RAGDiffusion: Faithful Cloth Generation via External Knowledge Assimilation

Yuhan Li[1*†], Xianfeng Tan[2*], Wenxiang Shang[2], Yubo Wu[2], Jian Wang[2],
Xuanhong Chen[1], Yi Zhang[1], Hangcheng Zhu[2‡], Bingbing Ni[1‡]
[1]Shanghai Jiao Tong University [2]Alibaba Group
https://colorful-liyu.github.io/RAGDiffusion-page/

Abstract

Standard clothing asset generation involves restoring forward-facing flat-lay garment images displayed on a clear background by extracting clothing information from diverse real-world contexts, which presents significant challenges due to highly standardized structure sampling distributions and clothing semantic absence in complex scenarios. Existing models have limited spatial perception, often exhibiting structural hallucinations and texture distortion in this high-specification generative task. To address this issue, we propose a novel Retrieval-Augmented Generation (RAG) framework, termed RAGDiffusion, to enhance structure determinacy and mitigate hallucinations by assimilating knowledge from language models and external databases. RAGDiffusion consists of two processes: (1) Retrieval-based structure aggregation, which employs contrastive learning and a Structure Locally Linear Embedding (SLLE) to derive global structure and spatial landmarks, providing both soft and hard guidance to counteract structural ambiguities; and (2) Omni-level faithful garment generation, which introduces a coarse-to-fine texture alignment that ensures fidelity in pattern and detail components within the diffusing. Extensive experiments on challenging real-world datasets demonstrate that RAGDiffusion synthesizes structurally and texture-faithful clothing assets with significant performance improvements, representing a pioneering effort in high-specification faithful generation with RAG to confront intrinsic hallucinations and enhance fidelity.

1. Introduction

While the high quality of images generated by diffusion models [20, 45] has significantly lowered the barriers for individuals to engage in content creation, the generation of 2D standard clothing assets from in-the-wild data has been relatively underexplored [43]. Standard clothing asset generation refers to generating forward-facing flat-lay garment images [49] on a clear background by recovering clothing information from arbitrary real-world scenes (*e.g.*, garments hung on hangers, worn by models, or casually laid on chairs in Fig. 1). Standard garments serve as a crucial intermediary variable, connecting various downstream applications such as e-commerce catalogs maintenance, garment design, product marketing and virtual fitting [19, 64, 65].

Figure 1. RAGDiffusion synthesizes structurally and pattern-wise faithful standard garments by assimilating retrieved knowledge.

The garment generation task presents dual technical challenges distinct from conventional virtual try-on paradigms. First, standard clothing generation necessitates *standard structure and frontal positioning* while inherently **limiting stylistic flexibility**, constrained by the narrow manifold of **high-specification garment distributions** [62]. Second, **the complexity of in-the-wild condi-**

* Joint first authors.
†Work done during an internship at Alibaba.
‡Joint corresponding authors.

tions involving occlusions, multi-layer outfits, truncations, or extreme lateral viewpoints notably exacerbates *structural uncertainty and pattern ambiguity*. Collectively, experimental evaluations reveal that existing methods [21, 51] fail to achieve practical usability thresholds with structural and texture distortion, particularly manifesting a pathological ***structure hallucinations*** in *inaccurate structure, fitness and length estimation* under challenging scenes (see Fig. 3 and Fig. 5), which we attribute to the limited spatial reasoning capacity inherent in Stable Diffusion (SD) architectures.

Despite its critical commercial relevance, the garment restoration problem remains insufficiently addressed in literature [43, 49, 51, 57]. Early GAN-based frameworks [57] suffer from pattern distortions constrained by their generative priors. Recent approaches [43, 49, 51] employing pretrained SD exhibit several limitations: (i) inadequate handling of structural hallucinations in hard cases (occluded, multi-layer, *etc*.), (ii) commercially unfaithful logo/texture reproduction which is extremely sensitive to e-commerce sellers. In summary, no existing solution achieves the dual objectives of high-structural integrity and photorealistic texture synthesis under real-world complexity.

We propose RAGDiffusion, as a Retrieval-Augmented Generation (RAG) [38] paradigm including two processes: information aggregation based on retrieval, and conditional generation with omni-level fidelity, to address the issue of structural and texture distortion respectively. (1) **Retrieval-based Structure Aggregation**: Our key insight lies in enhancing ***structural determinacy*** through the assimilating of external structural landmarks and world knowledge, thereby rectifying visual models' misconceptions regarding garment structure, fitness and proportions. During RAG process, contrastive learning [10, 18] is firstly introduced to train a dual tower network to extract multi-modal structure embeddings from images of two branches as well as attributes derived from a frozen large language model (LLM) [1, 2]. Considering potential encoding bias in real-world data, we propose a structure retrieval named Structure Locally Linear Embedding (SLLE) to project the predicted structure embedding towards a standard manifold [41, 56] as well as offer a silhouette landmark. The remapped latent structure embedding and the landmarks provide *soft* and *hard* structure guidance respectively through Embedding Prompt Adapter and Landmark Guider to eliminate structural hallucinations. (2) **Omni-level Faithful Garment Generation**: RAGDiffusion establishes coarse-to-fine texture alignment for ***pattern faithfulness*** and ***detail faithfulness*** for generated garment. Specifically, we ensure the generated pattern matches the conditioning through ReferenceNet [21], a conditioning UNet isomorphic to the denoising UNet. However, problematic detail and logo artifacts persist partially due to resolution limitations of VAE decoders, rendering the generated images commercially unusable for e-commerce sellers, which is an under-addressed issue across prior research [43, 49, 51]. To mitigate reconstruction distortions inherent in original VAE [25], we propose Parameter Gradual Encoding Adaptation (PGEA) to adapt the SDXL [35] backbone to a more powerful VAE.

