "Cross-modal retrieval: a systematic review of methods and future directions." arXiv preprint arXiv:2308.14263 (2023).

Cross-modal Retrieval Models

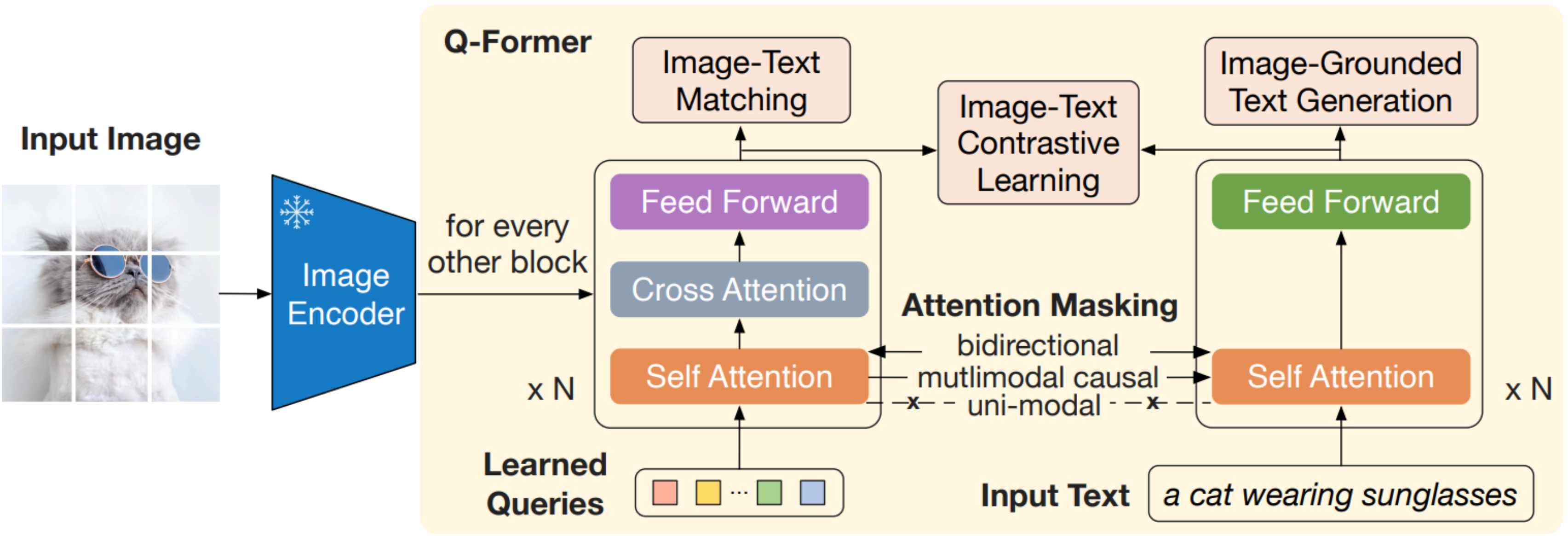
- **Dual-stream architecture: CLIP**



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *ICML* 2021.

Cross-modal Retrieval Models

- **Dual-stream architecture: BLIP-2**
- Q-Former jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text. So, it has both retrieval and generation abilities.

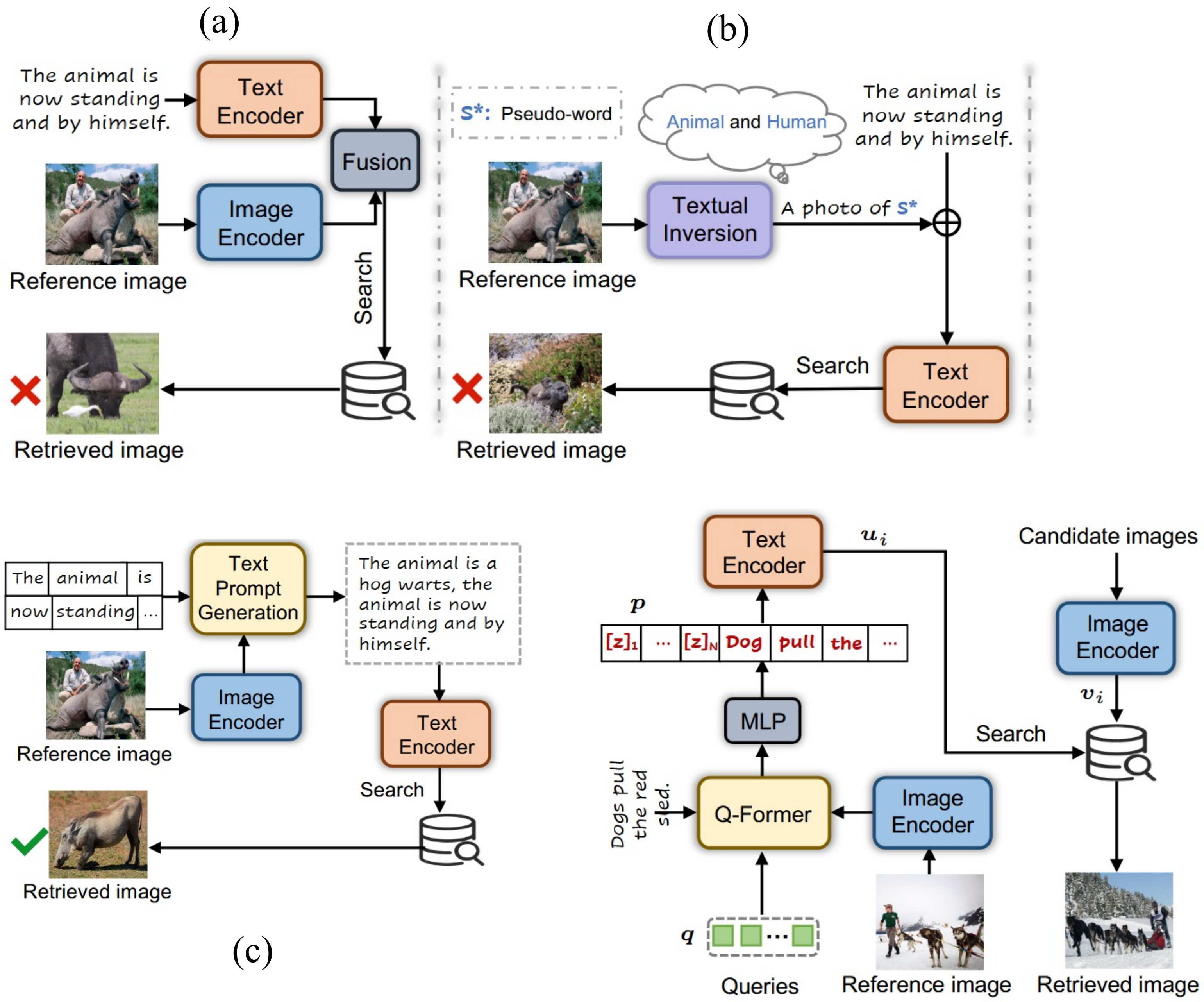


Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	**100.0**	81.5	95.6	97.8	84.8	96.5	98.3	67.2	**87.7**	**92.8**
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder reranking													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	**100.0**	**100.0**	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	**100.0**	**100.0**	88.6	97.6	**98.9**	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-g	1.2B	**97.6**	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**	**85.4**	**97.0**	**98.5**	**68.3**	**87.7**	92.6

Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML, 2023.

Composed Image Retrieval Models

- (a) Late fusion, (b) pseudo-word embedding, and (c) prompt-based method. Late fusion and pseudo-word embedding are limited in handling the cases where multiple objects are involved in the reference image and complex changes, e.g., object removal or attribute modification, are included in the relative caption.

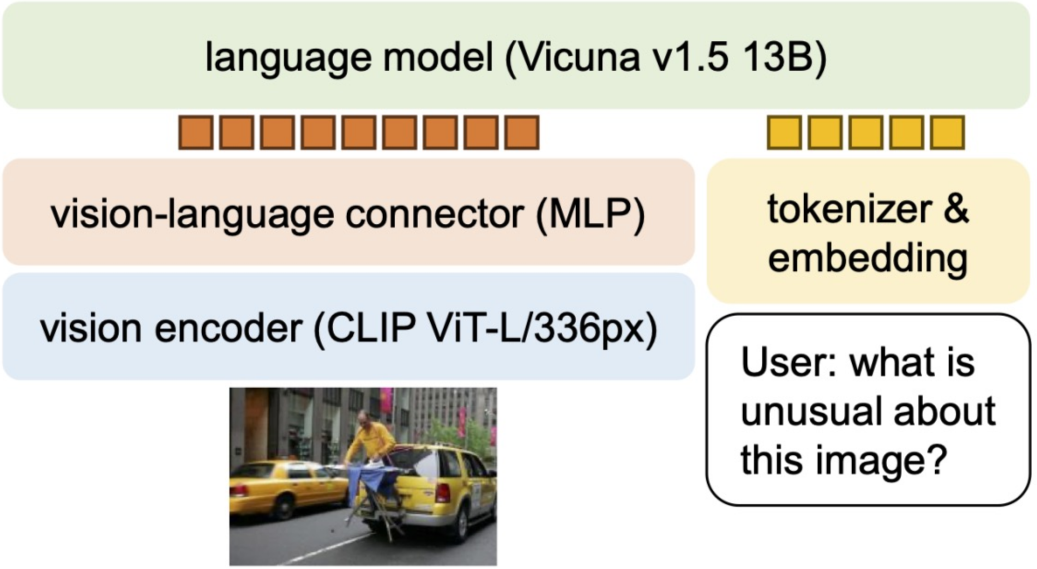


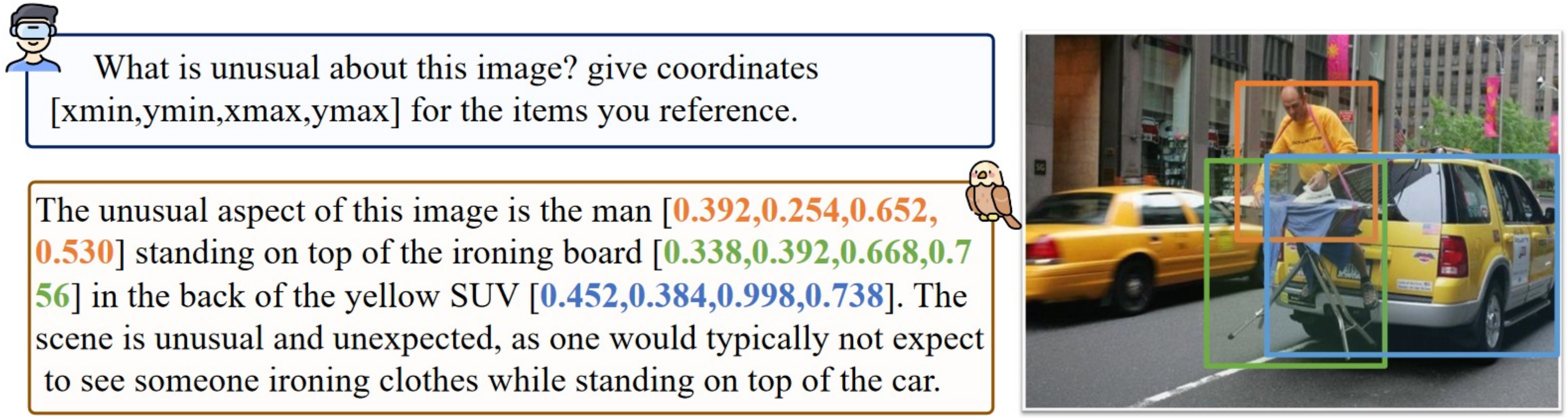
Bai, Yang, et al. "Sentence-level prompts benefit composed image retrieval." ICLR 2024.

