

# LOTS of Fashion!

## Multi-Conditioning for Image Generation via Sketch-Text Pairing

Federico Girella<sup>1</sup> Davide Talon<sup>2</sup> Ziyue Liu<sup>1,3</sup> Zanxi Ruan<sup>1</sup>  
 Yiming Wang<sup>2</sup> Marco Cristani<sup>1,4</sup>

<sup>1</sup>University of Verona, <sup>2</sup>Fondazione Bruno Kessler,  
<sup>3</sup>Polytechnic Institute of Turin, <sup>4</sup>University of Reykjavik

<https://intelligolabs.github.io/lotsof/>

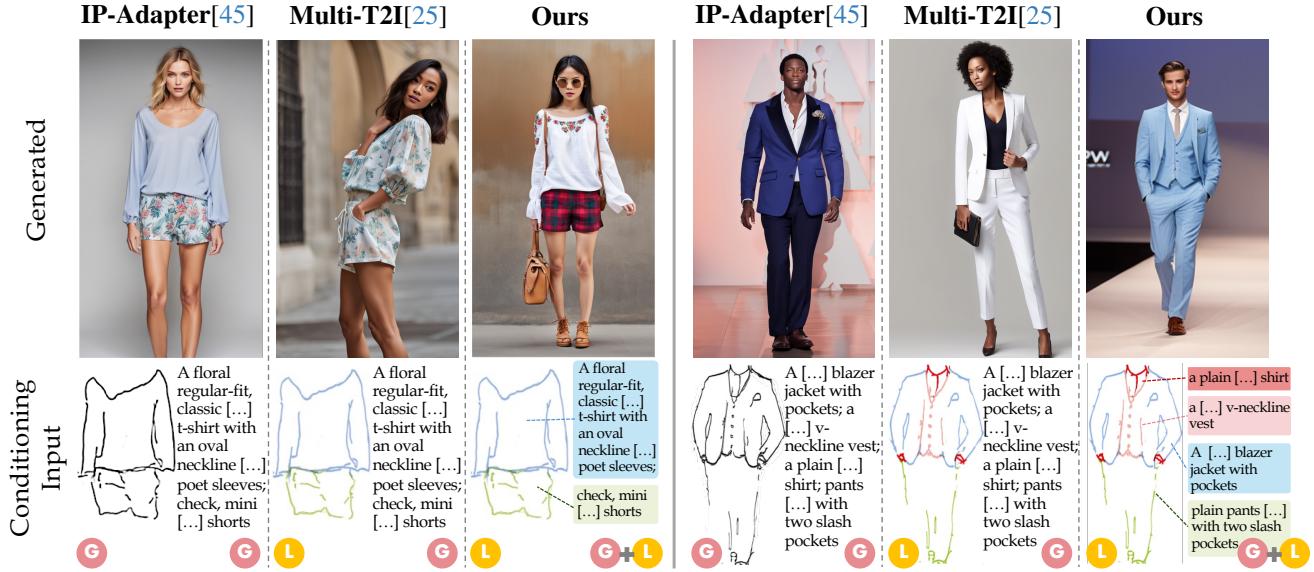


Figure 1. We present LOTS, enabling fashion image generation with an unprecedented level of control. LOTS represents the natural evolution of fashion design methodologies, progressing from global text and sketches (IP-Adapter [45]) to localized sketches with global text (Multi-T2I [25]). Our approach leverages a global description (omitted here for brevity) alongside a set of localized sketch-text pairs (the coloured boxes), effectively defining both the layout and appearance of individual garment items.

## Abstract

*Fashion design is a complex creative process that blends visual and textual expressions. Designers convey ideas through sketches, which define spatial structure and design elements, and textual descriptions, capturing material, texture, and stylistic details. In this paper, we present Localized Text and Sketch for fashion image generation (LOTS), an approach for compositional sketch-text based generation of complete fashion outlooks. LOTS leverages a global description with paired localized sketch + text information for conditioning and introduces a novel step-based merging strategy for diffusion adaptation. First, a Modularized*