To the best of our knowledge, RAGDiffusion stands as a pioneering multi-modal RAG method to solve intrinsic hallucination and unfaithfulness during image synthesis. Furthermore, we discover two additional benefits brought by RAG: **zero-shot generalizability** for unseen scenarios via retrieval database expansion, and **human-interpretable control** mechanisms through landmark manipulation. The core contributions of this work are threefold:

- We propose RAGDiffusion, a RAG framework for clothing generation with representation learning and SLLE to offer structure determinacy and eliminate hallucination.
- We employ a coarse-to-fine texture alignment in RAGDiffusion, addressing pattern and detail aspects, ensuring faithfulness of clothing on this high-specification task.
- Comprehensive experiments on challenging in-the-wild sets demonstrate RAGDiffusion is capable of synthesizing both structure and texture faithful clothing assets, outperforming current methods by a substantial margin. Furthermore, the ablation study has validated the effectiveness of different parts of RAGDiffusion.

2. Related works

Retrieval-augmented generative models. Retrieval-augmented strategies leverage external databases to enhance the capabilities of generative models across a variety of tasks. For instance, the RETRO [5] modifies an LLM to effectively utilize external databases. In the realm of image synthesis, retrieval has been employed in both GANs [7, 47] and diffusion models [4, 44], accommodating 3D generation [42], video generation [60] and artistic styles [40]. The essence of these works lies in retrieving similar images to serve as mimetic references for generating specified content, particularly in cases where models are insufficiently trained. Our RAGDiffusion not only builds upon this generalizability through convenient retrieval expansion, but also aggregates structural information, thereby mitigating intrinsic hallucinations on the high-standard tasks.

Controllable text-to-image diffusion models. To attain conditional control in text-to-image diffusion models, ControlNet [59], T2I-Adapter [31] and IP-Adapter [55] incorporate additional trainable modules to fuse conditions on feature maps. Additionally, recent investigations have utilized a variety of prompt engineering techniques [26, 54, 63] and implemented cross-attention constraints [9, 23, 52, 66] to facilitate more controllable generative processes. Furthermore, some studies further investigate multi-condition or multi-modal generation [22, 27, 36, 67]. However, these methods rely on the associative capabilities of stable diffu-

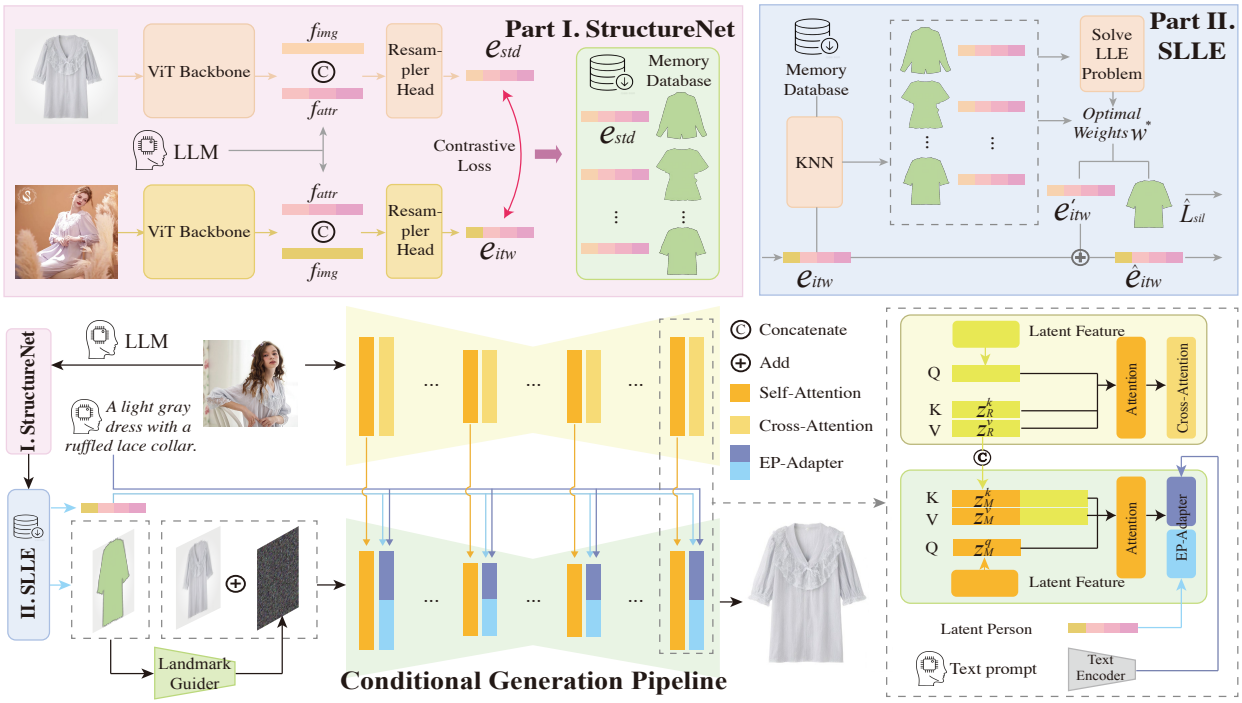


Figure 2. Overall framework of RAGDiffusion. StructureNet provides latent structural embeddings, while SLLE facilitates the embedding fusion along with landmark retrieval. The generative model synthesizes multiple conditions to achieve omni-level high-fidelity generation.

sion, which cannot guarantee fidelity, particularly with the highly standardized sampling of standard garments.

Garment restoration. Standard clothing generation involves restoring flat-lay garment images on from real-world contexts. TileGAN [57] pioneered a two-stage GAN [16] framework, suffering from backbone capability limitations. Recent works including TryOffDiff [49], TryOffAnyone [51], and IGR [43] have adopted pretrained SD as base backbones, yielding obvious improvements through enhanced generative priors. Specifically, TryOffDiff utilizes SigLIP [58] to extract garment features; TryOffAnyone employs direct image concatenation for conditional generation; IGR implements ReferenceNet-inspired [21] architectures for pattern consistency. However, these methods exhibit three critical shortcomings: (1) They ignore structural hallucinations and distortions arising from real-world complexities (multi-layer, extreme viewpoints, *etc*.). (2) Their detail fidelity fails to meet commercial standards. (3) Adaptation to new garment categories/scenarios requires costly paired data and retraining. Our RAGDiffusion systematically resolves these limitations through RAG paradigm and a coarse-to-fine alignment pipeline.

