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

Anonymous ACL submission

Abstract

The fashion domain encompasses a variety of real-world multimodal tasks, including multimodal retrieval and multimodal generation. The rapid advancements in artificial intelligence generated content, particularly in technologies like large language models for text generation and diffusion models for visual generation, have sparked widespread research interest in applying these multimodal models in the fashion domain. However, tasks involving embeddings, such as image-to-text or text-to-image retrieval, have been largely overlooked from this perspective due to the diverse nature of the multimodal fashion domain. And current research on multi-task single models lack focus on image generation. In this work, we present UniFashion, a unified framework that simultaneously tackles the challenges of multimodal generation and retrieval tasks within the fashion domain, integrating image generation with retrieval tasks and text generation tasks. UniFashion unifies embedding and generative tasks by integrating a diffusion model and LLM, enabling controllable and high-fidelity generation. Our model significantly outperforms previous single-task state-of-the-art models across diverse fashion tasks, and can be readily adapted to manage complex vision-language tasks. This work demonstrates the potential learning synergy between multimodal generation and retrieval, offering a promising direction for future research in the fashion domain.

1 Introduction

The fashion domain presents a range of real-world multimodal tasks, encompassing multimodal retrieval (Gao et al., 2020; Wu et al., 2021; Bai et al., 2023) and multimodal generation (Yang et al., 2020) tasks. Such tasks have been utilized in diverse e-commerce scenarios to enhance product discoverability, seller-buyer interaction, and customer conversion rates after catalog browsing (Han et al., 2023; Zhuge et al., 2021). The remarkable progress in the field of artificial intelligence generated content (AIGC), particularly in technologies like large language models (LLMs) (Chiang et al., 2023; Touvron et al., 2023; Brown et al., 2020) for text generation and diffusion models (Rombach et al., 2022; Nichol et al., 2022; Saharia et al., 2022) for visual generation, has sparked widespread research interest in applying these multimodal models in the fashion domain.

Multimodal large language models (Liu et al., 2023a; Dai et al., 2023; Dong et al., 2023) (MLLMs) seem to emerge as a promising direction for a single multi-task model. However, due to the heterogeneous nature of the multimodal fashion tasks (Han et al., 2023), existing MLLMs lack the capability to be directly applied to the fashion domain, such as embedding ability. For example, in the fashion domain, retrieval tasks that rely on embedding ability, like image-to-text or text-to-image retrieval, have been largely neglected from this aspect. Furthermore, MLLMs lack the ability to solve composed image retrieval (CIR) (Liu et al., 2021; Baldrati et al., 2022) task, which composes the reference image and related caption into a joint embedding to calculate similarities with the candidate images and is particularly relevant in fashion recommendation systems (Han et al., 2017).

Drawing inspiration from GRIT (Muennighoff et al., 2024), which successfully combined embedding and generative tasks into a unified model for text-centric applications and showed improved embedding performance through the addition of a generative objective, it becomes clear that investigating task correlations and integrating embedding with generative models in the fashion domain is is both necessary and promising.

While previous works (Han et al., 2023; Zhuge et al., 2021) in the fashion domain have also proposed using a single model for solving multiple tasks, they ignore the image generation tasks. Besides, for fashion tasks such as try-on (Choi et al.,