MLLM for Generation Tasks

- **LLaVA:** Visual instruction tuning LLMs

Data	Size	Response formatting prompts
LLaVA [36]	158K	–
ShareGPT [46]	40K	–
VQAv2 [19]	83K	Answer the question using a single word or phrase.
GQA [21]	72K	
OKVQA [41]	9K	
OCRVQA [42]	80K	
A-OKVQA [45]	66K	Answer with the option's letter from the given choices directly.
TextCaps [47]	22K	Provide a one-sentence caption for the provided image.
RefCOCO [24, 40]	48K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG [25]	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

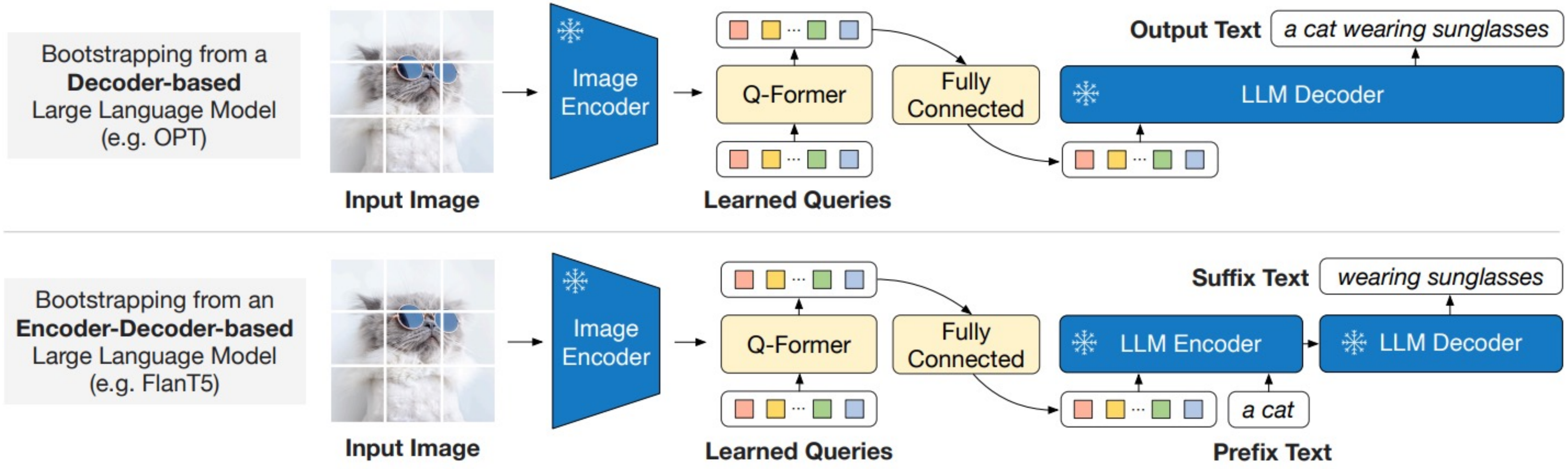




Liu, Haotian, et al. "Improved baselines with visual instruction tuning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

MLLM for Generation Tasks

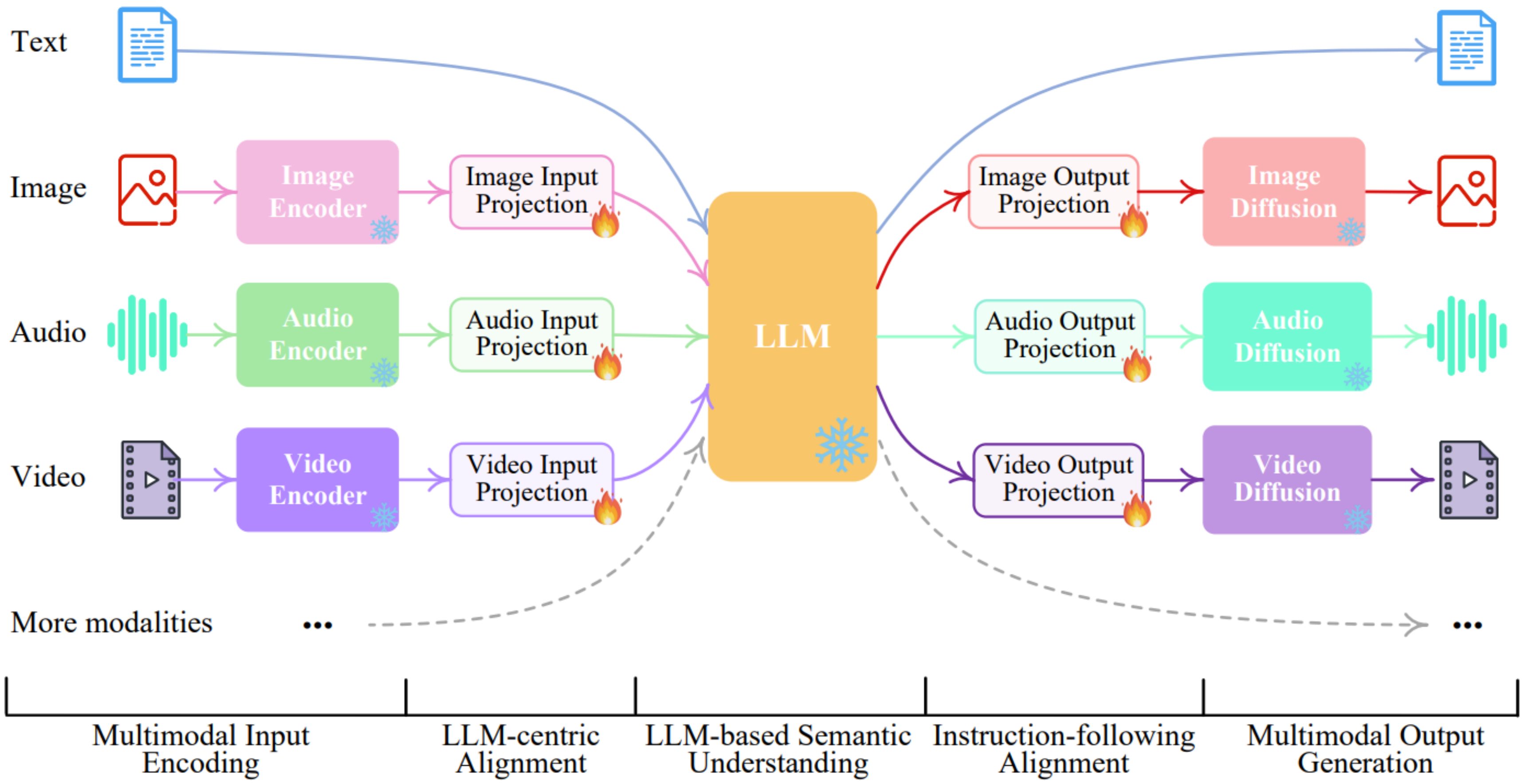
- **BLIP2** connect Q-Former (with the frozen image encoder attached) to a frozen LLM to harvest the LLM's generative language capability.



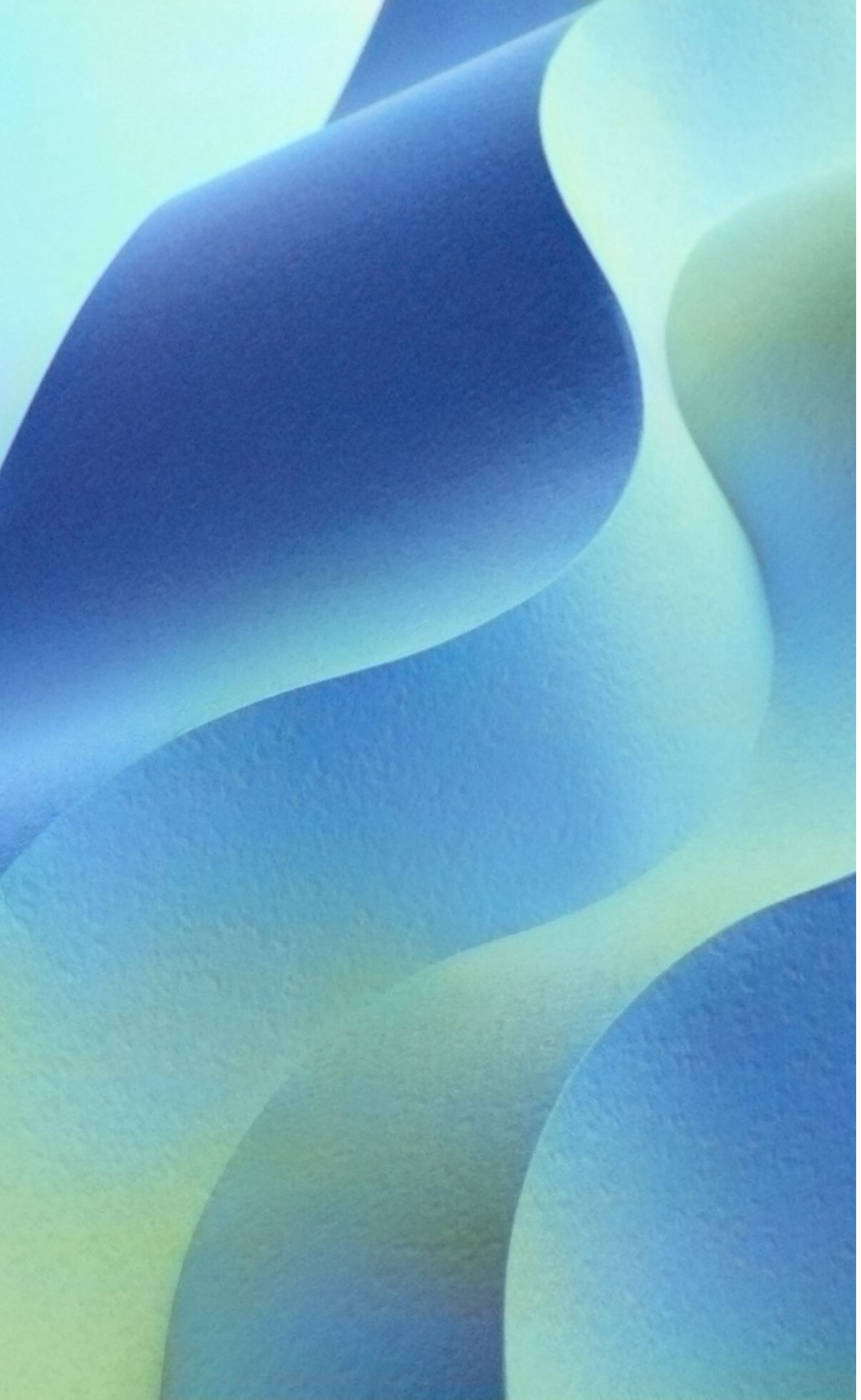
Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

MLLM for Generation Tasks

- **Any-to-any LLM:** NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output by connecting LLM with multimodal adaptors and diffusion decoders.



Wu, Shengqiong, et al. "Next-gpt: Any-to-any multimodal llm." arXiv preprint arXiv:2309.05519 (2023).