3. Method

An overview of the RAGDiffusion is presented in Fig. 2. The backbone of RAGDiffusion employs the SDXL [39], with the preliminary detailed in Appendix. Given an in-the-wild clothing image $x_{itw} \in \mathbb{R}^{H\times W\times 3}$, RAGDiffusion is aimed to generate an authentic standard flat lay in-shop garment image x_{std}. A dual tower StructureNet extracts latent structure embeddings by contrastive learning as detailed in Sec. 3.1, and SLLE retrieves and fuses structure embeddings and landmarks as in Sec. 3.2. Coarse-to-fine alignment generation involves ReferenceNet,and PGEA for pattern, and decoding faithfulness described in Sec. 3.3.

3.1. Dual-tower embeddings extraction

To eliminate structural hallucinations, a straightforward approach is to extract structural features from images and feed them into generative networks, as seen in StyleGAN [24] and LADI-VTON [30], which use additional latent coding to module conditions. Accordingly, we extract latent structure embeddings through contrastive learning [37], using garments that share similar canny but are texture-dissimilar as training pairs. This process exploits the structural similarities between specialized pairs, which are not emphasized during the conventional training of diffusion.

Contrastive learning. Given a batch of N (in-the-wild clothing x_{itw}, standard clothing x_{std}) pairs, StructureNet g_θ is trained to predict which of the $N \times N$ possible (x_{itw}, x_{std}) pairings across a batch actually occurred according to structure similarity. To do this, StructureNet learns a multi-modal embedding space by jointly training a dual tower encoder tailored for x_{itw} and x_{std} to maximize the cosine similarity of the embeddings of the N real pairs (marked with superscript $+$) in the batch while minimizing the similarity of the $N^2 - N$ incorrect pairs(marked with superscript

$-$). To be specific, the LLM is used to extract 10 types of discrete attributes of clothing (*e.g.*, Category, FitType, CollarTechnique, IfTopTuckIn), which incorporates general in-context image knowledge from the language model [1, 2], *eliminating semantic deficiencies and ambiguities present in vision*. Then we assign a learnable embedding f_{attr} to each discrete attribute. The image structure features f_{img} are extracted by the twin ViT [14] encoder and concatenated with the attributes embeddings f_{attr} to form the final latent structure embeddings e after non-linear Resampler [55] head. We optimize the InforNCE loss [10] over these latent structure embeddings $\{e_{itw}, e_{std}\}$ from $\{x_{itw}, x_{std}\}$ as:

$$\mathcal{L} = \frac{-1}{N}\sum_{i=1}^{N}\log\frac{\exp(e_{itw,i}\odot e^{+}_{std}/\tau)}{\exp(e_{itw,i}\odot e^{+}_{std}/\tau)+\sum_{e^{-}_{std}}\exp(e_{itw,i}\odot e^{-}_{std}/\tau)}, \tag{1}$$

where $\odot$ is cosine similarity between two vectors, N is the batch size, and τ is a temperature scalar.

3.2. Retrieval-augmented SLLE

By integrating structural knowledge from StructureNet into the SD, we have improved the style and structure of generated flat-lay clothing as shown in Sec. 4.3. However, due to the limited spatial perception inherent in the SD [6, 28], it often struggles to accurately represent the length and contours of clothing, particularly in hard cases (occlusions, multi-layer, lateral viewpoints, *etc*.), as in Fig. 5. Furthermore, since StructureNet is trained on a limited set of contrastive pairs, it may perform poorly with out-of-distribution samples, making it unreliable for real-world applications.

We have observed that creating a comprehensive database of standard flat-lay clothing for retrieval is easier than gathering comprehensive in-the-wild samples (numerous scenarios). In this context, we set standard clothing e_{std} as basic vectors and propose Structure Locally Linear Embedding (SLLE), which is a manifold projection [41] to correct each predicted structure embedding e_{itw} into the target space of standard embedding e_{std} to mitigate potential error. Meanwhile, the retrieval database provides structure landmarks to strengthen explicit spatial constraints.

Memory database. To execute SLLE, we establish a retrieval memory database. We first encode the collected standard flat-lay clothing image dataset into a series of latent structure embeddings e_{std} with StructureNet g_θ. The silhouette landmarks L_{sil} are also extracted to designate areas for generated content. Thus an external memory database $\mathcal{D}$ that consists of embedding-landmark pairs (e_{std}, L_{sil}) is obtained. These structure embeddings are utilized in matching algorithms [17], enabling a given in-the-wild garment to find the most compatible standard features e_{std} and landmarks during inference. Notice that landmarks can be any structural figure. Here, we use the outline of clothing, primarily to tackle the challenges of limited spatial perception in SD (*e.g.* length and contour) [6, 28].