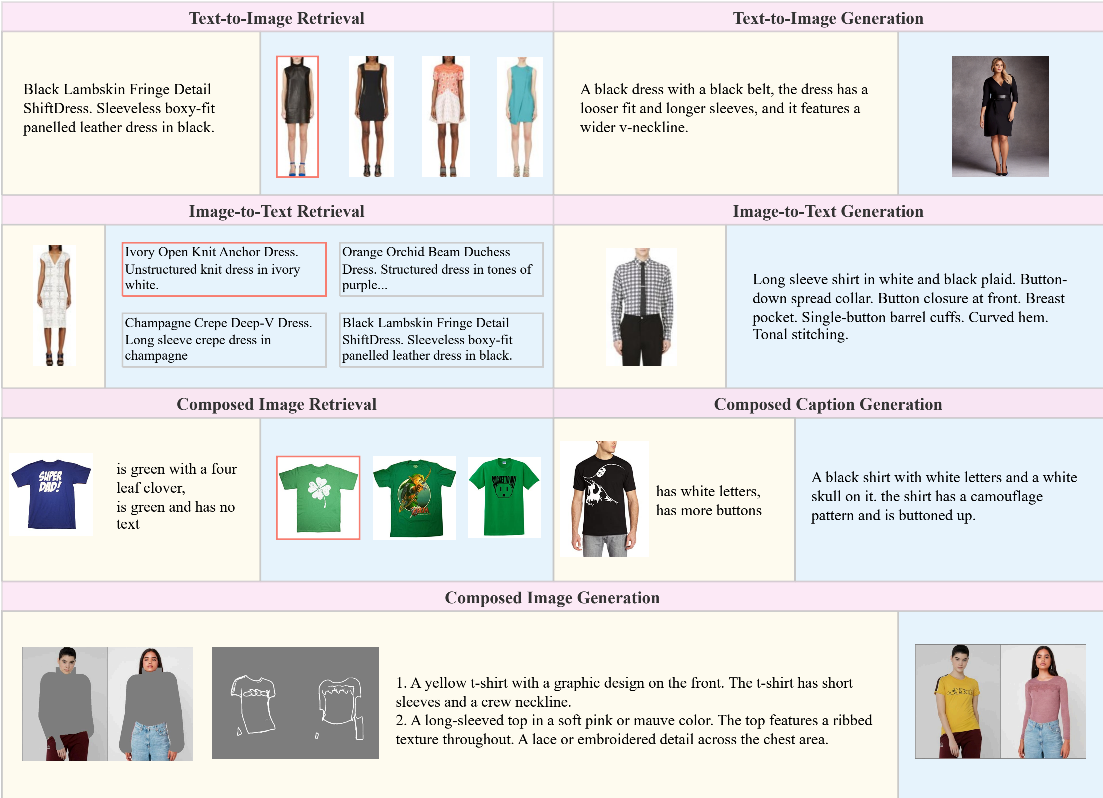

Figure 1: Illustration of the fashion tasks encompassed in our UniFashion framework: cross-modal retrieval, text-guided image retrieval, fashion image captioning, and fashion image generation. Model inputs highlighted with a light yellow background and outputs denoted by a light blue background.

2021) and fashion design (Baldrati et al., 2023b), it is generally required to generate target images based on multimodal input. However, previous works (Baldrati et al., 2023b) in fashion image generation typically adopt the CLIP text encoder to encode text information, which may not be capable of effectively understanding the textual context due to their weaker text encoder, as noted in Saharia et al. (2022). Hence, we posit that current studies have not fully exploited the potential in learning synergy between generation and retrieval.

In this work, we propose UniFashion, which unifies retrieval and generation tasks by integrating LLMs and diffusion models, as illustrated in Figure 2. UniFashion consists of three parts: The *Q-Former* is crucial for amalgamating text and image input, creating multimodal learnable queries. These queries, once refined through task-specific adapters, enable the *LLM* module to utilize them as soft prompts for generating captions for target images. Simultaneously, the *diffusion module* utilizes the learnable queries as conditions to guide the latent diffusion model in image synthesis and editing tasks. To enable controllable and high-fidelity generation, we propose a two-phase training strategy. In the first phase, we perform multimodal representation learning on image-text pairs datasets. We freeze Q-Former and fine-tune the LLM and diffusion modules, ensuring they develop the capability to comprehend the multimodal representations provided by Q-Former. Subsequently, in the second phase, we proceed to fine-tune UniFashion on datasets with multimodal inputs, such as Fashion-IQ, where we freeze the LLM and diffusion modules, only tuning Q-Former. This strategy ensures that Q-Former is adept at crafting multimodal representations that effectively integrate both reference images and text inputs.

UniFashion holds three significant advantages that address the challenges in multimodal fashion retrieval and generation:

For the first time, we conduct an in-depth study of the synergistic modeling of multimodal retrieval and generation tasks within the fashion domain, thoroughly exploiting the inter-task relatedness. Further, we introduce UniFashion, a versatile, unified model that can handle all fashion tasks.

Secondly, our model enhances performance via mutual task reinforcement. Specifically, the caption generation module aids the CIR task, while jointly training the generation and retrieval tasks improves the multimodal encoder for the diffusion module.

Third, extensive experiments on diverse fashion tasks—including cross-modal retrieval, composed image retrieval, and multimodal generation—demonstrate that our unified model significantly surpasses previous state-of-the-art methods.

2 Preliminaries and Related Works

2.1 Fashion Tasks

Fashion tasks encompass a range of image and language manipulations, including cross-modal retrieval, composed image retrieval, fashion image captioning and generation, etc. The representative tasks can be briefly divided into the following two groups:

Fashion Retrieval generally consists of Cross-Modal Retrieval (CMR) (Ma et al., 2022; Rostamzadeh et al., 2018) and composed image retrieval (CIR) tasks (Baldrati et al., 2023a; Bai et al., 2023). CMR requests to efficiently retrieve the most matched image/sentence from a large candidate pool $\mathcal{D}$ given a text/image query. CIR is a special type of image retrieval with a multimodal query (a combination of a reference image and a modifying text) matched against a set of images. It retrieves a target image from a vast image database based on a reference image and a text description detailing changes to be applied to the reference image. In this scenario, a query pair $p = \{I_R, t\}$ is provided, where I_R is the reference image and t is the text describing the desired modifications. The challenge for this task is to accurately identify the target image I_T that best matches the query among all potential candidates in the image corpus $\mathcal{D}$.

Fashion Generation consists of Fashion Image Captioning (FIC) and Fashion Image Generation (FIG). FIC (Yang et al., 2020) aims to generate a descriptive caption for a product based on the visual and/or textual information provided in the input. FIG aims to generate images based on the multimodal input, such as try-on (Choi et al., 2021; Gou et al., 2023) and fashion design (Baldrati et al., 2023b).

2.2 Multimodal Language Models

Recent research has witnessed a surge of interest in multimodal LLMs, including collaborative models (Wu et al., 2023; Yang et al., 2023b; Shen et al., 2023) and end-to-end methods (Alayrac et al., 2022; Zhao et al., 2023; Li et al., 2022; Bao et al., 2021; Wang et al., 2022b,a,a). More recently, some works also explore training LLMs with parameter-efficient tuning (Li et al., 2023b; Zhang et al., 2023b) and instruction tuning (Dai et al., 2023; Liu et al., 2023a; Ye et al., 2023; Zhu et al., 2023a; Li et al., 2023a). They only focus on generation tasks, while UniFashion is built upon a unified framework that enables both retrieval and generation tasks.