*Pair-Centric representation encodes sketches and text into a shared latent space while preserving independent localized features; then, a Diffusion Pair Guidance phase integrates both local and global conditioning via attention-based guidance within the diffusion model's multi-step denoising process. To validate our method, we build on Fashionpedia to release Sketchy, the first fashion dataset where multiple text-sketch pairs are provided per image. Quantitative results show LOTS achieves state-of-the-art image generation performance on both global and localized metrics, while qualitative examples and a human evaluation study highlight its unprecedented level of design customization.*

## 1. Introduction

In fashion design, designers need to express their abstract inspirations through forms that are natural to humans, *e.g.*, sketches or natural language. For example, a designer can sketch the shape of a t-shirt, with its design idea expressed in natural language description “*a floral, regular-fit, classic t-shirt with an oval neckline and wrist-length poet sleeves*” or the shorts as “*check, mini, symmetrical, gathering shorts with a regular fit*”, as shown in Fig. 1 (left). Sketches and natural language descriptions associated with the same garment convey complementary information for depicting the final design: sketches provide spatial information for the garment outline, such as its fit (*e.g.*, the tight fit jacket in Fig. 1 (right)), while the language description conveys richer details in terms of accessories and styles (*e.g.*, “*a blazer jacket with a pocket*”). Since a complete design is composed of several clothing garments, multiple natural descriptions are often collected together to outline an outfit. Each description specifies a *localized part* of the design, in terms of silhouette shapes, materials, and textual details, allowing fine-grained localized control over the generation.

In this paper, we aim to facilitate the process of concretizing creative ideas into graphical output, as expressed in the form of localized conditions, *i.e.*, sketch-text pairs, to support the designing process for the fashion industry. We frame this problem as a conditional image generation task, where the conditioning consists of a set of localized text-sketch pairs. In addressing such multi-localized-conditioning generation, the state-of-the-art methods fall short. Recent diffusion adapters for sketch-to-image generation [25, 37, 45, 46] allow for multi-spatial conditions but underperform when providing fine-grained textual information, such as neckline type, pocket/button style, as shown in Fig. 1 (Multi-T2I [25]). We argue that this limitation stems from the use of a *single global description* to inject textual condition: all relevant details about different parts of the garment are considered in a monolithic fashion, leading to incorrect localization of attributes to parts [25, 37, 45, 46]. We refer to this problem as “attribute confusion”, where properties of one item are instead generated for another item, *e.g.*, “*a floral t-shirt with check short*” could bring the floral pattern in the shorts Fig. 1 (left).

We propose LOcalized Text and Sketch for fashion image generation (LOTS), a novel approach leveraging multiple localized sketch-text pairs for image conditioning. First, the *Modularized Pair-Centric Representation* module independently encodes sketches and text into a shared latent space, preserving localized features while limiting information leakage between pairs to reduce attribute confusion. Then, the *Pair-former* merges text and sketch information within each pair, ensuring spatially grounded descriptions that capture single-item attributes through the structural guidance of sketches. Next, during the *Diffusion Pair Guidance* phase, these localized representations serve as conditioning inputs to a pre-trained diffusion model, alongside a global textual representation specifying general appearance properties (style, background). Unlike prior methods, our approach defers condition inputs merging to the diffusion process itself, breaking down the task across multiple denoising steps via a cross-attention strategy.

For model training and comparative evaluation, we introduce Sketchy, a new dataset built on Fashionpedia [13] for localized sketch-to-image generation. In this dataset, garments within an outfit are enriched with sketches and treated as multiple conditioning inputs, each paired with fine-grained descriptions. LOTS achieves state-of-the-art performance in image quality and attribute localization, as evidenced by quantitative metrics (+3.4% GlobalCLIP wrt fine-tuned ControlNet), human evaluation studies (+3.1% F1 Score over SDXL), and qualitative analysis.

**Our contributions** are four-fold:

- We focus on localized sketch-text image generation, advancing state-of-the-art conditioning with localized sketches and a single global text.
- We introduce a novel method, LOTS, to mitigate attribute confusion by modularized attention-based processing per sketch-text pair, and deferring the multi-conditioning merge to the denoising process.
- We introduce a new dataset, Sketchy, in the fashion domain, to facilitate model training and evaluation for the localized text-sketch image generation problem.
- LOTS achieves state-of-the-art performance in sketch-text conditioning and attribute localization, as measured with both metrics and human evaluation.