Structure LLE algorithm. In this process, retrieval-augmented SLLE drags in-the-wild embedding closer to the standard flat-lay garment embeddings to void outliers during inference, as well as offer silhouette landmarks. Motivated by the successful practice of classic locally linear embedding in [3], we assume that the garment embedding or landmark and its nearby points are locally linear on the manifold, eliminating the need to project them into a higher-dimensional space as vanilla LLE [41] does. Specifically, given an extracted garment embedding e_{itw}, the goal of SLLE is to reconstruct embedding e'_{itw} with standard embeddings e_{std} as basis vectors. We start by searching the K nearest standard embeddings $\{e^1_{std}, ..., e^K_{std}\}$ from the standard garment memory database $\mathcal{D}$ using cosine similarity. Thus K corresponding flat-lay cloth silhouette landmarks $\{L^1_{sil}, ..., L^K_{sil}\}$ are obtained as well. Next, a linear combination of these neighbors is sought to reconstruct e'_{itw} by minimizing the error $\|e'_{itw} - e_{itw}\|_2$, which could be formulated as the least-squared optimization problem:

$$\min \left\| e_{itw} - \textstyle\sum_{i=1}^{K} w_i \cdot e^i_{std} \right\|_2, \quad s.t. \textstyle\sum_{i=1}^{K} w_i = 1, \tag{2}$$

where w_i is the barycentric weight of the i-th nearest embedding e^i_{std}. The optimal weights $\{w_1, \ldots, w_K\}$ can be determined by solving Eq. (2). Subsequently, we can reconstruct the shape embedding as $e'_{itw} = \sum_{i=1}^{K} w^*_i \cdot e^i_{std}$. Ideally, the reconstructed embedding e'_{itw} serves as an in-domain data point that retains structural accuracy and appropriate fitness, albeit with some information loss. In practice, we use a linear combination of original structure embedding e_{itw} and reconstructed one e'_{itw} as the final structure representation during inference:

$$\hat{e}_{itw} = \alpha \cdot \textstyle\sum_{i=1}^{K} w^*_i \cdot e^i_{std} + (1-\alpha)\cdot e_{itw}, \tag{3}$$

where $\alpha \in [0, 1]$ controls the trade-off and is set to be 0.5. The final landmark $\hat{L_{sil}}$ is also fused with optimal weights.

Structure conditioning. Inspired by IP-Adapter [55], we adopt a similar ***Embedding Prompt Adapter (EP-Adapter)*** to condition the high-level structure semantics. While maintaining the text branch unchanged, fused structure embeddings $\hat{e}_{itw}$ after SLLE are fed into additional projection layers to generate key and value matrices for the structure representations. Two parallel cross-attention layers process the *text modality* and the *structure embedding modality*, with results being summed to replace the original single text one: $\text{Attention}(Q, K_{text}, V_{text}) + \text{Attention}(Q, K_{emb}, V_{emb})$. The EP-Adapter enhances the global/inner structural faithfulness by incorporating prior knowledge from LLM and structural training pairs as a soft constraint. Meanwhile, the contour landmark $\hat{L_{sil}}$ involves external spatial structural information. A Landmark Guider [21] with 4 convolution layers (4×4 kernels, 2×2 strides, $16, 32, 64, 128$ channels) is incorporated to align the silhouette mask with z_t as an explicit and hard constrain.

3.3. Omni-Level faithful garment generation

Pattern faithfulness. Inspired by success in human editing [8, 21], we introduce an additional UNet encoder (*i.e*. ***ReferenceNet*** [21]) to precisely preserve the fine-grained details of clothing assets, which is isomorphic to the main generative UNet (*i.e*. MainNet) and shares same initial parameter weights. The latent of in-the-wild garment image is passed through ReferenceNet to obtain the intermediate *key* and *value* features $\{z_R^k, z_R^v\} \in \mathbb{R}^{N\times l\times d}$ in self-attention, which are concatenated with the features $\{z_M^k, z_M^v\} \in \mathbb{R}^{N\times l\times d}$ from the MainNet along the l dimension to obtain the final $\{z_C^k, z_C^v\} \in \mathbb{R}^{N\times 2l\times d}$. Then we compute the self-attention on the concatenated features as:

$$\text{Attention}\left\{z_M^q, z_C^k, z_C^v\right\} = \text{softmax}(\frac{z_M^q z_C^{kT}}{\sqrt{d}})z_C^v, \quad (4)$$

where z_M^q represents *Query* features in self-attention from MainNet. ReferenceNet plays an important role in texture preserving of garments when it has complicated patterns.

Detail faithfulness. The cloth restoration technique primarily serves e-commerce sellers for product catalog maintenance and marketing purposes. The requirement for brand logo fidelity is so stringent that existing methods' results [43, 49, 51] are practically unusable even with ReferenceNet, partially because of the fine-grained logo decoding degradation caused by the high compression ratio of SDXL's VAE, as shown in Fig. 5. Although the VAE in FLUX [15] greatly alleviates the loss by inflating latent channels, challenges intrinsic to the DiT framework [34], such as slow convergence, and high data requirements, hinder its widespread adoption in the community.

To address this, we propose a versatile three-stage ***parameter gradual encoding adaption (PGEA)*** to align the SDXL UNet with the FLUX VAE. Specifically, we expand the channels in *conv in* and *conv out* layers (from 4 to 16) in UNet to match FLUX VAE, enabling direct modification on the SD config file to load adapted weights. Stage I: we focus on distilling the knowledge from the original *conv in* layer to the adapted *conv in* layer. The input image is encoded through different VAEs and passed into the 4 and 16-channel *conv in* layers respectively, applying a reconstruction loss between their output to update the 16-channel *conv in* layer for fast adaption. Stage II: we train the UNet on a standard text-to-image task for 20,000 steps, where only the 16-channel *conv in* and *conv out* layers are updated. Stage III: we train and update the entire UNet on the standard text-to-image task for 200,000 steps. The complete training of PGEA lasts for 4 days on 8 H20 GPUs. It is noteworthy that the adapted general UNet with extremely low decoding loss can be applied to various downstream tasks. Crucially, despite marginal gains in quantitative metrics, RAGDiffusion's breakthrough in brand logo fidelity makes it commercially viable, the first in industrial garment restoration.

Figure 3. RAGDiffusion delivers both loyal structures and superior details on challenging layered and side-view situations.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	DISTS ↓
PBE	0.5278	0.4821	26.55	10.08	0.352
ControlNet	0.5225	0.4792	23.02	9.231	0.345
IP-Adapter	0.6067	0.4728	14.50	3.747	0.264
TryOffDiff	0.6124	0.5007	34.34	14.19	0.335
TryOffAnyone	0.6575	0.4335	18.86	7.919	0.245
ReferenceNet	0.6546	0.3979	10.70	1.399	0.224
RAGDiffusion	**0.6963**	**0.3684**	**9.990**	**1.092**	**0.192**

Table 1. Quantitative comparisons on the STGarment dataset.