2.3 Diffusion Models

Diffusion generative models (Rombach et al., 2022; Ramesh et al., 2021; Nichol et al., 2022; Ruiz et al., 2023) have achieved strong results in text conditioned image generation works. Among contemporary works that aim to condition pretrained latent diffusion models, ControlNet (Zhang et al., 2023a) proposes to extend the Stable Diffusion model with an additional trainable copy part for conditioning input. In this work, we focus on the fashion domain and propose a unified framework that can leverage latent diffusion models that directly exploit the conditioning of textual sentences and other modalities such as human body poses and garment sketches.

2.4 Problem Formulation

Existing fashion image retrieval and generation methods are typically designed for specific tasks, which inherently restricts their applicability to the various task forms and input/output forms in the fashion domain. To train a unified model that can handle multiple fashion tasks, our approach introduces a versatile framework capable of handling multiple fashion tasks by aligning the multimodal representation into the LLM and the diffusion model. This innovative strategy enhances the model's adaptability, and it can be represented as:

$$I_{\text{out}}, T_{\text{out}} = \mathcal{F}_{\mathcal{T}_{\text{Ret}}, \mathcal{T}_{\text{Gen}}}(I_{\text{in}}, T_{\text{in}}; \Theta), \quad (1)$$

where $\mathcal{F}_{\mathcal{T}}$ represents the unified model parameterized by Θ, it consists of retrieval module $\mathcal{T}_{Ret}$ and generative module $\mathcal{T}_{Gen}$.

3 Proposed Model: UniFashion

In this section, we introduce the UniFashion to unify the fashion retrieval and generation tasks into

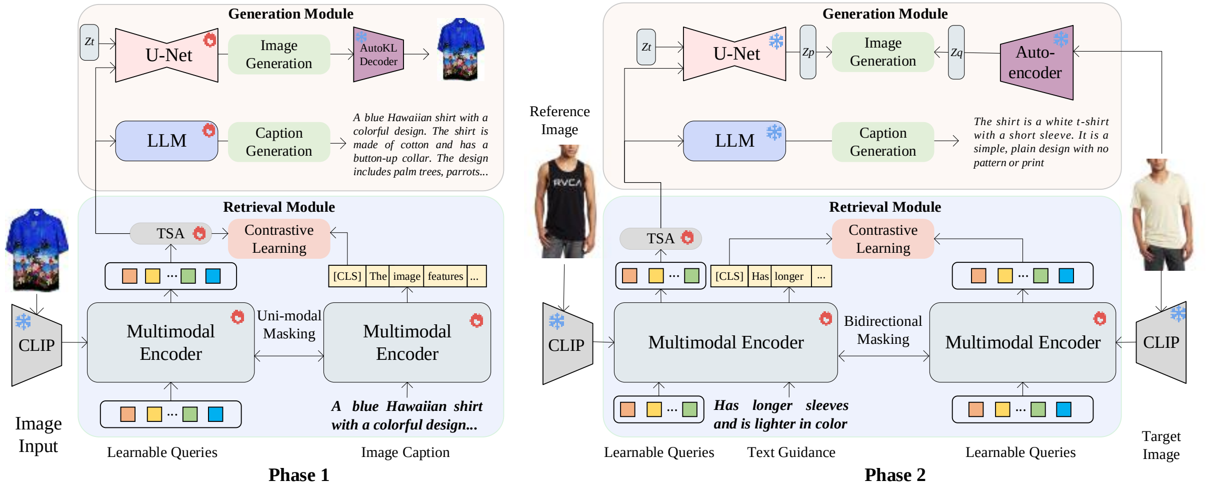

Figure 2: Overview of the training framework of our UniFashion model. **Phase 1** - Cross-modal Pre-training: UniFashion acquires robust cross-modal fashion representation capabilities through pre-training, leveraging both the language model and the diffusion model. **Phase 2** - Composed Multimodal 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 with the LLM and the diffusion model for enhanced performance.

a single model. By combining **retrieval and generative modules**, the proposed UniFashion employs a **two-stage** training strategy to capture relatedness between image and language information. Consequently, it can seamlessly switch between two operational modes for cross-modal tasks and composed modal tasks.

3.1 Phase 1: Cross-modal Pre-training

In the first stage, we conduct pre-training on the retrieval and generation modules to equip the Large Language Model (LLM) and diffusion model with strong cross-modal fashion representation capabilities for the next phase.

3.1.1 Cross-modal Retrieval

For cross-modal retrieval tasks, given a batch of image caption pairs $p = \{I, C\}$, we first calculate their unimodal representations using an independent method. In particular, we adopt a lightweight Querying Transformer, i.e., Q-Former in BLIP-2 (Li et al., 2023b), to encode the multimodal inputs, as it is effective in bridging the modality gap. To avoid information leaks, we employ a unimodal self-attention mask (Li et al., 2023b), where the queries and text are not allowed to see each other:

$$\begin{aligned} Z_I &= \text{Q-Former}(I, q), \\ Z_C &= \text{Q-Former}(C). \end{aligned} \tag{2}$$