## 2. Related Works

In this section, we focus on text-to-image, sketch-to-image, and controllable diffusion-based generation.

**Text-to-Image Generation.** Recent advances in Text-to-Image (T2I) generation have been driven by diffusion models [7, 8, 36], which generate high-quality images from textual prompts [27, 32, 34, 35] through a noise-based forward and denoising reverse process. Conditioning techniques enhance control: GLIDE [27] employs classifier-free guidance, DALLE-2 [32] uses a two-stage CLIP-based approach, and Imagen [35] integrates large-scale language models for improved realism and semantic alignment. Stable Diffusion (SD) [34] refines conditioning via cross-attention while optimizing efficiency through latent-space diffusion. Building on SD-like models, we extend control beyond textual descriptions, focusing on multiple sketch-text pairs for localized and fine-grained conditioning.

**Sketch-to-Image Generation.** Sketch-to-Image generation has evolved from GAN-based methods [4, 12, 16, 22, 33] to pre-trained diffusion models [24, 41, 42]. On this line, PITI [42] maps sketches to semantic latents of a large-

scale diffusion model, while SDEdit [24] employs sketch perturbation and denoising to guide generation. Finally, LGP [41] enforces consistency between noisy features and spatial sketch guidance. Recent approaches for sketch-to-image generation have implemented different techniques to control the downstream diffusion model that address the problem of spatial conditioning [25, 45, 46], sketch abstraction [17, 26], or professional sketches [43]. Unlike prior work [17, 24–26, 41–43, 45, 46], which conditions on global sketches, we introduce localized sketch control.

**Controllable diffusion-based generation.** While textual prompts enable high-quality image generation in T2I models, they often lack fine-grained control. To enhance controllability, various methods integrate additional conditioning elements [10, 20, 37, 45–47], including bounding boxes [20], blobs [28], and segmentation masks [5, 14]. GLIGEN [20] conditions the diffusion model with bounding box coordinates to localize textual information, but does not allow for paired sketch-text localization. ControlNet [46] introduces zero-convolution modulation in a frozen diffusion model, while subsequent works propose multi-modal control [9] and all-in-one control adapters [47], though both rely on fixed-length input channels. AnyControl [37] enables multi-control conditioning but requires a trainable copy of the diffusion model. Alternative methods, such as T2I [25] and IP [45] adapters, employ residual feature maps and cross-attention, respectively, to aggregate multiple control signals before conditioning. However, these approaches depend on global textual prompts and are limited by the 77-token constraint of text encoders. In contrast, we couple localized textual descriptions with their corresponding sketches to improve fine-grained generation. Our adapter enables a pre-trained T2I diffusion model to condition on a variable number of sketch-text pairs while remaining lightweight to train. Furthermore, recent approaches [1, 44] employ image editing in the Fashion domain. Multimodal Garment Designer[1] requires a starting image as input for the edit, while LOTS performs image generation from scratch. HieroFashDiff [44] is a concurrent work presenting a two-stage pipeline performing generation and iterative editing. Differently, LOTS performs one-shot generation and allows for additional sketch conditioning. [Not sure about this] Since direct comparison would require non-trivial modifications, risking unfair or misleading results, we will not compare to editing-based approaches.

### 3. Method

In this section we present LOTS: *LOcalized Text and Sketch for fashion image generation*. We start with a formalization of the task, defining the input-output of the problem, and later introduce the key modules of our adapter strategy.

**Localized sketch-to-image generation.** The input and desired output of the problem are as follows. **Input:** Let

$\mathcal{C}$  be a set of sketch-text pairs  $\mathcal{C} = \{C_1, \dots, C_{N_i}\}$  where  $C_i = (S_i, T_i)$  denotes the  $i$ -th sketch-text pair with sketch conditioning  $S_i \in \{0, 1\}^{H \times W}$  and textual description  $T_i$ , with  $N_i$  the number of pairs associated with the  $i$ -th sample and  $H, W$  denoting the width and height of the desired generated image, respectively. We assume that the provided sketches have a globally coherent spatial layout. In other words, local sketches associated with an item should satisfy the constraint  $\sum_i^N S_i = S$ , where  $S$  is the global sketch of the desired image. To enable global information conditioning in natural language, we further allow for the global textual representation  $T_g$  to prompt the model with general appearance information that the desired image should have, e.g., the fashion style or the background. **Output:** Localized sketch-to-image generation aims to synthesize with the generative model  $\phi$  an image  $X \in \mathbb{R}^{3 \times H \times W}$  based on the conditioning input  $\mathcal{C}$  as  $X = \phi(\mathcal{C}, T_g)$ . The resulting generation should faithfully preserve the spatial constraints of sketches and the semantic details of text descriptions while ensuring global coherence between all localized pairs and global description. In particular, the specified properties of the textual description  $T_i$  associated with the  $i$ -th item should be correctly reflected in the location described by its associated spatial information  $S_i$ , while not leaking to other local parts of the image  $S_j$ , with  $i, j = 1..N, j \neq i$ .