4. Experiments

4.1. Experimental setup

Datasets. We collected 65,131 pairs of (in-the-wild upper clothing, standard flat lay upper clothing) for training and 1969 pairs for testing, named STGarment. Among the in-the-wild clothing, there are three main displays: clothing worn on a person, clothing laid indoors, and clothing hung on hangers. We use Qwen2-VL-7B [2] to provide prompt annotations for each pair along with 10 discrete attributes (3.5 seconds per image). Additionally, we matched each pair of images with the most structurally similar flat-lay clothing for StructureNet training based on canny image similarity. We employed a clustering approach to filter and construct a memory database for retrieval, which contains 4,000 embedding-landmark pairs from the training set. Please refer to the Appendix for more details.

Implementation details. We initialize the RAGDiffusion with a pre-trained SDXL model and train it on STGarment using an AdamW optimizer with the learning rate of $5e-5$ at a resolution of 768×768. The models are trained for 5

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	DISTS ↓
Vanilla Model	0.6576	0.3961	10.65	1.405	0.222
+ emb	0.6614	0.3917	10.55	1.354	0.217
+ emb + sil	0.6926	0.3690	10.22	**0.940**	0.210
Generating mask	0.6426	0.4186	14.55	3.570	0.239
+ emb + sil + PGEA (**Full**)	**0.6963**	**0.3684**	**9.990**	1.092	**0.192**

Table 2. Quantitative ablation study of each component.

Dataset	Viton-HD		DC-upper		DC-lower		DC-dress	
Method	FID	KID	FID	KID	FID	KID	FID	KID
PBE	80.8	52.9	45.0	21.1	158.8	128.3	105.1	74.6
ControlNet	69.9	43.2	48.9	23.6	174.7	161.7	118.3	95.4
IP-Adapter	46.8	17.8	28.8	11.1	111.3	62.7	47.8	19.0
TryOffDiff*	14.0	3.31	38.9	16.7	158.7	130.0	74.3	48.6
TryOffAnyone*	**11.4**	1.97	24.4	9.30	114.2	79.6	38.4	20.4
ReferenceNet	15.2	3.23	17.8	6.24	124.5	90.2	51.2	22.4
Ours	12.3	**1.45**	**15.9**	**4.51**	**40.5**	**19.4**	**23.1**	**6.22**

Table 3. Cross dataset evaluation on Viton-HD, DressCode. Model with * means inner-dataset test which is trained on Viton-HD.

Scale	Top-1 Acc. ↑	Top-5 Acc. ↑	IOU ↑
1000	76.6%	93.4%	0.872
2000	78.2%	93.2%	0.885
4000	**80.6%**	**93.7%**	0.902
8000	79.7%	93.4%	**0.905**

Table 4. Retrieval accuracy at different scales of external database.

days on 8 H20 GPUs with DeepSpeed [29] ZeRO-2 to reduce memory usage, at a batch size of 15. $K = 4$ in SLLE in Eq. (2). The StructureNet (*i.e.* embedding encoder) utilizes CLIP-ViT-L/14 image encoder as the backbone and is fine-tuned on nearly 2 million various garment images following DinoV2 [32]. StructureNet is further trained on STGarment with contrastive learning for 4 days on 4 H20 GPUs at a batch size of 128. At inference time, we run RAGDiffusion on a single NVIDIA RTX 3090 GPU for 30 sampling steps with the DDIM sampler [45]. Please refer to the Appendix for more details.

Evaluation protocols. About **generation quality**, we utilize LPIPS [61], SSIM [50] to assess the reconstruction accuracy, DISTS [13] to evaluate image similarity on both perceptual and structural level. Additionally, we employ FID [33] and KID [46] metrics to evaluate the realism and authenticity of the generated distributions. In terms of **retrieval ability**, we evaluate the top-1 accuracy and top-5 accuracy of retrieval results from the memory database $\mathcal{D}$ across 1969 test samples, alongside the average Intersection over Union (IoU) of the corresponding silhouette masks with GT ones. Given the absence of GT masks of test samples in the memory database $\mathcal{D}$, we define the retrieved masks to be accurate if their IoU exceeds 0.85.

Baselines. We compare RAGDiffusion with recent works TryOffDiff [49] and TryOffAnyone [51] with official checkpoints. We also train 4 classic conditional generation methods on STGarment with SDXL backbone for fair comparison: IP-Adapter [55], ReferenceNet [21], ControlNet [59] and Paint-by-Example [53] (PBE) as baselines.

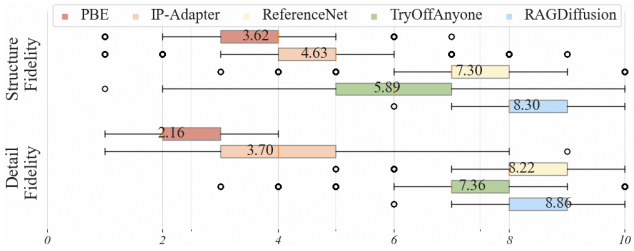


Figure 4. Boxplot illustration of user study. RAGDiffusion demonstrates better performance and stability across scenes.

4.2. Generation comparisons with baselines

Qualitative results. Fig. 3 presents a qualitative comparison between RAGDiffusion and baselines on the STGarment dataset. TryOffDiff has yielded entirely unsuccessful outputs. Meanwhile, naive ReferenceNet and TryOffAnyone manage to maintain correct textures in challenging scenarios, but encounter structural confusion and distortion. This phenomenon may stem from TryOffDiff and ReferenceNet's over-reliance on local visual texture features, lacking the benefits of a global perspective in challenging cases. RAGDiffusion employs a retrieval-aggregate approach to capture structural information and integrates conditional controls from EP-Adapter, Landmark Guider and PGEA, yielding promising performance in both structurally faithful and detail-oriented garment conversion. Notably, while trained on the same dataset with naive ReferenceNet, RAGDiffusion assimilates high-quality contour landmarks and structure embeddings as external prior to producing visually compelling results that enhance depth and realism.