where the output sequence Z_I is the encoding result of an initialized learnable query q with the input image and Z_C is the encoded caption, which contains the embedding of the output of the [CLS] token e_{cls}, which is a representation of the input caption text. Since Z_I contains multiple output embeddings (one from each query), we first compute the pairwise similarity between each query output and e_{cls}, and then select the highest one as the image-text similarity. In our experiments, we employ 32 queries in q, with each query having a dimension of 768, which is the same as the hidden dimension of the Q-Former. For cross-modal learning objective, we leverage the Image-Text Contrastive Learning (ITC) and Image-Text Matching (ITM) method. The first loss term is image-text contrastive loss, which has been widely adopted in existing text-to-image retrieval models. Specifically, the image-text contrastive loss is defined as:

$$\mathcal{L}_{\text{ITC}}(X, Y) = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp[\lambda(X_i^T \cdot Y^i)]}{\sum_{j=1}^{B}\exp[\lambda(X_i^T \cdot Y^j)]}, \tag{3}$$

where λ is a learnable temperature parameter. ITM aims to learn fine-grained alignment between image and text representation. It is a binary classification task where the model is asked to predict whether an image-text pair is positive (matched) or negative (unmatched), it is defined as,

$$\mathcal{L}_{\text{ITM}}(X, Y) = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp f_\theta(X_i, Y_i)}{\sum_{j=1}^{B}\exp f_\theta(X_j, Y_i)}, \tag{4}$$

Then, we maximize their similarities via symmetrical contrastive loss:

$$\mathcal{L}_{\text{cross}} = \mathcal{L}_{\text{ITC}}(t_c, Z_I) + \mathcal{L}_{\text{ITM}}(Z_C, Z_I), \quad (5)$$

3.1.2 Cross-modal Generation

As depicted in Fig. 2, after the learnable queries q pass through the multimodal encoder, they are capable of integrating the visual information with textual guidance. However, in Section 3.1.1, we did not specify a learning target for q. Empirically, the q that has been merged with the reference image and edited text information should be equivalent to the encoding of the target image. This implies that we should be able to reconstruct the target image and its caption based on q. In this section, we will employ generative objectives to improve the representation of augmented q.

In the first stage, we connect the Q-Former (equipped with a frozen image encoder) to a Large Language Model (LLM) to harness the LLM's prowess in language generation, and to a diffusion model to exploit its image generation capabilities. Notably, we exclusively train the model using image-text pairs throughout this process. As depicted in Figure 2, we employ a Task Specific Adapter (TSA) layer to linearly project the output query embeddings q to match the dimensionality of the embeddings used by the LLM and diffusion model. In this stage, we freeze the parameters of the Q-Former and fine-tune only the adapter layers, connecting LLM and diffusion models. This approach allows us to develop a discriminative model that can evaluate whether queries q can generate the target image and its corresponding caption.

Target caption generation. The adapter layer is placed before the LLM to map the output of Q-Former to the text embedding space of the LLM. To synchronize the space of Q-Former with that of the LLM, we propose to use the image-grounded text generation (ITG) objective to drive the model to generate texts based on the input image by computing the auto-regressive loss:

$$\mathcal{L}_{\text{ITG}} = -\frac{1}{L}\sum_{l=1}^{L} \log p_\phi(w_l^g | w_{<l}^g, f_\theta(q)), \quad (6)$$

where $w^g = (w_1^g, ..., w_L^g)$ represents the ground-truth caption of image I with length L, $q =$ Q-Former(I, q), ϕ denotes the LLM's parameters, and θ denotes the text adapter layers' parameters.

Target image generation. In the first stage, our task also aims to reconstruct the image $\hat{I_T}$ from q. As in standard latent diffusion models, given an encoded input $\mathbf{x}$, the proposed denoising network is trained to predict the noise stochastically added to $\mathbf{x}$. The corresponding objective function can be specified as:

$$\mathcal{L}_{\text{q2I}} = \mathbb{E}_{\boldsymbol{\epsilon}^y, \mathbf{x}_0}[||\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_\eta^x(\mathbf{x}_{t^x}, f_\zeta(q), t^x)\|^2], \quad (7)$$

where η denotes the u-net models' parameters and ζ denotes the image adapter layers' parameters. The overall loss in the first stage can be expressed:

$$\mathcal{L}_{\text{ph1}} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{ITG}} + \mathcal{L}_{\text{q2T}}. \quad (8)$$

After the first training stage, we can leverage the LLM and diffusion model as discriminators to guide the generation of composed queries.

3.2 Phase 2: Composed Multimodal Fine-tuning

In this phase, the inputs are reference image and guidance text, and we fine-tune the model for composed multimodal retrieval and generation tasks.

3.2.1 Composed Image Retrieval

For CIR task, the target image I_T generally encompasses the removal of objects and the modification of attributes in the reference image. To solve this problem, as depicted in Fig. 2, the multimodal encoder is utilized to extract features from the reference image and the guide text. It joint embeds the given pair $p = \{I_R, t\}$ in a sequential output. Specifically, a set of learnable queries q concatenated with text guidance t is introduced to interact with the features of the reference image. Finally, the output of Q-Former is the multimodal synthetic prompt Z_R. We use a bi-directional self-attention mask, similar to the one used in BLIP2 (Li et al., 2023b), where all queries and texts can attend to each other. The output query embeddings Z_R thus capture multimodal information:

$$\begin{aligned} Z_R &= \text{Q-Former}(I_R, t, q_R), \\ Z_T &= \text{Q-Former}(I_T, q_T). \end{aligned} \quad (9)$$