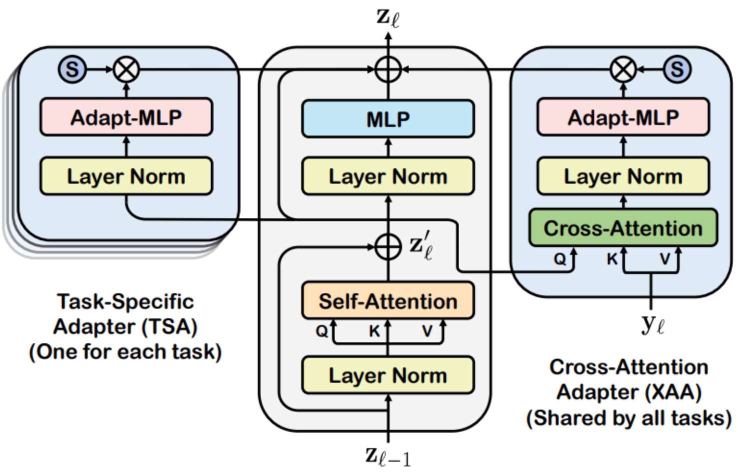
Outline

Multimodal Retrieval & Generation Tasks

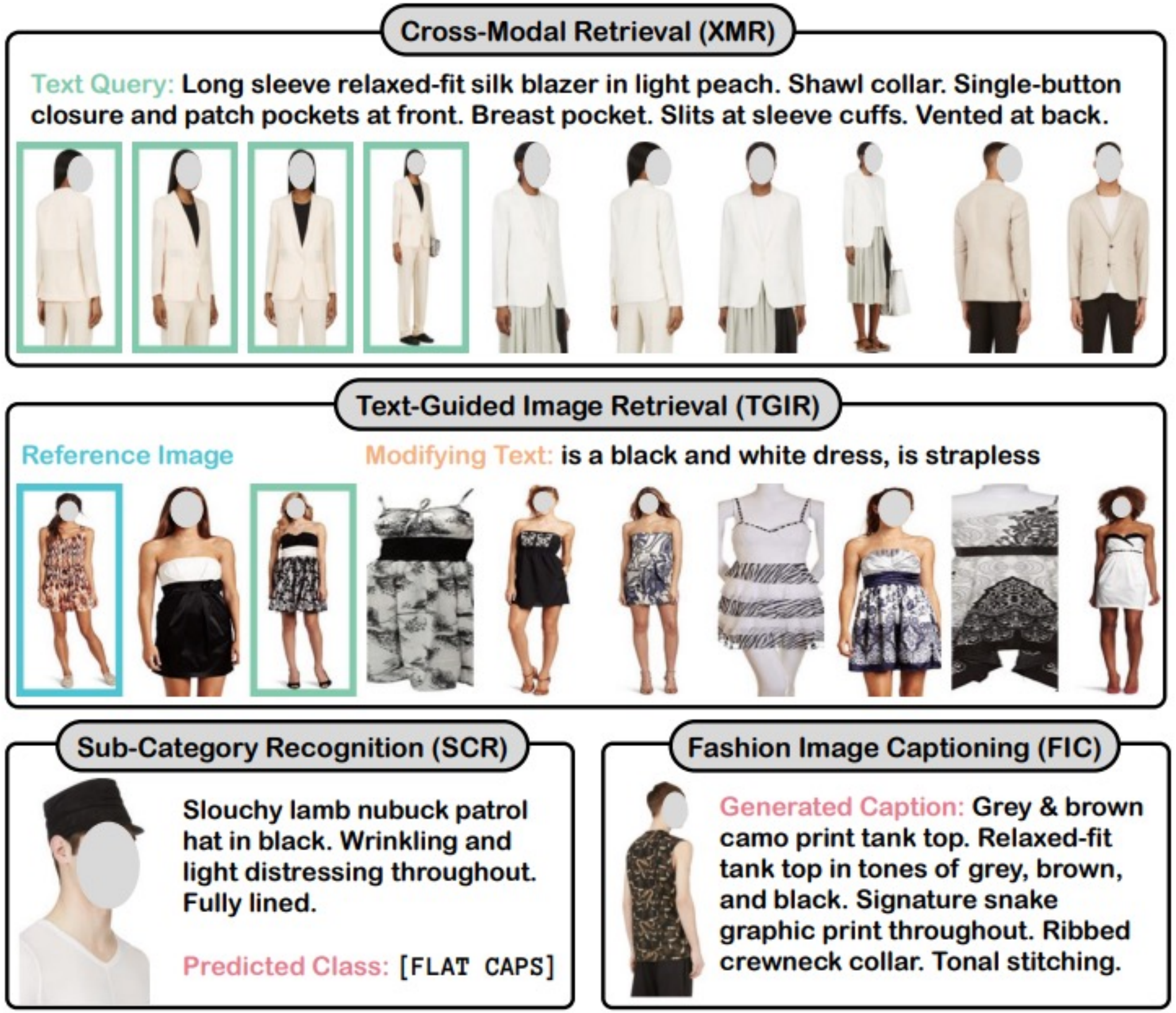
Vision-language Models for Multimodal Retrieval & Generation

Unified Vision-language Models for Fashion Retrieval & Generation

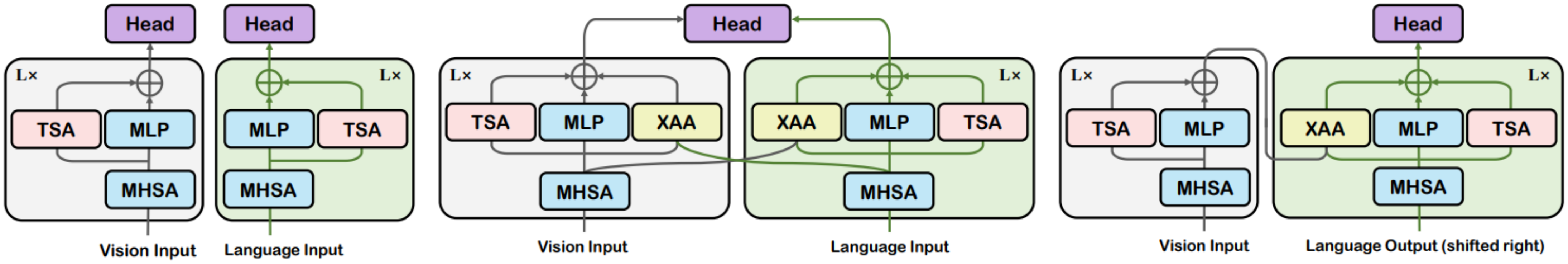
Unified VL Models for Fashion Tasks



Task-versatile Transformer layer equipped with two adapters: cross-attention adapter (XAA) and task-specific adapter (TSA).



Han, Xiao, et al. "Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks." CVPR 2023.



(a) Contrastive Mode (b) Fusion Mode (c) Generative Mode

Unified VL Models for Fashion Tasks

- Contrastive mode supports Cross-Modal Retrieval tasks.
- Fusion mode: Both XAA and TSA modules are enabled in this mode. Given an input image-text pair, a fusion encoder producing two cross-modal attended representations for the Composed Image Retrieval task.
- Generative mode works as a seq2seq model performing the generative tasks auto-regressively, e.g., Fashion Image Captioning.

Han, Xiao, et al. "Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks." CVPR 2023.

Motivation

- Previous works have not thoroughly explored multimodal generation and retrieval tasks within a unified model.
- Investigating task correlations and integrating retrieval tasks with generation tasks is both necessary and promising.

XMR: Cross-modal retrieval tasks; **CIR**: Composed image retrieval task.

Model Types	Task Domain	Model	Main Structure	XMR	CIR	Text Generation	Image Generation
Cross-modal Retrieval	General	CLIP (2021)	Dual-stream Transfomer	✔	✗	✗	✗
	Fashion	FashionBERT (2020)	Single-stream Transfomer	✔	✗	✗	✗
Multimodal LLM	General	LLaVA (2023)	CLIP, LLM	✗	✗	✔	✗
Composed Image Retrieval	General	SPRC (2024)	CLIP,Qformer	✗	✔	✗	✗
Conditional Diffusion	General	ControlNet (2023)	Stable diffusion	✗	✗	✗	✔
	Fashion	StableVITON (2023)	Stable diffusion	✗	✗	✗	✔
Unified Model	General	NExT-GPT (2023)	ImageBind,LLM,Diffusion	✗	✗	✔	✔
	Fashion	FAME-ViL (2023)	Dual-stream Transfomer	✔	✔	✔	✗
	General	BLIP2 (2023)	CLIP,Qformer,LLM	✔	✗	✔	✗
Unified Model (Ours)	Fashion	UniFashion	CLIP,Qformer,LLM,Diffusion	✔	✔	✔	✔

UniFashion: A Unified Vision-Language Model for Multimodal Fashion Retrieval and Generation

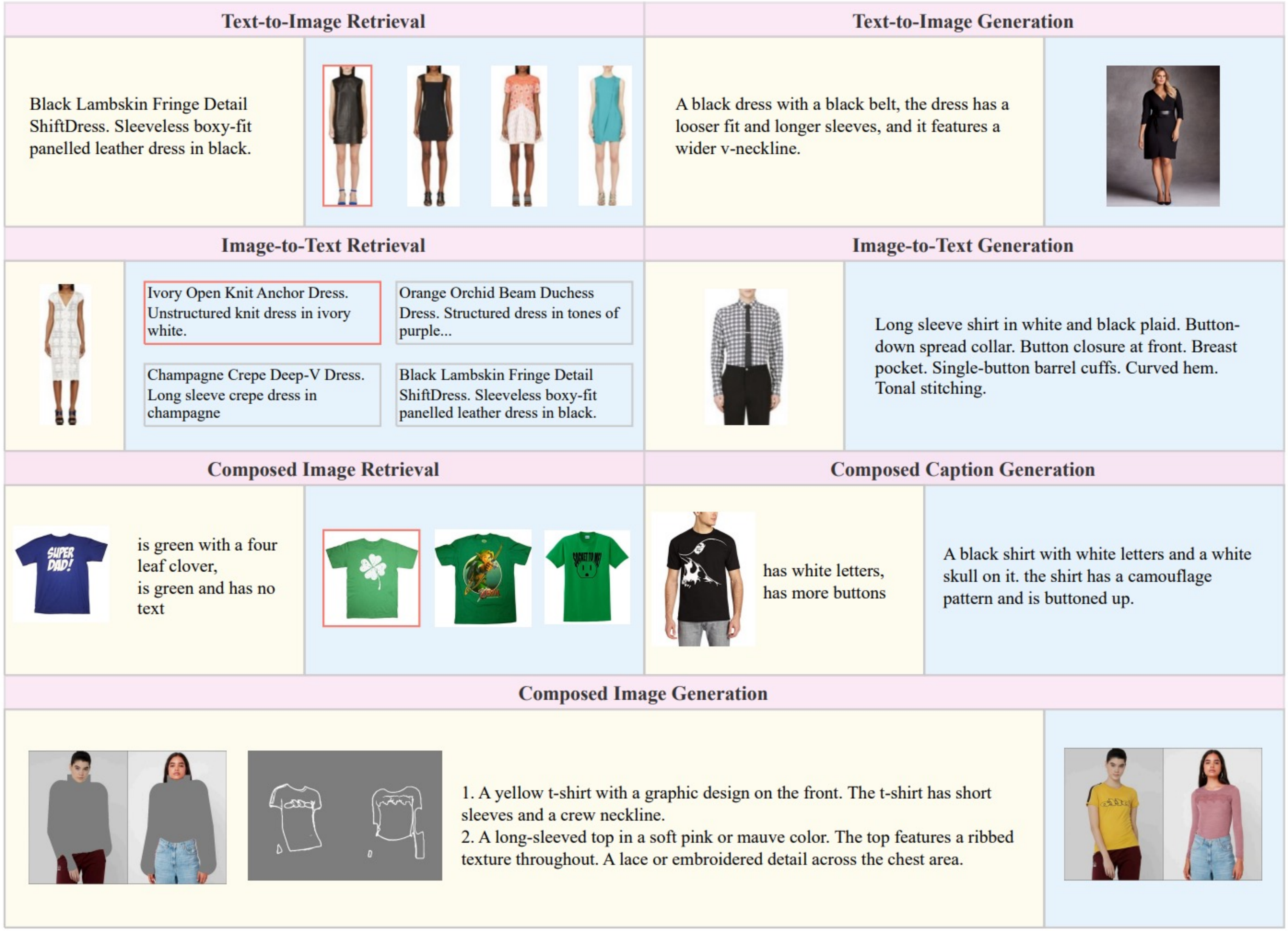
Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, Xiao-Ming Wu