**Method overview.** LOTS enables sketch-text conditioned image generation through a two-phase process. An illustration of the proposed approach is presented in Fig. 2. In the *Modularized Pair-Centric Representation* phase (Section 3.1), sketch and text inputs are independently encoded using pre-trained modality-specific encoders and then projected into a multi-modal shared latent space via the *Pair-Former module*, ensuring localized feature extraction without interference between pairs. To address *attribute confusion* when merging multiple conditioning inputs, the *Diffusion Pair Guidance* phase (Section 3.2) employs attention-based conditioning within the multi-step denoising process, integrating both local and global conditioning effectively.

#### 3.1. Modularized Pair-Centric Representation

Localized sketch-to-image generation requires generating an image starting from the set of local conditionings  $\mathcal{C} = \{C_1, \dots, C_2\}$  while ensuring no semantic information from  $C_i$  is leaking to unrelated parts of the image. To address this, we propose to independently process the input information in a pair-centric fashion: each pair is considered independent, or in other words, pairs do not influence or see each other. Consider a single localized representation in the form of a sketch-text pair  $C_i = (S_i, D_i), i = 1..N$ . We embed conditioning information via modality-specific en-

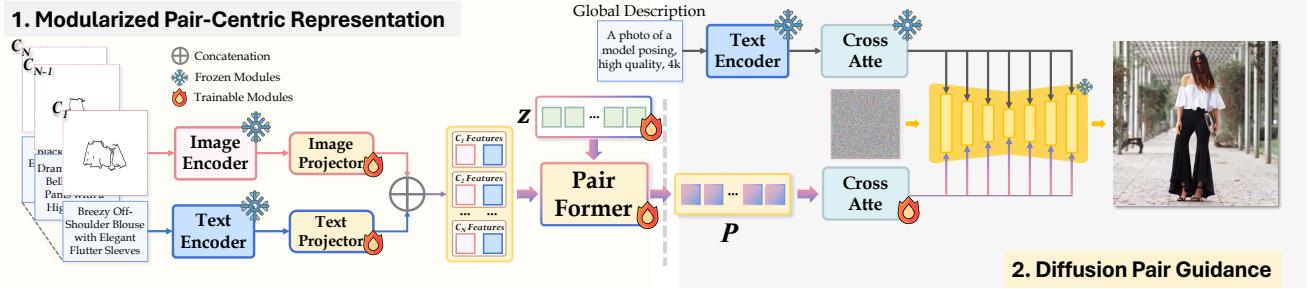


Figure 2. LOTS mitigates attributes confusion building on paired sketch-text conditioning for image generation. **1.** In an initial phase, the modularized Pair-Centric Representation (Sec. 3.1) independently processes available pairs by first embedding the different modalities with pre-trained modality-specific encoders, and later localizing the semantic textual information according to the associated sketch structure in the Pair-Former. **2.** In the second Diffusion Pair Guidance phase (Sec. 3.2), pair representations are directly injected into the downstream diffusion model. By breaking down the merge task within the denoising diffusion steps, LOTS avoids explicit merge of pair representations that lead to attribute confusion.

coders:

$$h_i^T = f^T(T_i) \quad (1)$$

$$h_i^S = f^S(S_i) \quad (2)$$

where  $f^T$  and  $f^S$  denote respectively the text and sketch encoders, and  $h_i^T, h_i^S$  their associated output latent representations,  $i = 1..N$ .