Noting that the output sequence Z_R consists of learnable queries q and encoded text guidance $\boldsymbol{t}$, which includes e_{cls}, the embedding of the output of the [CLS] token. On the other hand, the target image's output sequence Z_T consists only of learnable queries. Therefore, we can use Z_R as a representation that incorporates information from the reference image and the guidance text and align it with the features of the target image Z_T. Moreover, as UniFashion acquires the ability to generate captions for images from Sec. 3.1.2, we can generate captions for the candidate images and use e_{cls} to retrieve the caption Z_C of the target image. Then, the final contrastive loss for the CIR task is:

$$\mathcal{L}_{\text{cir}} = \mathcal{L}_{\text{ITC}}(e_{cls}, Z_T) + \mathcal{L}_{\text{ITC}}(e_{cls}, Z_C) + \mathcal{L}_{\text{ITM}}(\boldsymbol{t}, Z_T), \tag{10}$$

3.2.2 Composed multimodal Generation

For these generation tasks, we freeze the LLM parameters and tune the parameters of the task-specific adapters, the diffusion model, and the Q-Former. The loss function for the target image's caption generation is formulated in a way that is similar to Eq. 6:

$$\mathcal{L}_{\text{ITG}} = -\frac{1}{L}\sum_{l=1}^{L}\log p_\phi(w_l^g | w_{<l}^g, f_\theta(q_R)), \tag{11}$$

The loss function for the target image generation is formulated in a way that is similar to Eq. 7:

$$\mathcal{L}_{\text{q2I}} = \mathbb{E}_{\boldsymbol{\epsilon}^y, \mathbf{x}_0}[\|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_\eta^x(\mathbf{x}_{t^x}, f_\zeta(q_R), t^x)\|^2], \tag{12}$$

The overall loss in the second stage can be expressed as:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{cir}} + \mathcal{L}_{\text{ITG}} + \mathcal{L}_{\text{q2I}}. \tag{13}$$

4 Experiments

4.1 Experimental Setup

We initialize the multimodal encoder from BLIP2's Q-Former and MLLM from LLaVA-1.5. As for the diffusion module, following, StableVITON, we inherit the autoencoder and the denoising U-Net of the Stable Diffusion v1.4. We initialize the weights of the U-Net from the Paint-by-Example and for more refined person texture, we utilized a VAE fine-tuned on the VITONHD dataset from StableVITON. The statistics of the two-stage datasets can be found in Table 6. For cross-modal retrieval, we evaluated UniFashion on FashionGen validation set. For the image captioning task, UniFashion is evaluated in the FashionGen dataset. For the composed image retrieval task, we evaluated the Fashion-IQ validation set. To maintain consistency with previous work, for the composed image generation task, we fine-tuned UniFashion and evaluated it on the VITON-HD and MGD datasets. More details can be found in Appendix D.

Phase 1: For multimodal representation learning, we follow BLIP2 and pretrain the Q-Former on fahsion image-text pairs. To adapt the model for multimodal generation, we freeze the parameters of Q-Former and fine-tune the MLLM and diffusion model with their task specific adapters separately. Due to the different styles of captions in different fashion datasets, we adopt the approach of instruction tuning to train the LLM so that it can generate captions of different styles. More details can be found in Appendix E.

Phase 2: In order to make UniFashion have the composed retrieval and generation abilities, we freeze the parameters of LLM and diffusion model, only fine-tune the multimodal encoder.

4.2 Evaluation Methods

We compare our models with previous state-of-the-art methods on each task. For extensive and fair comparisons, all prior competitors are based on large-scale pre-trained models.

Cross-modal retrieval evaluation: We consider both image-to-text retrieval and text-to-image retrieval with random 100 protocols used by previous methods. 100 candidates are randomly sampled from the same category to construct a retrieval database. The goal is to locate the positive match depicting the same garment instance from these 100 same-category negative matches. We utilize Recall@K as the evaluation metric, which reflects the percentage of queries whose true target ranked within the top K candidates.

Fashion image captioning evaluation: For evaluating the performance of caption generation, we utilize BLEU-4, METEOR, ROUGE-L, and CIDEr as metrics.

Composed fashion image retrieval evaluation: We compare our UniFashion with CIR methods and the FAME-ViL model of V + L that is oriented towards fashion in the original protocol used by Fashion-IQ. For this task, we also utilize Recall@K as the evaluation metric.

Composed fashion image generation evaluation: We compare our UniFashion with try-on methods on VITON-HD dataset and fashion design works on MGD dataset. To evaluate the quality of image generation, we use the Frechet Inception Distance (FID) score to measure the divergence between two multivariate normal distributions and employ the CLIP Score (CLIP-S) provided in the TorchMetrics

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

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 2: Image captioning task performance on the FashionGen dataset.

library to assess the adherence of the image to the textual conditioning input (for fashion design task).

4.3 Comparative Analysis of Baselines and Our Method

UniFashion performs better compared to baselines in all data sets. Tab. 1 presents the evaluation results for each baseline and our models in FashionGen data sets for cross-modal retrieval. UniFashion outperforms most of the baseline models on both the text-to-image and image-to-text tasks. Following FAME-ViL, we also adopt a more challenging and practical protocol that conducts retrieval on the entire product set, which is in line with actual product retrieval scenarios. In Tab. 2, we performed a comparison between our UniFashion and other baselines on the FashionGen dataset for the image captioning task. By integrating the powerful generative ability of the LLM, our model performed significantly better than the traditional multimodal models in this task. In Tab. 4, we conducted a comparison between our UniFashion and CIR-specialist methods. Our findings are in line with those of Tab. 1.