https://arxiv.org/abs/2408.11305

UniFashion: A Unified VL Model for Fashion Retrieval & Generation

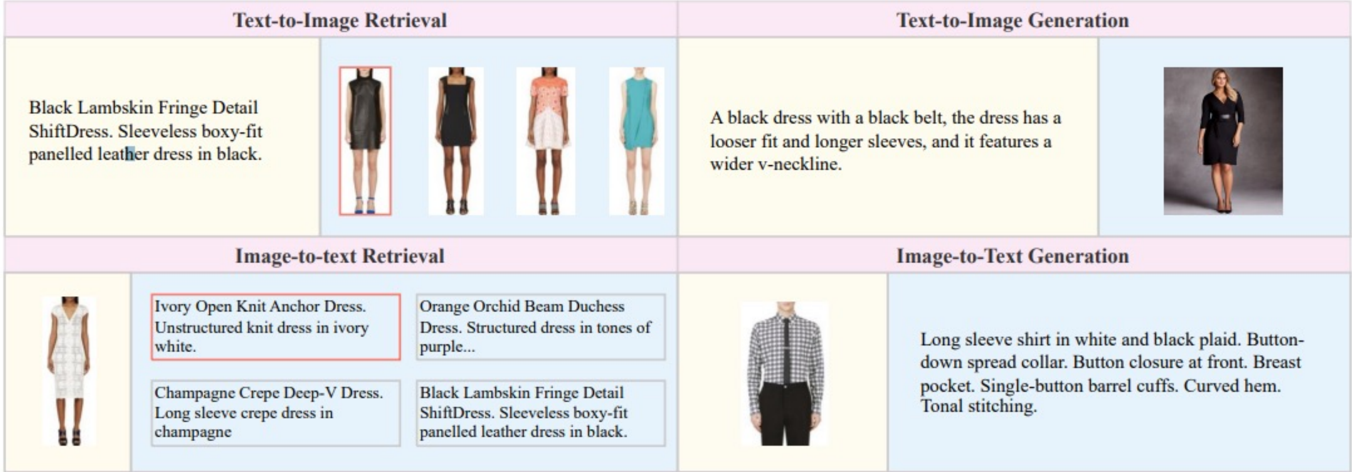
- **Encompass all multimodal retrieval & generation tasks:**

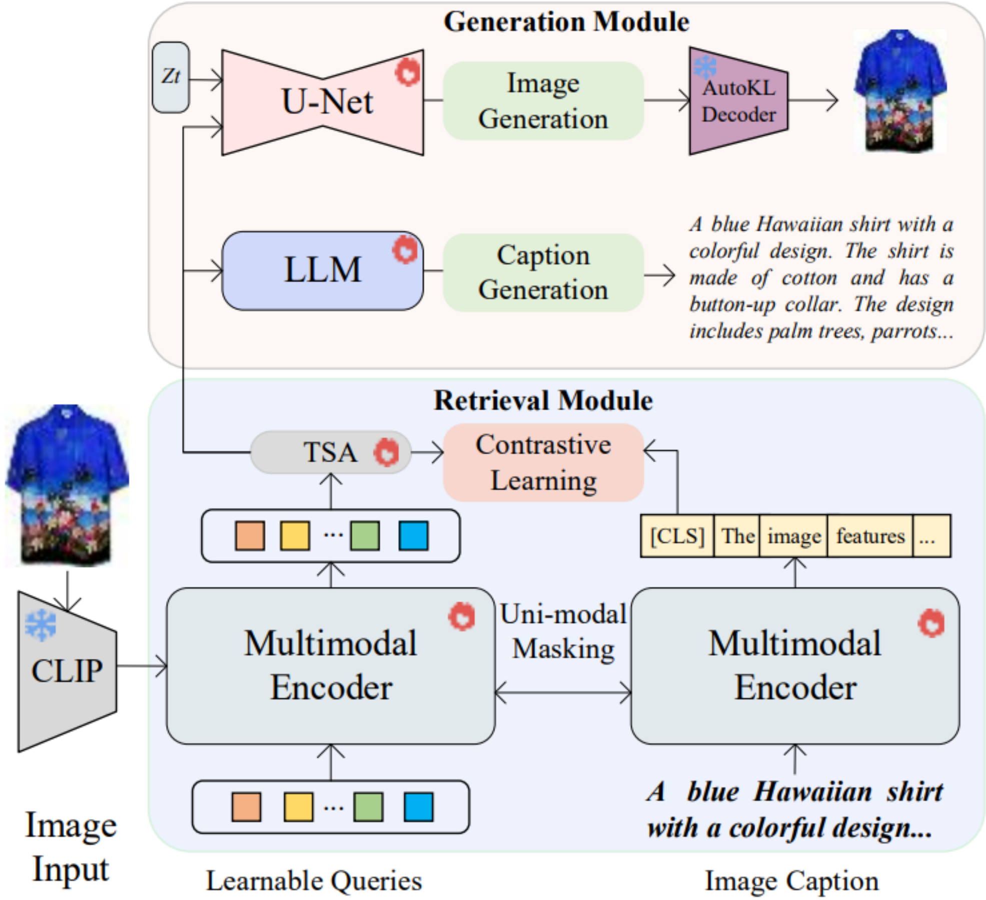
1. Cross-modal retrieval
2. Text-guided image retrieval
3. Fashion image captioning
4. Fashion image generation



Phase 1: Cross-modal Pre-training

- UniFashion acquires robust cross-modal fashion representation capabilities through pre-training, leveraging both the LLM and the diffusion model.
- Leverage Q-Former as the multimodal encoder.

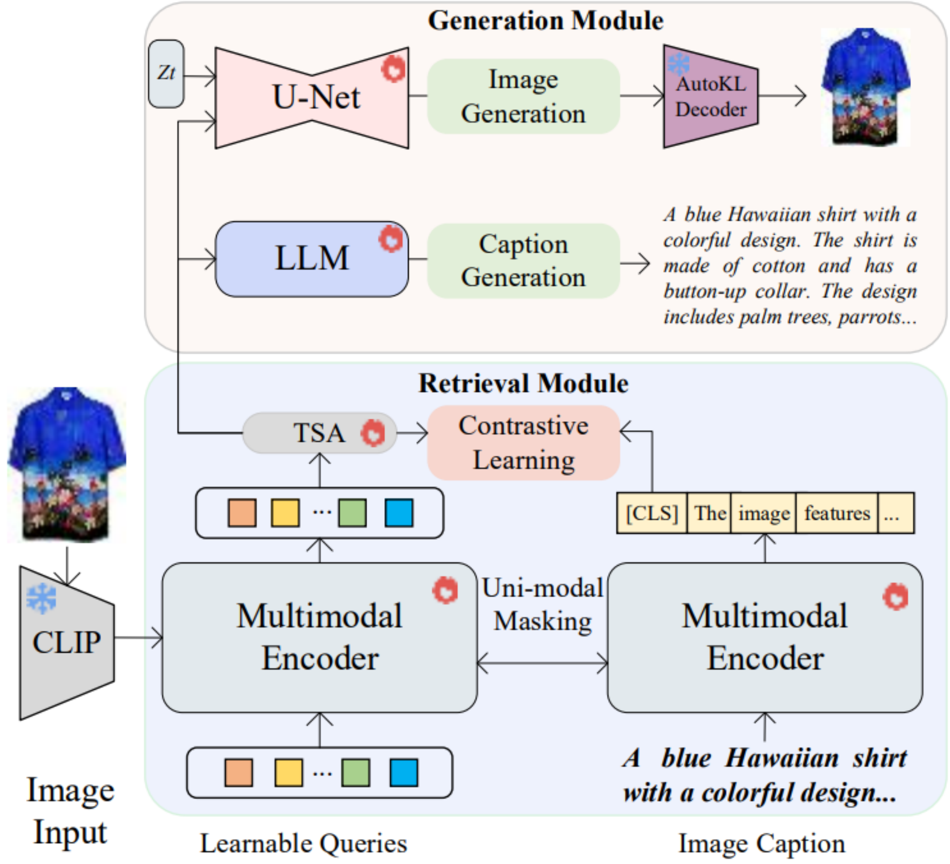




Phase 1

Phase 1: Cross-modal Pre-training

- UniFashion acquires robust cross-modal fashion representation capabilities through pre-training, leveraging both the LLM and the diffusion model.
- Leverage Q-Former as the multimodal encoder.



Phase 1

- Cross-modal Retrieval

$$\mathcal{L}_{\mathrm{ITC}}(X,Y) = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp[\lambda(X_i^T \cdot Y^i)]}{\sum_{j=1}^{B}\exp[\lambda(X_i^T \cdot Y^j)]}$$

- Cross-modal Generation
  - Target caption generation

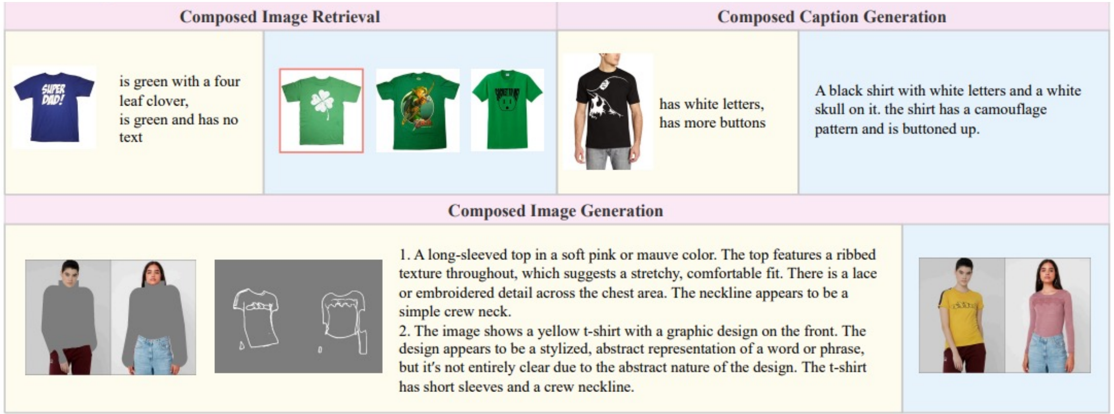
$$\mathcal{L}_{\mathrm{ITG}} = -\frac{1}{L}\sum_{l=1}^{L}\log p_\phi(w_l^g | w_{<l}^g, f_\theta(q))$$