Quantitative results. As indicated in Tab. 1, we quantitatively evaluate the *generation quality* of various methods on the STGarment dataset, demonstrating that RAGDiffusion significantly outperforms all baseline approaches. This confirms RAGDiffusion's ability to deliver superior and accurate garment generation across diverse scenarios. Retrieval, as a crucial component of RAGDiffusion for assimilating external knowledge, plays a significant role. Using the evaluation metrics described in Sec. 4.1, we report the *retrieval accuracy* at different sizes of the external memory database $\mathcal{D}$ in Tab. 4. Specifically, we employ clustering and downsampling algorithms to iteratively eliminate outliers and highly similar samples from the original memory database, constructing retrieval libraries of four different scales: 1000, 2000, 4000, and 8000 samples. When the memory database $\mathcal{D}$ is too large, outliers may adversely affect quality if embeddings are inaccurate, while redundant samples reduce retrieval efficiency. Conversely, if the memory database $\mathcal{D}$ is too small, the retrieval library may lack sufficient representativeness and completeness, adversely affecting the generation performance for specific categories. Notably, a memory database comprising 4000 samples achieved the best trade-off in performance. Please refer to the Appendix for more technical details.

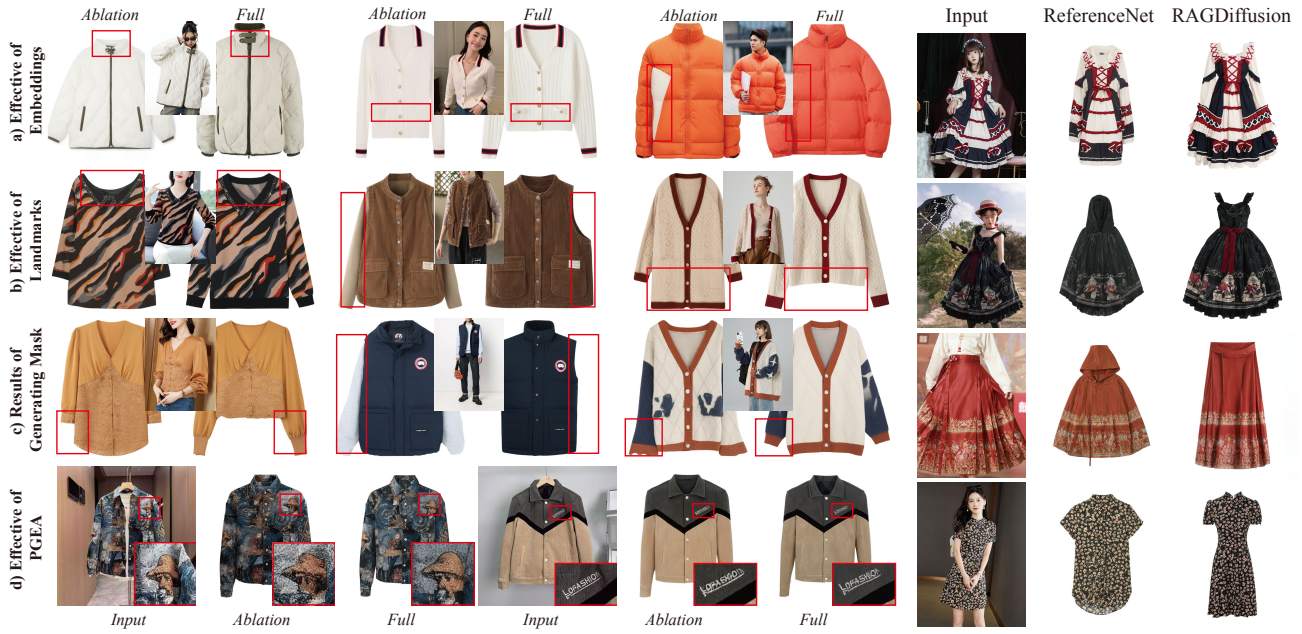


Figure 5. Ablation study. Without landmarks, RAGDiffusion suffers inaccurate shape. Embeddings from StructureNet improve inner structure, while PGEA enhances detail preservation.

Figure 6. RAG significantly improves generalization on untrained categories.

User study. When metrics like FID assess the realism and authenticity of standard garment generation, they tend to be less sensitive to high-frequency information such as contours, collar variations, buttons, and color shifts [11]. We conducted user studies involving 100 participants to evaluate different methods based on user preferences across 200 randomly selected image pairs from the test set. Participants were asked to assign a preference score (ranging from $1 \sim 10$) in terms of structural fidelity and detail fidelity for each sample generated by anonymized methods. As illustrated in Fig. 4, we report the distribution of the scores, encompassing medians, means, quartiles, and outliers. Our findings reveal that RAGDiffusion is significantly preferred over all baseline methods in terms of structural fidelity ($mean = 8.34$) and detail fidelity ($mean = 8.52$). Additionally, the narrow range of outliers in our method indicates a more stable generation across various scenes.

4.3. Ablation study

To validate our core contributions, we define a Vanilla Model that removes structure embedding $\hat{e}_{itw}$, silhouette landmarks $\hat{L}_{sil}$ and PGEA. Specifically, its $\hat{e}_{itw}$ and $\hat{L}_{sil}$ are set to zero vectors while using original SDXL VAE.