After fine-tuning UniFashion on different modal input composed image generation/editing tasks, it also demonstrates excellent performance. Tab. 3 evaluates the quality of the generated image of UniFashion in the VITON-HD unpaired setting. In order to verify that our model can achieve good results in a variety of modal inputs, we have conducted tests respectively on the traditional try-on task and the fashion design task proposed in MGD. For a fair evaluation with baselines, all the models are trained at a 512 × 384 resolution. To confirm the efficacy of our approach, we assess the realism using FID and KID score on all the tasks and using CLIP-S score for fashion design task. As can be seen, the proposed UniFashion model consistently outperforms competitors in terms of realism (i.e., FID and KID) and coherence with input modalities (i.e., CLIP-S), indicating that our method can better encode multimodal information. Meanwhile, although our model is slightly lower than StableVITON on the try-on task, this is because we froze the parameters of the diffusion model on the try-on task and only fine-tuned the Q-former part, but it can still achieve top2 results.

4.4 Ablation Study

Our model completes the multimodal composed tasks in more aspects. In Tab. 4, we also carry out ablation studies on different retrieval methods. Since UniFashion is capable of generating captions, for the CIR task, we initially utilize UniFashion to generate the captions of candidate images and then conduct the image retrieval task (denoted as UniFashion w/o cap) and the caption retrieval task (denoted as UniFashion w/o img). We find that our single-task variant has already achieved superior performance in the relevant field. Furthermore, due to the generative ability of our model, the pregenerated candidate library optimizes the model's performance in this task. For specific implementation details, please refer to Appendix C.

We researched the impact of different modules in UniFashion on various fashion tasks. In Tab. 5, we perform an ablation study on the proposed model architecture, with a focus on LLM and diffusion models. For comparison on the cross-

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)	✗	✗	✓	✓	8.42	0.67	-
fashion design task							
SDEdit (Meng et al., 2021)	✓	✓	✓	✗	15.12	5.67	28.61
MGD (Baldrati et al., 2023b)	✓	✓	✓	✗	12.81	3.86	30.75
UniFashion (Ours)	✓	✓	✓	✗	**12.43**	**3.74**	31.29

Table 3: Performance analysis of unpaired settings on VITON-HD and MGD datasets across different input modalities.

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	72.43	55.64	73.89	59.35	78.58	54.92	74.97	64.85
UniFashion w/o cap	49.65	72.17	56.88	74.12	59.29	78.11	55.27	74.80	65.04
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 4: 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.

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

Table 5: 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).

modal retrieval task (CMR), we design the base model as directly fine-tuning BLIP2 without any new modules. The results indicate that the base model performs relatively well on this task and that the introduction of other modules does not lead to significant improvements. However, in the CIR task, the introduction of LLM and diffusion models as supervision can lead to significant improvements, especially when utilizing the captions pre-generated by the UniFashion to assist in retrieval, resulting in greater benefits. At the same time, we note that, after introducing the diffusion model, it may have some negative impact on the model's image captioning ability, possibly due to the inherent alignment differences between LLM and the diffusion model.

5 Conclusion

We have introduced UniFashion, a unified framework that addresses the challenges in multimodal generation and retrieval tasks within the fashion domain. By unifying embedding and generative tasks with a diffusion model and LLM, UniFashion enables controllable and high-fidelity generation, significantly outperforming previous single-task state-of-the-art models across diverse fashion tasks. Our model's ability to readily adapt to manage complex vision-language tasks demonstrates its potential for enhancing various e-commerce scenarios and fashion-related applications. The findings of this study highlight the importance of exploring the potential learning synergy between multimodal generation and retrieval, and provide a promising direction for future research in the fashion domain.

6 Limitations

This section aims to highlight the limitations of our work and provide further insight into the research in this area. Our model relies on diffusion for multimodal interaction, which means that the composed image generation processes may take longer. In our experiments, we tested the performance of our model on one A100 (80G) GPU. During inference, using 1000 examples from VITON-HD dataset, UniFashion took approximately 3.15 seconds for each image generation. We believe it would be beneficial to explore more efficient sampling methods, such as DPM-Solver++ (Lu et al., 2022), to improve the overall efficiency of UniFashion.

References

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

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. 2023. Sentence-level prompts benefit composed image retrieval. *arXiv preprint arXiv:2310.05473*.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2023a. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–24.

Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023b. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23393–23402.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.

Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115.

Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 2023. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7599–7607.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2023. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2669–2680.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2024. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2023. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)(2021)*, pages 2105–2114.

Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. 2023b. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *arXiv preprint arXiv:2305.16304*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18051–18061.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu

Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5270–5279.

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

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

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. 2023. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 915–923.

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 preprint arXiv:2303.04671*.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317.

Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. 2023. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023a. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 1–17. Springer.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Xiangyu Zhao, Bo Liu, Qijiong Liu, Guangyuan Shi, and Xiao-Ming Wu. 2023. Making multimodal generation easier: When diffusion models meet llms. *arXiv preprint arXiv:2310.08949*.

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

Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. 2023b. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–22.

Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12647–12657.