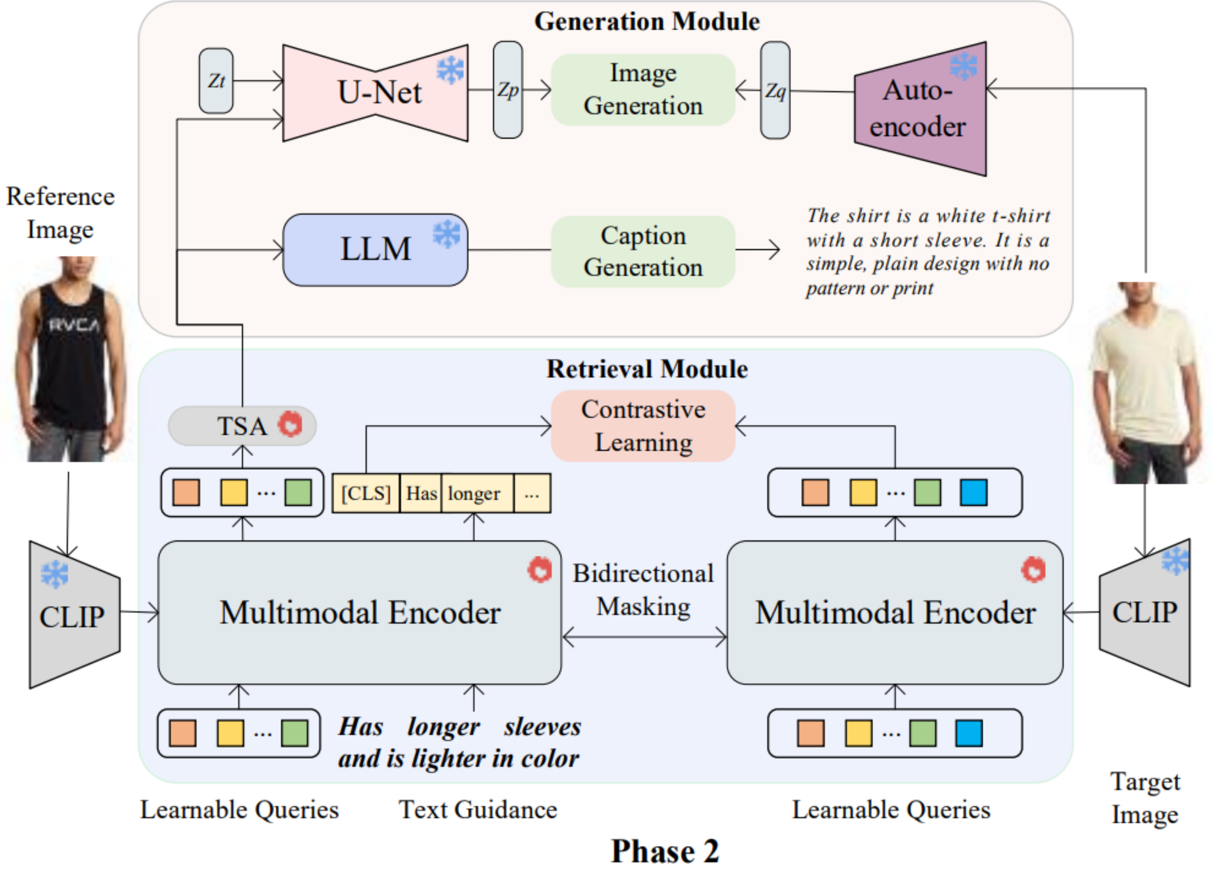
  - Target image generation

$$\mathcal{L}_{\mathrm{q2I}} = \mathbb{E}_{\boldsymbol{\epsilon}^y,\mathbf{x}_0}[\|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_\eta^x(\mathbf{x}_{t^x}, f_\zeta(q), t^x)\|^2]$$

Phase 2: Composed Multimodal Fine-tuning

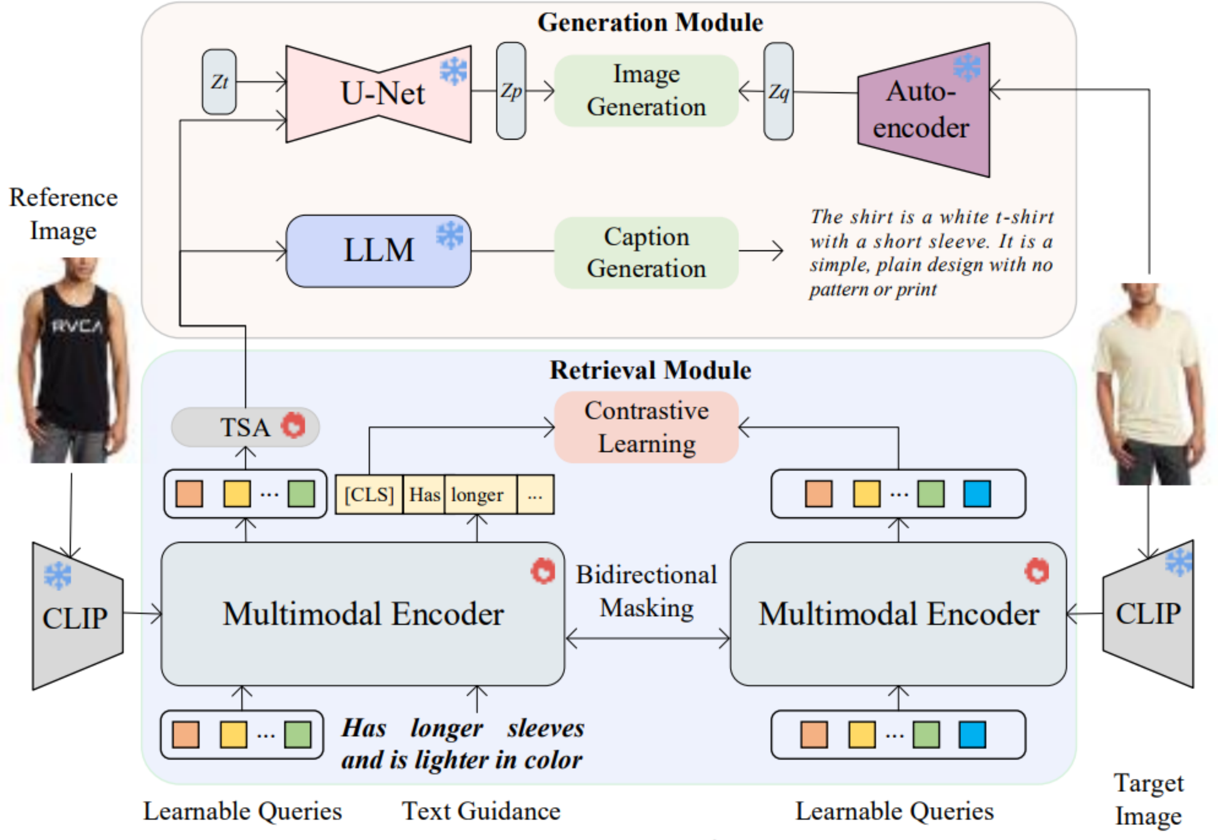
- The model undergoes fine-tuning to process both image and text inputs, refining its ability to learn composed modal representations.
- This is achieved by aligning the multimodal encoder (Q-Former) with the LLM and the diffusion model for enhanced performance.





Phase 2: Composed Multimodal Fine-tuning

- The model undergoes fine-tuning to process both image and text inputs, refining its ability to learn composed modal representations.
- This is achieved by aligning the multimodal encoder (Q-Former) with the LLM and the diffusion model for enhanced performance.



Phase 2

- Composed Image Retrieval

The output sequence of Multimodal Encoder consists of learnable queries and encoded text guidance, which includes e_{cls}, the embedding of the output of the [CLS] token. Z_T and Z_C is the target image/caption's output sequence from Multimodal Encoder:

$$\mathcal{L}_{\text{cir}} = \mathcal{L}_{\text{ITC}}(e_{cls}, Z_T) + \mathcal{L}_{\text{ITC}}(e_{cls}, Z_C)$$

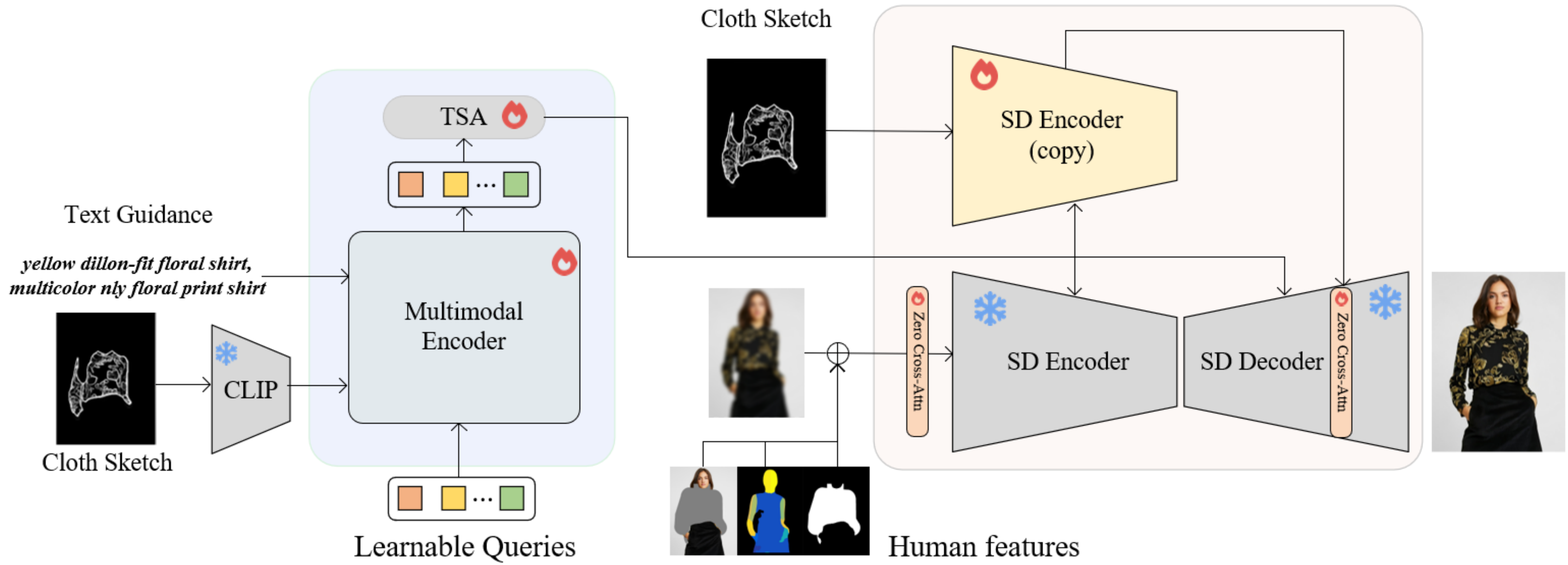
- Composed multimodal Generation

$$\mathcal{L}_{\text{ITG}} = -\frac{1}{L}\sum_{l=1}^{L} \log p_\phi(w_l^g | w_{<l}^g, f_\theta(q_R))$$

$$\mathcal{L}_{\text{q2I}} = \mathbb{E}_{\boldsymbol{\epsilon}^y, \mathbf{x}_0}[\|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_\eta^x(\mathbf{x}_{t^x}, f_\zeta(q_R), t^x)\|^2]$$

Fine-tuning for Fashion Image Editing/Try-on Tasks

- The diffusion model receives multimodal encoder's output, cloth sketch and human features as input, then generate the target images.
- We provide the cloth sketch and text guidance as a multimodal input to the encoder, such that the extracted image sketch and text features are more relevant to the ground truth.