Soft guidance of structure embeddings. Since structure embeddings serve as prerequisites for retrieval in SLLE and inputs to EP-Adapter, we first add structure embeddings $\hat{e}_{itw}$ into EP-Adapter of vanilla model (denoted as "+ emb"). As in Tab. 2 and Fig. 5, vanilla model lacking structure embedding frequently exhibits artifacts in clothing internal structures, such as inaccurate collars, missing pockets, and occlusion confusion. This demonstrates that structure embedding provides global guidance and plays a fundamental role in optimizing apparel internal structures.

Hard guidance of silhouette landmarks. The core of RAG is to retrieve silhouette landmarks that enhance explicit spatial constraints. Thus we add landmarks on "+ emb" version as the ablation named "+ emb + sil" in Tab. 2. The model without retrieval exhibits significant contour errors, particularly around the back collar, sleeve length, and garment length. This is attributed to the fact that retrieval-augmented SLLE incorporates priors from LLM, which transcend visual limitations and aid in accurate spatial structures through the use of landmarks. Furthermore, we investigate the role of SLLE in embedding remapping in Sec. 4.5.

Alternative to RAG: generating masks via UNet. As RAG requires a memory database and contrastive learning for embeddings, we try a simpler alternative by employing an SDXL UNet to directly predict landmarks. Specifically, we modify the input of the alternative UNet to a three-channel in-the-wild image x_{itw}, with outputs being one-channel predicted landmarks L_{sil} that are directly fed into RAGDiffusion's denoising UNet. However, this alternative produces landmarks with noticeable errors and degraded generation quality, due to limitations in visual understanding and the absence of proper pre-trained weights.

PGEA. PGEA is employed to mitigate information loss during the VAE encoding and decoding process. We add PGEA on "+ emb + sil" version as Full Model. The full model utilizing PGEA achieves clearer and more accurate edges in high-frequency patterns simulation; additionally, it shows significant improvements in the recovery of logos and text. PGEA greatly alleviates the long-standing issue of distortion in detail and makes RAGDiffusion commercially

viable, the first in industrial garment restoration.

4.4. Additional rationale for introducing RAG

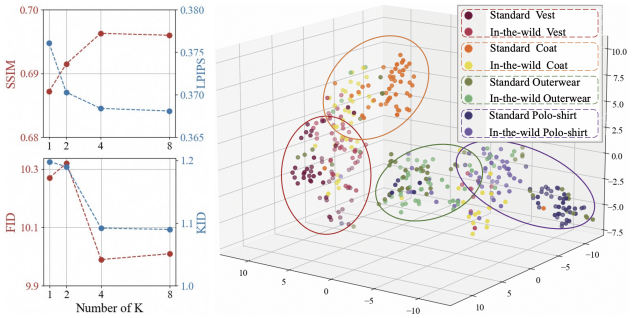
In addition to structural faithfulness, RAGDiffusion demonstrates **excellent generalizability and robustness** across untrained categories and scenes simply by updating retrieval database, without the need for collecting expensive paired data and retraining. Besides, please refer to Appendix for another rationale of **human-interpretable control**.

Cross category evaluation. RAGDiffusion is trained on upper-body clothing data and has not encountered lower-body garments during training. To validate the generalizability, we collect 856 flat-lay lower-body/dress images (no in-the-wild image is needed) and incorporate them with silhouette masks into the external memory database through StructureNet. Subsequently, we test it on 50 lower-body in-the-wild clothing samples, using a ReferenceNet as baseline. The results in Fig. 6 demonstrate that retrieval significantly improves generalization capabilities, serving as a cost-effective zero-shot generalizing solution. The quantitative results on lower-body/dress categories in Tab. 3 also validate the generalizability boost.

Cross dataset evaluation. Despite zero exposure to the Viton-HD [12] and DressCode [19] datasets during training, RAGDiffusion shows strong out-of-distribution (OOD) compatibility over baselines without tuning illustrated in Tab. 3. This boost in generalization increases the operational maturity of RAGDiffusion, enabling it to effectively handle various OOD images submitted by users.

4.5. Component analysis

Latent structure embedding distributions. We further visualize the learned latent embedding distribution in Fig. 7(b) to gain an in-depth comprehension of how the embeddings work in retrieval and the EP-Adapter. For this purpose, we sample $N = 200$ pairs of (in-the-wild cloth embedding e_{itw}, standard cloth embedding e_{std}) pairs from our dual-tower StructureNet. We then employ t-SNE [48] to project each feature representation into a point within Fig. 7. Notably, 1) the dual-branch embeddings (e_{itw}, e_{std}) corresponding to the same category exhibit clustering behavior, indicating that the learned priors in StructureNet effectively perform structural alignment cross domains, thus facilitating retrieval based on structural similarity; and 2) representations from different categories are clearly dispersed, demonstrating that the embeddings encompass discriminative structural information, which aids in generation through the EP-Adapter. Lastly, we also visualize the distribution of cosine similarity between a given sample and images from the external memory database $\mathcal{D}$ in Fig. 8. The cosine similarities present a normal distribution, affirming that the constructed embedding library is representative and comprehensive. Furthermore, the examples visualized for different similarity retrieval results effectively illustrate the efficacy of the landmark retrieval mechanism.



(a) Embedding remapping sensitivity to K in Eq.(3)

(b) Visualization of embedding distribution with t-SNE. Observe intra-class dual-tower coherence and inter-class divergence.

Figure 7. Illustrations of component analysis.

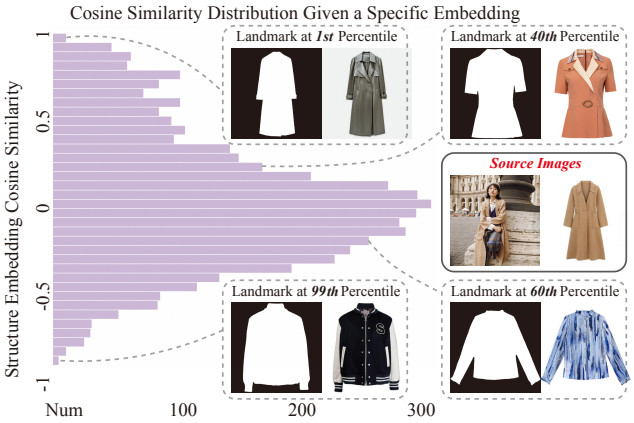


Figure 8. Distribution of cosine similarity between a given sample and images from the external memory database.