**Pair-Former.** Starting from the modality-specific representations, for each pair, we integrate the sketch spatial guidance and the text semantic information into a shared feature space, where textual information is localized according to the sketch structure. Inspired by recent vision-language advancements [18], we devise a new self-attention strategy aiming to (i) compress the sparse sketch representation in a limited number of tokens and (ii) merge multi-modal information. Let  $z \in \mathbb{R}^{k \times d}$  be a set of  $k$   $d$ -dimensional learnable tokens prepended to the concatenation of visual and textual embedding representations. We obtain the pair tokens by first computing the self-attention with sketch and text representations:

$$h_i = \text{Self-attn}([z; h_i^S; h_i^T]), \quad (3)$$

and restricting  $p_i$  to be the first  $k$  tokens of  $h_i$ , i.e., the output tokens associated to  $z$ , where  $[\cdot; \cdot]$  denotes the concatenation operation along the token dimension. Hence, self-attention allows the prepended token  $z$  to pool the informative content from both the sketch and textual modalities in a coordinated manner. We remark that while the  $z$  learnable tokens are shared between the different pairs, the Pair-former is independently processing the  $N$  input pairs.

### 3.2. Diffusion Pair Guidance

Merging information from multiple conditioning pairs is challenging, as it requires effective coordination while

avoiding information leakage. Existing approaches encode multiple guidance signals in a single pooling step, which may cause interference among pairs. In contrast, we defer the merging process to the pre-trained diffusion model, leveraging its iterative denoising steps to progressively integrate conditioning pairs, rather than combining them in a single step. Given  $P \in \mathbb{R}^{N \times k \times d}$  as the concatenated sequence of the  $N$  pooling tokens  $p_i, i = 1..N$  extracted by our Pair-Former, we avoid explicitly merging the pairs. Instead, we inject the entire sequence into the diffusion model, allowing the merge operation to occur throughout the entire diffusion process. This allows for more dynamic interaction between the multiple given conditionings during the reverse diffusion process. Inspired by previous work [45], we rely on cross-attention layers to guide diffusion generation and introduce an additional learnable cross-attention layer  $\hat{w}$  after each pre-existing cross-attention layer  $w$  in the frozen denoising model. These new layers inject the conditioning sequence  $P$  into the model features at each diffusion timestep, allowing for the iterative merging of information. Let  $x$  denote the input features of a global text-conditioning cross-attention layer in the denoising network, the conditioned output  $x'$  of the paired cross-attention layers is computed as:

$$x' = w(x, h^{T_g}) + \alpha \hat{w}(x, P), \quad (4)$$

where  $w(\cdot, \cdot)$  represents the cross-attention between the two input token sequences and  $h^{T_g}$  denotes the embedding representation of the global text prompt guiding the model with semantic information that should be globally represented, e.g., the style or the background. Here,  $\alpha$  is a scaling hyperparameter, constrained to the range  $[0, 1]$ , that regulates the strength of the guidance from  $P$ . During training, we set  $\alpha$  to 1 to fully enable the cross-attention layers  $\hat{w}$  to learn the appropriate merging behavior. Notably, the attention-based nature of these adapter blocks allows an arbitrary number of

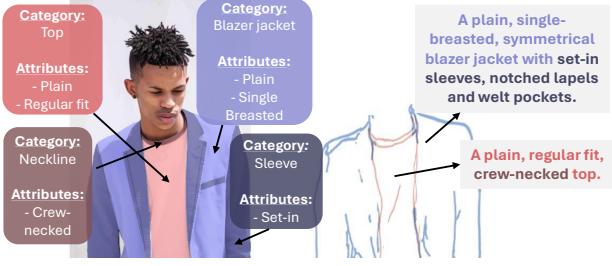


Figure 3. Example of the hierarchical structure of Sketchy. Starting from whole-body item (light colors) and garment parts (dark shades) annotations, we build a hierarchical structure by pairing the garment part annotations to their related whole-body garment. Then, we use this structure to generate garment-level sketches and natural language descriptions by relying on off-the-shelf models.

conditioning tokens, enabling LOTS to work with a variable number of conditioning pairs.