Cross-modal Retrieval & Generation Tasks

Model	Image to Text			Text to Image			Mean
	R@1	R@5	R@10	R@1	R@5	R@10	
FashionBERT (Li et al., 2022)	23.96	46.31	52.12	26.75	46.48	55.74	41.89
OSCAR (Alayrac et al., 2022)	23.39	44.67	52.55	25.10	49.14	56.68	41.92
KaledioBERT (Li et al., 2023b)	27.99	60.09	68.37	33.88	60.60	68.59	53.25
EI-CLIP (Li et al., 2023b)	38.70	72.20	84.25	40.06	71.99	82.90	65.02
MVLT (Dai et al., 2023)	33.10	77.20	91.10	34.60	78.00	89.50	67.25
FashionViL (Zhu et al., 2023a)	65.54	91.34	96.30	61.88	87.32	93.22	82.60
FAME-ViL (Liu et al., 2023a)	65.94	91.92	97.22	62.86	87.38	93.52	83.14
UniFashion (Ours)	**71.44**	**93.79**	**97.51**	**71.41**	**93.69**	**97.47**	**87.55**

Table 3

Model	Image Captioning			
	BLEU-4	METEOR	ROUGE-L	CIDEr
FashionBERT	3.30	9.80	29.70	30.10
OSCAR	4.50	10.90	30.10	30.70
KaleidoBERT	5.70	12.80	32.90	32.60
FashionViL	16.18	25.60	37.23	39.30
FAME-ViL	30.73	25.04	55.83	150.4
UniFashion	**35.53**	**29.32**	**54.59**	**169.5**

Table 4

Table 3: Performance comparison of UniFashion and baseline models on the FashionGen dataset for cross-modal retrieval tasks.

Table 4: Image captioning task performance on the FashionGen dataset.

Composed Image Retrieval Tasks

Model	Dress		Shirt		Toptee		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Avg.
FashionVLP (Goenka et al., 2022)	32.42	60.29	31.89	58.44	38.51	68.79	34.27	62.51	48.39
CASE (Levy et al., 2023)	47.44	69.36	48.48	70.23	50.18	72.24	48.79	70.68	59.74
AMC (Zhu et al., 2023b)	31.73	59.25	30.67	59.08	36.21	66.06	32.87	61.64	47.25
CoVR-BLIP (Ventura et al., 2024)	44.55	69.03	48.43	67.42	52.60	74.31	48.53	70.25	59.39
CLIP4CIR (Baldrati et al., 2023a)	33.81	59.40	39.99	60.45	41.41	65.37	38.32	61.74	50.03
FAME-ViL (Han et al., 2023)	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	58.29
TG-CIR (Wen et al., 2023)	45.22	69.66	52.60	72.52	56.14	77.10	51.32	73.09	58.05
Re-ranking (Liu et al., 2023b)	48.14	71.43	50.15	71.25	55.23	76.80	51.17	73.13	62.15
SPRC (Bai et al., 2023)	49.18	<u>72.43</u>	55.64	73.89	<u>59.35</u>	<u>78.58</u>	54.92	<u>74.97</u>	64.85
UniFashion w/o cap	<u>49.65</u>	72.17	<u>56.88</u>	<u>74.12</u>	59.29	78.11	<u>55.27</u>	74.80	<u>65.04</u>
UniFashion w/o img	32.49	49.11	44.70	59.63	43.16	60.26	40.12	56.33	48.22
UniFashion	**53.72**	**73.66**	**61.25**	**76.67**	**61.84**	**80.46**	**58.93**	**76.93**	**67.93**

Table 5: Comparative evaluation of UniFashion and variants and baseline models on the Fashion-IQ dataset for composed image retrieval task. Best and second-best results are highlighted in bold and underlined, respectively.

Fashion Image Editing/Try-on Tasks

Model	Modalities				Metrics		
	Text	Sketch	Pose	Cloth	FID↓	KID ↓	CLIP-S
try-on task							
VITON-HD (Choi et al., 2021)	✗	✗	✓	✓	12.12	3.23	-
Paint-by-Example (Yang et al., 2023a)	✗	✗	✓	✓	11.94	3.85	-
GP-VTON (Xie et al., 2023)	✗	✗	✓	✓	13.07	4.66	-
StableVITON (Kim et al., 2024)	✗	✗	✓	✓	**8.23**	**0.49**	-
UniFashion (Ours)	✗	✗	✓	✓	<u>8.42</u>	<u>0.67</u>	-
fashion design task							
SDEdit (Meng et al., 2021)	✓	✓	✓	✗	15.12	5.67	28.61
MGD (Baldrati et al., 2023b)	✓	✓	✓	✗	<u>12.81</u>	<u>3.86</u>	<u>30.75</u>
UniFashion (Ours)	✓	✓	✓	✗	**12.43**	**3.74**	**31.29**

Table 6: Performance analysis of unpaired settings on VITON-HD and MGD datasets across different input modalities.

Findings

- UniFashion highlights the benefits of exploiting inter-task relatedness to improve overall performance. For example, the caption generation task enhances the performance of the image retrieval task.

- UniFashion extends the capability to address multimodal problems. For example, in the composed image retrieval task, the generative ability of our model enables the use of the pre-generated captions to enhance the performance.

- UniFashion integrates multiple complex modules (such as Q-Former, LLM, and diffusion models), potentially increasing computational complexity.

Thank You！ Have a nice day!

Additional Slides

Findings

- Unified model could enhance multimodal retrieval task by using more loss functions. Generally, the *dual-stream model* is trained with the contrastive learning loss. For example, CLIP. By combining the loss of the generative model, that is, aligning the embedding after image encoding to LLM to generate captions, better embeddings can be trained.

- Unified model could complete the multimodal composed tasks in more aspects. By introducing LLM, different modalities can be aligned in the form of text. That is, when Unifashion is doing the CIR task, it will generate the caption of the target image according to the reference image and the guiding text, so that retrieval can be carried out through the generated caption.

Training dataset

Description of datasets used in two stages:

Data types	Dataset	Size	Stage 1	Stage 2	Metrics
CMR	FashionGen (Lin et al., 2014) Fashion200K (Krishna et al., 2017)	260.5K 172K	✔ ✔	✔ ✗	R@K -
CIR	Fashion-IQ (Liu et al., 2023a)	18K	✗	✔	R@K
FIC	FashionGen (Liu et al., 2023a) Fashion-IQ-Cap	260.5K 60K	✔ ✔	✔ ✗	BLEU,CIDEr,METEOR,ROUGE-L -
FIG	VITON-HD (Goyal et al., 2017) MGD (Schwenk et al., 2022)	83K 66K	✗ ✗	✔ ✔	FID, KID FID,KID,CLIP-S

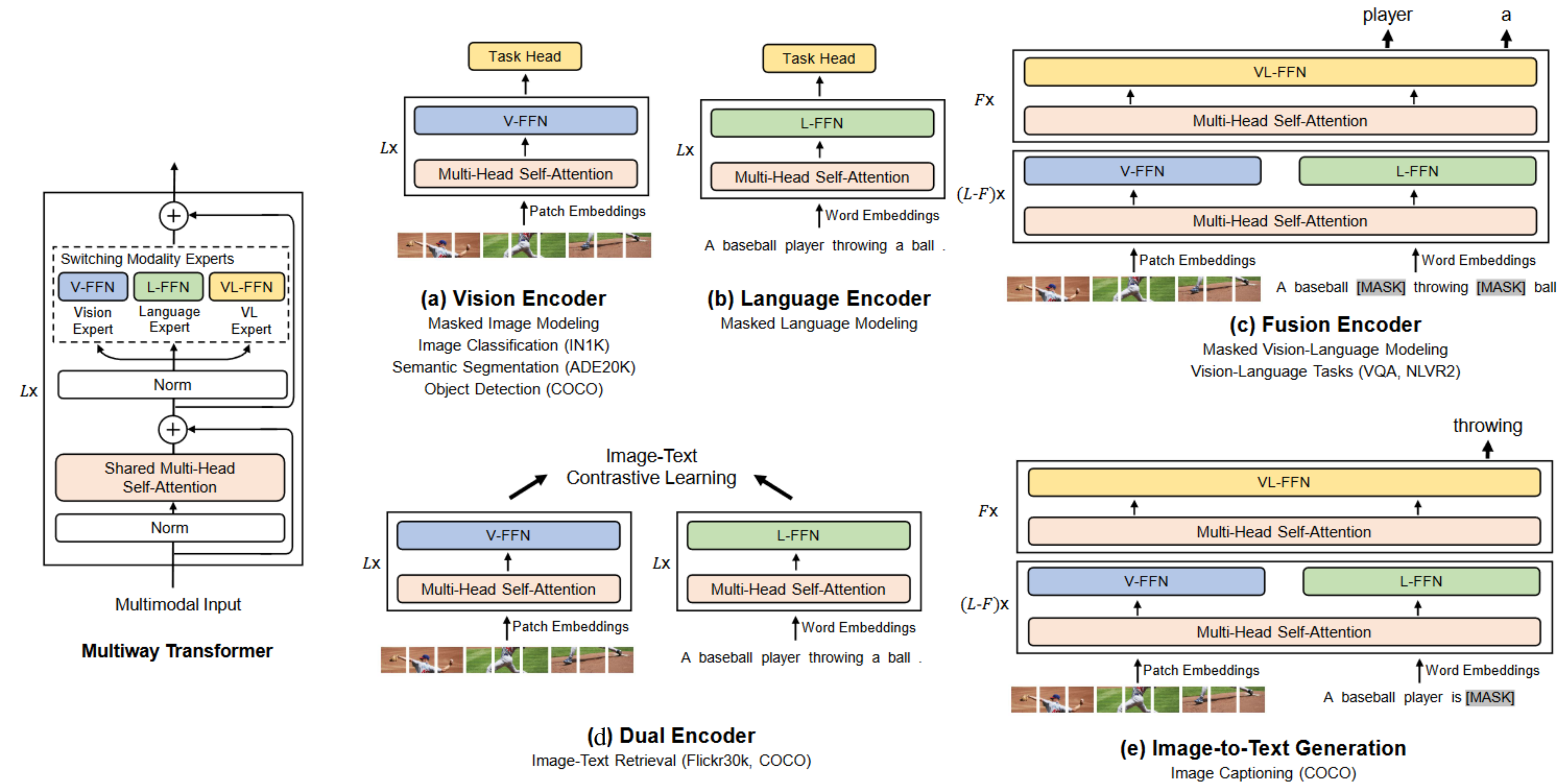
Table 1

Instruction-Tuning LLMs for Different Caption Style

Dataset	Instruction
Fashion200K	USER:<image>+Short description. Assistant:
FashionGen	USER:<image>+Write a detail and professional description for the cloth. Assistant:
Fashion-IQ-cap	USER:<image>+Describe the cloth's style, color, design... and other key points. Assistant:

Table 2

Multimodal Models – Cross-modal Retrieval Models



BEIT-3 can be transferred to various vision and vision-language downstream tasks. With a shared Multiway Transformer, it can reuse the model as (a)(b) vision or language encoders; (c) fusion encoders that jointly encode image-text pairs for deep interaction; (d) dual encoders that separately encode modalities for efficient retrieval; (e) sequence-to-sequence learning for image-to-text generation.