Number of retrieved embeddings. As the number K of retrieved nearest neighbors in Eq. (2) during SLLE plays a fundamental role in the properties of the final structure embeddings and landmarks, we demonstrate the fusion sensitivity to K in Fig. 7(a). A smaller K value facilitates the sharp edge of landmarks, whereas a larger K value enhances the accurate representation of the reconstructed embedding e'_{itw} in Eq. (3). It can be observed that the optimal performance occurs at $K = 4$, which also serves as the default parameter setting for other experimental parts.

5. Conclusion

In conclusion, RAGDiffusion presents a significant advancement in the generation of standardized clothing assets by effectively addressing the prevalent challenges of structural hallucinations and texture fidelity. By integrating retrieval-based structure aggregation to get soft and hard guidance, as well as omni-level faithful generation pipeline, RAGDiffusion is capable of synthesizing both structure and texture faithful clothing assets. Comprehensive experiments and in-depth analysis demonstrate notable performance improvements over existing models. We hope that RAGDiffusion helps open avenues for RAG in diverse high-specification faithful generation tasks, bringing us closer to the goal of professional content creation for all.

Acknowledgements

This work is supported by the Science and Technology Commission of Shanghai Municipality under research grant No. 25ZR1401187.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4, 5

[3] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 4

[4] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. 2

[5] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022. 2

[6] Ali Borji. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing*, 137:104771, 2023. 4

[7] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021. 2

[8] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*, 2023. 5

[9] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4

[11] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *arXiv preprint arXiv:2406.07547*, 2024. 7

[12] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 8

[13] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 6

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Black forest labs. Flux.1-dev. https://github.com/black-forest-labs/flux, 2024. 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 2014. 3

[17] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003. 4

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[19] Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. Dresscode: Autoregressively sewing and generating garments from text guidance. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024. 1, 8

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020. 1

[21] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2, 3, 4, 5, 6

[22] Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. Cocktail: Mixing multi-modality control for text-conditional image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[23] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023. 2

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[27] Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. *arXiv preprint arXiv:2405.18172*, 2024. 2

[28] Qihao Liu, Adam Kortylewski, Yutong Bai, Song Bai, and Alan Yuille. Intriguing properties of text-guided diffusion models. *arXiv preprint arXiv:2306.00974*, 2, 2023. 4

[29] Microsoft. Deepspeed. https://github.com/microsoft/DeepSpeed, 2023. 6

[30] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the ACM International Conference on Multimedia*, 2023. 3

[31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[32] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 6

[33] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[36] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 3

[38] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 3

[40] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022. 2

[41] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290 (5500):2323–2326, 2000. 2, 4

[42] Junyoung Seo, Susung Hong, Wooseok Jang, Inès Hyeonsu Kim, Minseop Kwak, Doyup Lee, and Seungryong Kim. Retrieval-augmented score distillation for text-to-3d generation. *arXiv preprint arXiv:2402.02972*, 2024. 2

[43] Le Shen, Rong Huang, and Zhijie Wang. Igr: Improving diffusion model for garment restoration from person image. *arXiv preprint arXiv:2412.11513*, 2024. 1, 2, 3, 5

[44] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 2

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 6

[46] JD Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference for Learning Representations*, 2018. 6

[47] Hung-Yu Tseng, Hsin-Ying Lee, Lu Jiang, Ming-Hsuan Yang, and Weilong Yang. Retrievegan: Image synthesis via differentiable patch retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 242–257. Springer, 2020. 2

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[49] Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Tryoffdiff: Virtual-try-off via high-fidelity garment reconstruction using diffusion models. *arXiv preprint arXiv:2411.18350*, 2024. 1, 2, 3, 5, 6

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 6

[51] Ioannis Xarchakos and Theodoros Koukopoulos. Tryoffanyone: Tiled cloth generation from a dressed person. *arXiv preprint arXiv:2412.08573*, 2024. 2, 3, 5, 6

[52] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained

diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 2

[53] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6

[54] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 2

[55] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 4, 6

[56] Zhenhui Ye, Jinzheng He, Ziyue Jiang, Rongjie Huang, Jiawei Huang, Jinglin Liu, Yi Ren, Xiang Yin, Zejun Ma, and Zhou Zhao. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*, 2023. 2

[57] Wei Zeng, Mingbo Zhao, Yuan Gao, and Zhao Zhang. Tilegan: category-oriented attention-based high-quality tiled clothes generation from dressed person. *Neural Computing and Applications*, 32:17587–17600, 2020. 2, 3

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 6

[60] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 2

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 6

[62] Shiyue Zhang, Zheng Chong, Xujie Zhang, Hanhui Li, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. Garmentaligner: Text-to-garment generation via retrieval-augmented multi-level corrections. *arXiv preprint arXiv:2408.12352*, 2024. 1

[63] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023. 2

[64] Xujie Zhang, Yu Sha, Michael C Kampffmeyer, Zhenyu Xie, Zequn Jie, Chengwen Huang, Jianqing Peng, and Xiaodan Liang. Armani: Part-level garment-text alignment for unified cross-modal fashion design. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4525–4535, 2022. 1

[65] Xujie Zhang, Binbin Yang, Michael C Kampffmeyer, Wenqing Zhang, Shiyue Zhang, Guansong Lu, Liang Lin, Hang Xu, and Xiaodan Liang. Diffcloth: Diffusion based garment synthesis and manipulation via structural cross-modal semantic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23154–23163, 2023. 1

[66] Xuanpu Zhang, Dan Song, Pengxin Zhan, Qingguo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Anan Liu. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. *arXiv preprint arXiv:2408.06047*, 2024. 2

[67] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2