A Ethics Statement

We adhere to the ACL Ethics Policy and have conducted our research using publicly available repositories and datasets. Our primary focus is on investigating the integration of diffusion models and LLMs for multimodal generation. Therefore, the results should be seen as AI-generated content. While we have not observed deliberate harmful content, the model has the potential to generate such content if triggered. We have taken steps to minimize this risk through fine-tuning on public datasets, but caution is still exercised. In future, we will prioritize improving downstream performance and exploring methods to enhance control over the generation process. To ensure reproducibility and support future research, we have made all resources publicly available and provided proper citations to previous research within the code.

B Basics of Diffusion Models

After the initial proposal of diffusion models by (Sohl-Dickstein et al., 2015), they have demonstrated remarkable capacity for generating high-quality and diverse data. DDPM (Ho et al., 2020) connects diffusion and score matching models through a noise prediction formulation, while DDIM (Song et al., 2020) proposes an implicit generative model that generates deterministic samples from latent variables.

Given a data point sampled from a real data distribution $x_0 \in q(x)$, during forward diffusion, x_0 is gradually "corrupted" at each step t by adding Gaussian noise to the output of step t-1. It produces a sequence of noisy samples $x_1, ..., x_T$. Each step is controlled by:

Stable Diffusion Model. In the field of image generation, diffusion models operate by progressively denoising a random variable that is sampled from a Gaussian distribution. Latent diffusion models (LDMs) operate in the latent space of a pre-trained autoencoder achieving higher computational efficiency while preserving the generation quality. Stable diffusion model is composed of an autoencoder with an encoder $\mathbb{E}$ and a decoder $\mathbb{D}$, a conditional U-Net denoising model ϵ_θ, and a CLIP-based text encoder. With the fixed encoder $\mathbb{E}$, an input image x is first transformed to a lower-dimensional latent space $z_0 = \mathbb{E}(x)$. The decoder $\mathbb{D}$ performs the opposite operation, decoding z_0 into the pixel space. When considering a latent variable z and its noisy counterpart z_t, which is obtained by incrementally adding noises to z over t steps, the latent diffusion models are designed to train the $\boldsymbol{\epsilon}_\theta(\cdot)$ to predict the added noise $\boldsymbol{\epsilon}$ using a standard mean squared error loss:

$$\mathcal{L} := \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon},t}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)||^2]. \quad (14)$$

Multimodal Conditional Generation. In the context of our current work, we have a particular focus on the pre-trained multimodal latent diffusion models. For a multimodal conditional generation, given a target image $\mathbf{x}_0$, in addition to the textual information, the input condition $\mathbf{y}_0$ also contains other constraints such as . The aim is to model the conditional data distribution $q(\mathbf{x}_0|\mathbf{y}_0)$, where $\mathbf{y}_0$ contains different modalities prompts. The conditioning mechanism is implemented by first encoding conditional information, then the denoising network $\boldsymbol{\epsilon}_\theta$ conditions on y_0 via cross-attention. The label y_0 in a class-conditional diffusion model $\boldsymbol{\epsilon}_\theta(x_t|y_0)$ is replaced with a null label $\emptyset$ with a fixed probability during training. At inference time, with a guidance scale s, the modified score estimate is further in the direction of $\boldsymbol{\epsilon}_\theta(x_t|y_0)$ and away from $\boldsymbol{\epsilon}_\theta(x_t|\emptyset)$ as follows:

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t|\mathbf{y}_0) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t|\emptyset) + \\ s \cdot (\boldsymbol{\epsilon}_\theta(\mathbf{x}_t|\mathbf{y}_0) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t|\emptyset)).\end{aligned}$$

C Datasets

We test the effectiveness of UniFashion by experimenting on different tasks including fashion image captioning, cross-modal retrieval, composed image retrieval and composed image generation. Table 6 shows the statistics of the datasets used for two stage in our training process.

We use the FashionGen and FshaionIQ (Lin et al., 2014) datasets for retrieval tasks. FashionGen contains 68k fashion products accompanied by text descriptions. Each product includes 1 - 6 images from different angles, resulting in 260.5k image-text pairs for training and 35.5k for testing. Fashion-IQ contains 18k training triplets (that is, reference image, modifying text, target image) and 6k validation triplets over three categories: Dress, Shirt, and Toptee. Each pair (reference image, target image) is manually annotated with two modifying texts, which are concatenated.

For fashion image captioning tasks, we utilize the FashionGen (Zang et al., 2021) dataset. Additionally, to enhance our model's capability in

Data types	Dataset	Size	Stage 1	Stage 2	Metrics
CMR	FashionGen (Lin et al., 2014)	60K	✓	✓	R@K
	Fashion200K (Krishna et al., 2017)	172K	✓	✗	-
CIR	Fashion-IQ (Liu et al., 2023a)	30K	✗	✓	R@K
FIC	FashionGen (Liu et al., 2023a)	80K	✓	✓	BLEU,CIDEr,METEOR,ROUGE-L
	Fashion-IQ-Cap (Krishna et al., 2017)	60K	✓	✗	-
FIG	VITON-HD (Goyal et al., 2017)	83K	✗	✓	FID, KID
	MGD (Schwenk et al., 2022)	66K	✗	✓	FID,KID,CLIP-S

Table 6: Description of datasets used in two stages.