Wang, Wenhui, et al. "Image as a foreign language: Beit pretraining for vision and vision-language tasks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Ablation study for UniFashion:

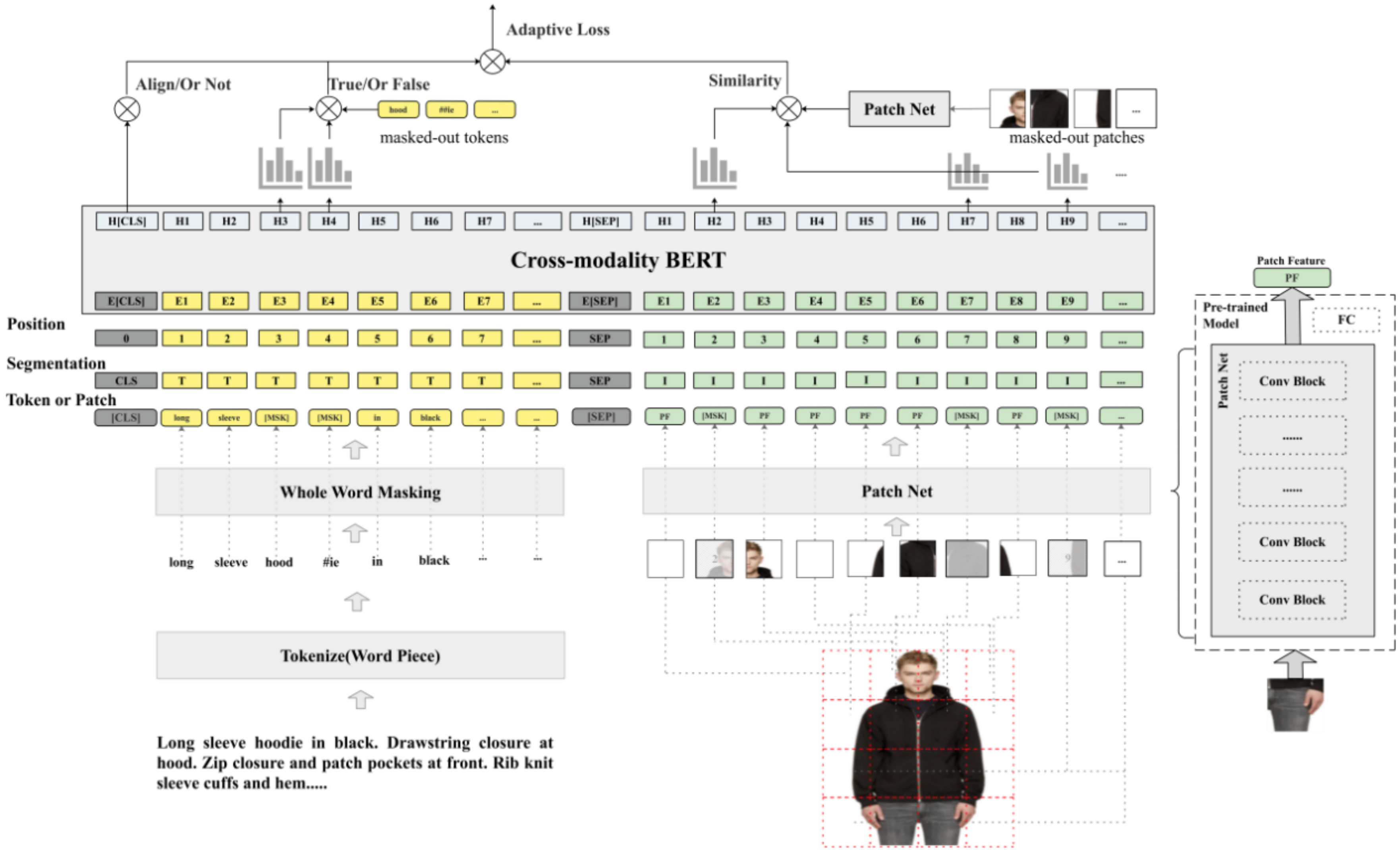
Model	CMR	CIR	FIC	FIG
Base	87.38	64.76	-	-
Base+LLM	87.49	65.04	**36.21**	-
Base+LLM w/ cap	87.49	66.83	**36.21**	-
Base+LLM+diff.	87.55	**67.93**	35.53	12.43

Ablation study and analysis of UniFashion across FashionGen, Fashion-IQ, and VITON-HD Datasets. Metrics reported include average image-to-text and text-to-image recall for cross-modal retrieval(CMR), average recall for composed image retrieval(CIR), BLEU-4 for Fashion Image Captioning, and FID for Fashion image generation (FIG).

Multimodal Models -- Vision-Language Pre-training models

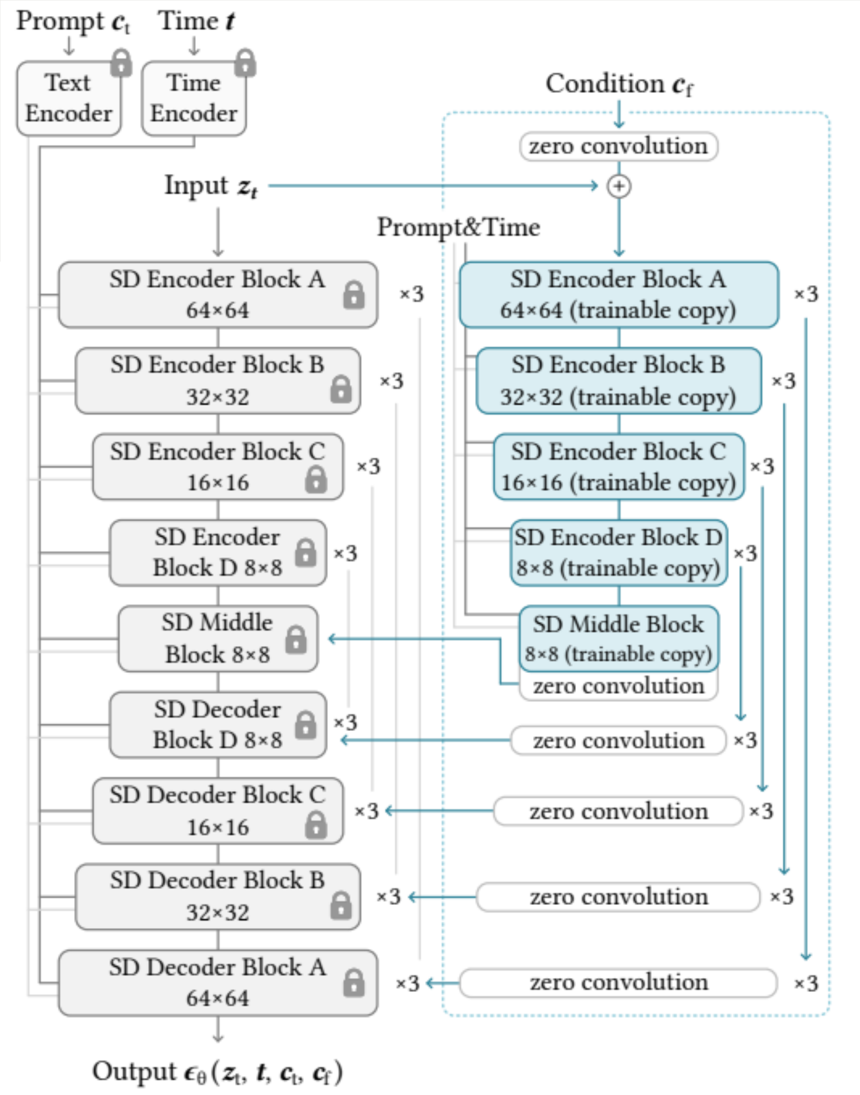
Single-stream architecture:

This technique overtly aligns multi-modal token embeddings, culminating in the generation of token-level matching scores for input image-text pairs.



Gao D, Jin L, Chen B, et al. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 2251-2260.

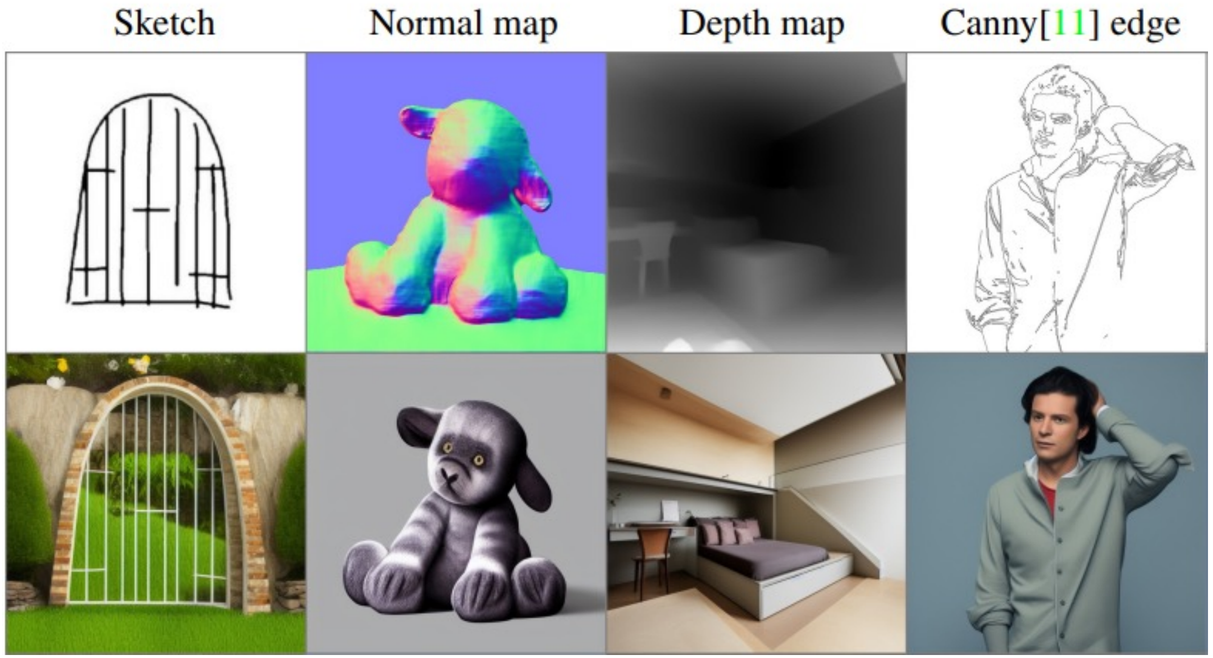
Multimodal Models -- Conditional Diffusion Models



(a) Stable Diffusion (b) ControlNet

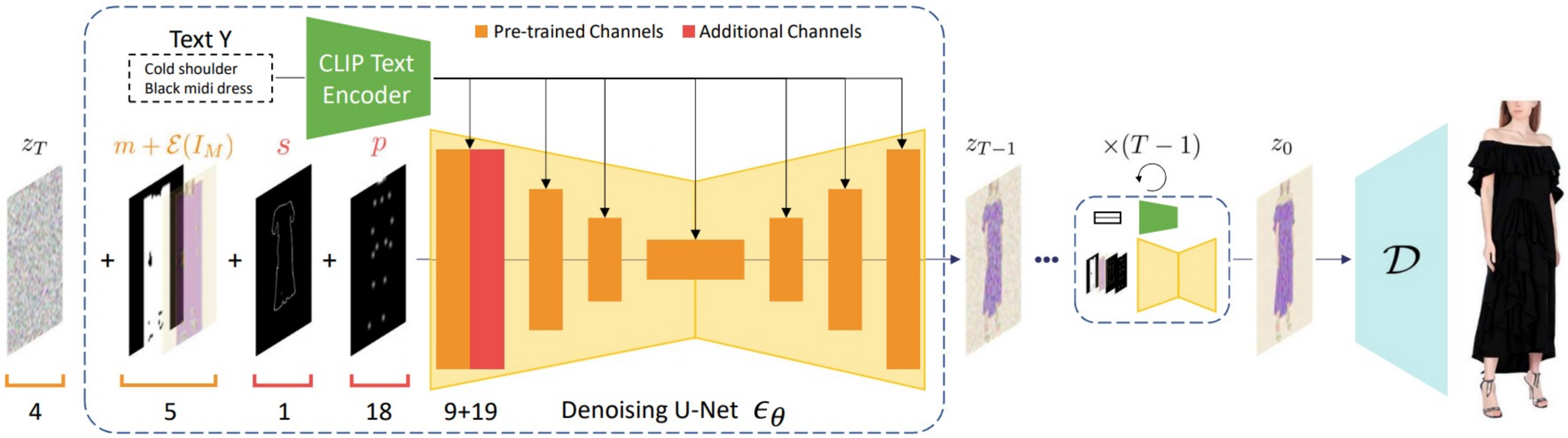
The output of the ControlNet block is combined with the output elements of the corresponding Unet Encoder(Middle) block, and then fed together via jumper to the corresponding Unet Decoder block.

The additional input is another image, such as, sketch, canny edge, etc. This additional input serves as a control condition of the Stable Diffusion model. It can control the image result generated by Stable Diffusion so that it conforms to the conditional image features we input.

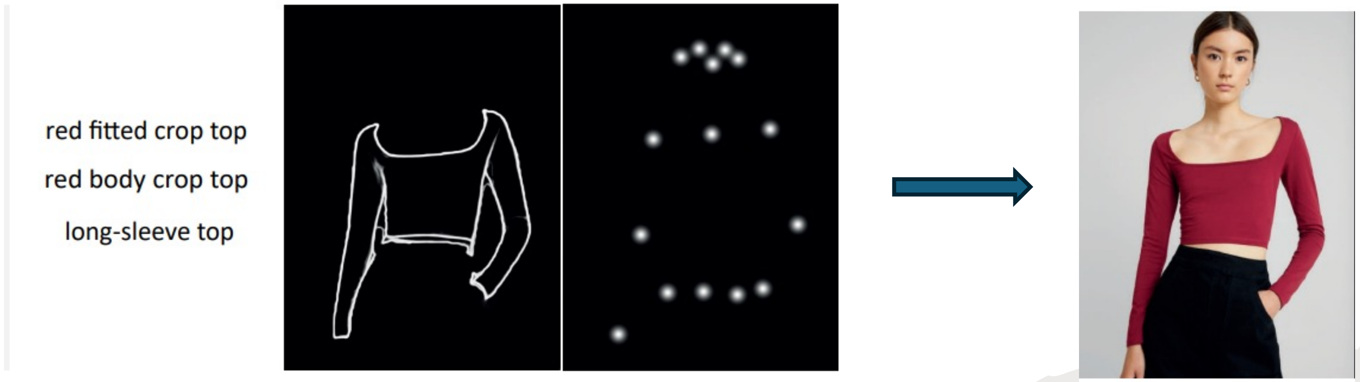


Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Multimodal Models – Conditional Diffusion Models



Overview of Multimodal Garment Designer (MGD), a human-centric latent diffusion model conditioned on multiple modalities (i.e. text, human pose, and garment sketch).



Baldrati, Alberto, et al. "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Multimodal Models -- Multimodal Language Models with Diffusion Models

Model	Source	Stream	Main structure	Backbone	Core technology
Oscar [47]	ECCV20	Single	Transformer	—	Mask mechanism
Uniter [48]	ECCV20	Single	Transformer	—	Mask mechanism
Unicoder [49]	AAAI20	Single	Transformer	—	Mask mechanism
SOHO [50]	CVPR21	Single	CNN, Transformer	—	Mask mechanism
ALIGN [51]	ICML21	Dual	CNN, Transformer	—	Contrastive learning
CLIP [52]	ICML21	Dual	CNN, Transformer	—	Contrastive learning
FashionBERT [53]	SIGIR20	Single	Transformer	BERT [218]	Mask mechanism
EI-CLIP [54]	CVPR22	Dual	Transformer	CLIP [52]	Contrastive learning
TEAM [55]	MM22	Single	Transformer	ViT [219], BERT [218]	Contrastive learning
COTS [56]	CVPR22	Dual	Transformer	ViT [219], BERT [218]	Contrastive learning
CSIC [57]	TCSVT23	Single	Transformer	Uniter [48]	Triplet ranking
AGREE [58]	WSDM23	Dual	Transformer	ViT [219], CLIP [52]	Contrastive learning

Summary of VLP model method for image-text matching.

Zhu, Lei, et al. "Cross-modal retrieval: a systematic review of methods and future directions." arXiv preprint arXiv:2308.14263 (2023).

Multimodal Models -- Multimodal Language Models

Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning	Main Tasks & Capabilities
BLIP-2 (Li et al., 2023g)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former	✗	Visual Dialogue, VQA, Captioning, Retrieval
FROMAGe (Koh et al., 2023b)	OPT-6.7B★	CLIP ViT-L	Linear	✗	Visual Dialogue, Captioning, Retrieval
Kosmos-1 (Huang et al., 2023b)	Magneto-1.3B◊	CLIP ViT-L	Q-Former*	✗	Visual Dialogue, VQA, Captioning
LLaMA-Adapter V2 (Gao et al., 2023)	LLaMA-7B▲	CLIP ViT-L	Linear	✗	VQA, Captioning
OpenFlamingo (Awadalla et al., 2023)	MPT-7B★	CLIP ViT-L	XAttn LLM	✗	VQA, Captioning
Flamingo (Alayrac et al., 2022)	Chinchilla-70B★	NFNet-F6	XAttn LLM	✗	Visual Dialogue, VQA, Captioning
PaLI (Chen et al., 2023j)	mT5-XXL-13B♦	ViT-e	XAttn LLM	✗	Multilingual, VQA, Captioning, Retrieval
PaLI-X (Chen et al., 2023h)	UL2-32B♦	ViT-22B	XAttn LLM	✗	Multilingual, VQA, Captioning
LLaVA (Liu et al., 2023e)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
MiniGPT-4 (Zhu et al., 2023a)	Vicuna-13B★	EVA ViT-g	Linear	✓	VQA, Captioning
mPLUG-Owl (Ye et al., 2023c)	LLaMA-7B▲	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA
InstructBLIP (Dai et al., 2023)	Vicuna-13B★	EVA ViT-g	Q-Former	✓	Visual Dialogue, VQA, Captioning
MultiModal-GPT (Gong et al., 2023)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
LaVIN (Luo et al., 2023)	LLaMA-13B▲	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
Otter (Li et al., 2023b)	LLaMA-7B★	CLIP ViT-L	XAttn LLM	✓	VQA, Captioning
Kosmos-2 (Peng et al., 2023)	Magneto-1.3B◊	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning, Referring, REC
Shikra (Chen et al., 2023f)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Clever Flamingo (Chen et al., 2023b)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
SVIT (Zhao et al., 2023a)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
BLIVA (Hu et al., 2024)	Vicuna-7B★	EVA ViT-g	Q-Former+Linear	✓	Visual Dialogue, VQA, Captioning
IDEFICS (Laurençon et al., 2024)	LLaMA-65B★	OpenCLIP ViT-H	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
Qwen-VL (Bai et al., 2023b)	Qwen-7B♦	OpenCLIP ViT-bigG	Q-Former*	✓	Visual Dialogue, Multilingual, VQA, Captioning, REC
StableLLaVA (Li et al., 2023i)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
Ferret (You et al., 2023)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, Captioning, Referring, REC, GroundCap
LLaVA-1.5 (Liu et al., 2023d)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
MiniGPT-v2 (Chen et al., 2023e)	LLaMA-2-7B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Pink (Xuan et al., 2023)	Vicuna-7B▲	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
CogVLM (Wang et al., 2023c)	Vicuna-7B♦	EVA ViT-E	MLP	✓	Visual Dialogue, VQA, Captioning, REC
DRESS (Chen et al., 2023l)	Vicuna-13B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning
LION (Chen et al., 2023d)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former+MLP	✓	Visual Dialogue, VQA, Captioning, REC
mPLUG-Owl2 (Ye et al., 2023d)	LLaMA-2-7B♦	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning
SPHINX (Lin et al., 2023b)	LLaMA-2-13B♦	Mixture	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Honeybee (Cha et al., 2023)	Vicuna-13B♦	CLIP ViT-L	ResNet blocks	✓	Visual Dialogue, VQA, Captioning
VILA (Lin et al., 2023a)	LLaMA-2-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
SPHINX-X (Gao et al., 2024)	Mixtral-8×7B♦	Mixture	Linear	✓	Visual Dialogue, Multilingual, VQA, Captioning, Referring, REC

Summary of MLLMs with components specifically designed for image generation and editing. (◊: training from scratch; ♦: fine-tuning; ▲: fine-tuning with PEFT techniques; ⋆: frozen). Gray color indicates models not publicly available.

Caffagni, Davide, et al. "The (r) evolution of multimodal large language models: A survey." arXiv preprint arXiv:2402.12451 (2024).

Multimodal Models -- Multimodal Language Models with Diffusion Models

Model	LLM	Visual Encoder	Supporting Model	Main Tasks & Capabilities
GILL (Koh et al., 2023a)	OPT-6.7B★	CLIP ViT-L	SD v1.5★	Visual Dialogue, Retrieval, Image Generation
Emu (Sun et al., 2023b)	LLaMA-13B♦	EVA ViT-g	SD v1.5♦	Visual Dialogue, VQA, Captioning, Image Generation
SEED (Ge et al., 2023a)	OPT-2.7B▲	EVA ViT-g	SD v1.4★	VQA, Captioning, Image Generation
DreamLLM (Dong et al., 2023)	Vicuna-7B♦	CLIP ViT-L	SD v2.1★	Visual Dialogue, VQA, Captioning, Image Generation, Interleaved Generation
LaVIT (Jin et al., 2023)	LLaMA-7B♦	EVA ViT-g	SD v1.5♦	VQA, Captioning, Image Generation
MGIE (Fu et al., 2024)	LLaVA-7B★	CLIP ViT-L	SD v1.5♦	Image Editing
TextBind (Li et al., 2023f)	LLaMA-2-7B♦	EVA ViT-g	SD XL★	Visual Dialogue, VQA, Captioning, Image Generation
Kosmos-G (Pan et al., 2023)	Magneto-1.3B◊	CLIP ViT-L	SD v1.5★	Image Generation, Compositional Image Generation
MiniGPT-5 (Zheng et al., 2023)	Vicuna-7B▲	EVA ViT-g	SD v2.1★	Visual Dialogue, Image Generation, Interleaved Generation
SEED-LLaMA (Ge et al., 2023b)	LLaMA-2-13B♦	EVA ViT-g	SD unCLIP★	Visual Dialogue, VQA, Captioning, Image Generation, Interleaved Generation
CoDi-2 (Tang et al., 2023)	LLaMA-2-7B▲	ImageBind	SD unCLIP★	Visual Dialogue, Audio Understanding, Image Generation, Image Editing
Emu2 (Sun et al., 2023a)	LLaMA-33B♦	EVA ViT-E	SD XL♦	Visual Dialogue, VQA, Captioning, Image Generation, Image Editing
LLMGA (Xia et al., 2023a)	LLaVA-13B♦	CLIP ViT-L	SD XL♦	Visual Dialogue, VQA, Image Generation, Image Editing
SmartEdit (Huang et al., 2023c)	LLaVA-13B▲	CLIP ViT-L	SD♦	Image Editing
VL-GPT (Zhu et al., 2023b)	LLaMA-7B▲	CLIP ViT-L	SD v1.5★	Visual Dialogue, VQA, Captioning, Image Generation, Image Editing
MM-Interleaved (Tian et al., 2024a)	Vicuna-13B♦	CLIP ViT-L	SD v2.1♦	VQA, Captioning, REC, Image Generation, Interleaved Generation
JAM (Aiello et al., 2024)	LLaMA*-7B♦	-	CM3Leon♦	Image Generation, Interleaved Generation

Summary of MLLMs with components specifically designed for image generation and editing. (◊: training from scratch; ♦: fine-tuning; ▲: fine-tuning with PEFT techniques; ⋆: frozen). Gray color indicates models not publicly available.

Caffagni, Davide, et al. "The (r) evolution of multimodal large language models: A survey." arXiv preprint arXiv:2402.12451 (2024).