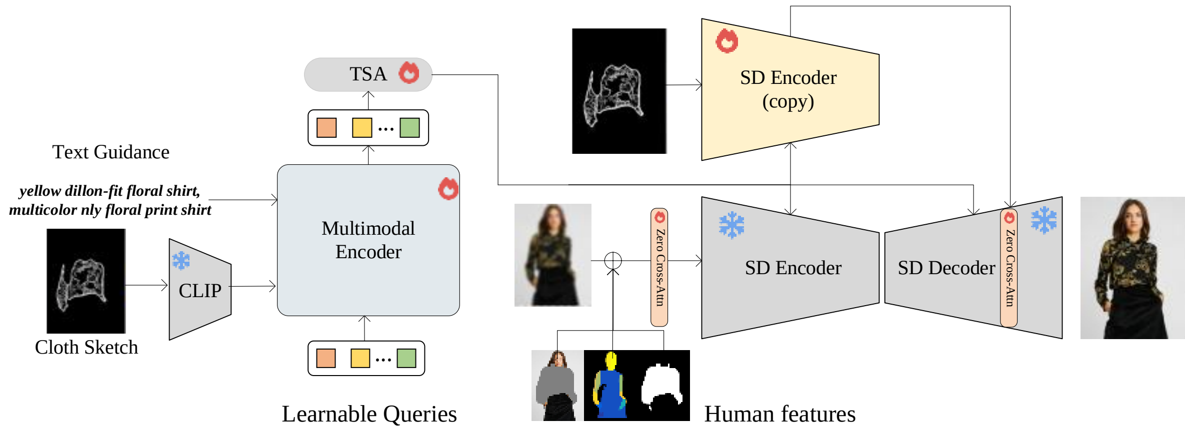

Figure 3: Model's architecture for MGD finetuning, where the diffusion model receives multimodal's output, cloth sketch and human features as input, then generate the target images. We provide the cloth sketch and text guidance as a multimodal input to the encoder, such that the extracted image sketch and text features are more relevant to the ground truth.

the CIR task, which involves the ability to retrieve captions for target images, we have annotated images from the training set of Fashion-IQ. Recognizing that manually annotating all the images would be time-consuming and resource-intensive, we draw inspiration from the success of recent MLLM models such as LLaVA in text-annotation tasks, and propose leveraging LLaVA 1.5 (13B) to semi-automatically annotate the dataset. We perform word lemmatization to reduce each word to its root form. Such pre-processing stage is crucial for the Fashion-IQ dataset, as the captions do not describe a single garment but instead express the properties to modify in a given image to match its target. As shown in Fig. 4, by analysis of the captions in Fashion-IQ, we extracted key words that describe clothing information such as color, sleeve, pattern, lace, etc., as prompts for MLLM (LLaVA 1.5). We then instructed the model to generate the corresponding captions referencing words that match the image features, as shown in Fig. 5. After this process, we got the captions for Fashion-IQ dataset. The trained UniFashion from this dataset (Fashion-IQ-cap) can generate captions for images in the evaluation set of Fashion-IQ to assist in the CIR task.

Figure 4: Vocabulary of the frequent words scaled by frequency for dresses.

D Implementation Details

LLM During the first phase, due to the flexibility brought by the modular architectural design of BLIP-2, we are able to adapt the model to a broad

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 7: Examples of task instruction templates. <image> represents the input image, <question> denotes the question in the VQA and LLaVA 80K dataset, and <photo> is the image description of the input image.

Figure 5: Illustration of Instruction-Following Data. The top section displays an image alongside its original captions from Fashion-IQ dataset. The bottom section presents detailed captions generated by LLaVA-1.5. The original captions are not prompts for generation but are provided for comparison with the newly generated caption.

spectrum of LLMs. In our experiments, in order to effectively utilize the capabilities of the existing MLLM models, we adopted LLaVA-1.5 as the LLM module of the model. Technically, we leverage LoRA to enable a small subset of parameters within UniFashion to be updated concurrently with two layers of adapter during this phase. Specifically, the lora rank is 128 and lora alpha is 256. We utilize the AdamW optimizer with $\beta_0 = 0.9$, $\beta_1 = 0.99$, and weight decay of 0. The LLMs are trained with a cosine learning rate of 2e-5 and a warmup rate of 0.03. We use a batch size of 32 for the tuned LLMs.

Diffusion Module Following, StableVITON, we inherit the autoencoder and the denoising U-Net of the Stable Diffusion v1.4. We initialize our denoising U-Net with the weights of the U-Net from the Paint-by-Example and for more refined person texture, we utilized a VAE fine-tuned on the VITONHD dataset from StableVITON. We train the model using an AdamW optimizer with a fixed learning rate of 1e-4 for 360k iterations, employing a batch size of 32. For inference, we employ the pseudo linear multi-step (PLMS) sampler, with the number of sampling steps set to 50.

E Instruction-Tuning LLMs for Different Caption Style

Liu et al.'s work shows that LLMs have the potential to handle multimodal tasks based on text description of images. Due to the different styles of captions in different fashion datasets, we adopt different instructions to tune the LLM so that it can generate captions of different styles.

We designed different instructions for different datasets and tasks, as shown in Table 7. General instruction template is denoted as follows:
USER: <Img><queries></Img> + Instruction. Assistant: <answer>.

For the <image> placeholder, we substitute it with the output of Multimodal Encoder. To avoid overfitting to the specific task and counteract the model's inclination to generate excessively short outputs, we have devised specific instructions, which enable the LLM to produce concise responses when necessary.