## 4. Experiments

We assess LOTS by comparing its performance with state-of-the-art sketch-to-image adapters in the fashion domain. First, in Sec. 4.2, we present our newly proposed dataset, detailing its construction and structure, along with the evaluation metrics and implementation details. In Sec. 4.3, we outline the experimental setup and baseline models, and analyze the quantitative and qualitative results. Finally, in Sec. 4.4, we conduct ablation studies to highlight key design choices of LOTS.

### 4.1. The Sketchy dataset

We construct a novel dataset, named Sketchy, to support training and evaluation on the task of localized sketch-to-image generation, with localized text-sketch pairs associated with high-quality ground truth images. In the following, we explain in detail how we organize the dataset based on garments and the localized text-sketch creation, as well as the dataset statistics.

**Local garments organization.** We build Sketchy on Fashionpedia [13], a dataset composed of 46k images for training and 1.2k for testing, where fashion experts annotate garments with fine-grained attributes and segmentation masks. While these masks include detailed part annotations (*e.g.*, pockets, zippers, sleeves), they lack a hierarchical structure linking garment components. To improve compositionality, we introduce a two-level hierarchical organization based on segmentation mask overlaps. Specifically, following Fashionpedia taxonomy, the 330k item annotations are first categorized into 14 “whole-body items” (*e.g.*, tops, shirts, and skirts) and 32 “garment parts” (*e.g.*, sleeves, pockets, and necklines). To ensure high-quality compositional annotations, we retain all whole-body categories, along with 21 sub-item categories from Fashionpedia, while removing 11

categories (31k annotations) that are rare, seldom overlap, or lack consistent overlap with any whole-body items, such as umbrellas, bags, and glasses. Then, for each image, we determine the overlap between each garment part’s mask and every whole-body item mask. Whole-body annotations are considered as top-level annotations, *i.e.*, a garment in the image, while part annotations are assigned to the whole-body item with which they have the greatest overlap, *i.e.*, they are considered sub-garment annotations referring to a property, such as sleeves, necklines, and pockets.

**Localized text-sketch creation.** While the Fashionpedia annotations are rich in attributes, they lack a coherent natural language description, which is essential for our text conditioning. Thus, we generate the textual description for each garment in the image by prompting a pre-trained Large Language Model [39] with the hierarchical annotation structure of each garment, along with some in-context learning examples of the desired reply format. We further augment the dataset by including localized garment-level sketches, generated from the ground-truth images, using a pre-trained Image-to-Sketch model [19]. We remove background information via masking to ensure each sketch contains only information about the associated item. We provide a global composition of all the garment sketches, which depicts the sketch of the entire outfit in the original image. Finally, we pre-process the images by resizing them to 512 pixels. We preserve their aspect ratio with white padding into a square format to maintain consistency across samples. Additional details are available in the *Supp. Mat.*.

**Dataset statistics.** Our Sketchy extends Fashionpedia, providing a total of 47k images and 79k garment-level annotations, resulting in an average of 1.7 annotations per image (min 1, max 6). As shown in Fig. 3, each annotation contains the associated sketch, hierarchical attributes, and natural language description of the item. The average word length of the descriptions is 16 words.

### 4.2. Experimental protocol

**Quantitative evaluation and metrics.** Following prior works [2, 3, 5, 23, 37], we adopt Fréchet Inception Distance (FID) [6] to assess the global fidelity of the generated images distribution, relative to ground truth images. Lower values of FID indicate better perceptual quality and stronger correspondence. To measure the semantic alignment, we rely on GlobalCLIP [31] score, calculated as the cosine similarity of image embeddings, in line with [2]. To further capture fine-grained and localized alignment, we adopt LocalCLIP score, building on [14, 23]. Specifically, we first use the masking annotations to obtain a crop for each garment in the ground truth and generated images. Then, we apply the CLIP [31] visual encoder to compute the cosine similarity between each pair of ground truth and generated image crops, averaging their score across all garments. Higher

scores correspond to better semantic alignment. In addition, we utilize the state-of-the-art compositional semantic alignment metric, VQAScore [21] that captures how well-generated images reflect complex text descriptions by relying on Visual Question Answering approaches to query the presence of desired properties. A larger VQAScore suggests improved compositional alignment to the provided prompt. Finally, we follow previous work [5, 43] and evaluate the proposed LOTS based on the Structural Similarity Index Measure (SSIM) assessing the sketch structural alignment and edge fidelity (the higher, the better).

**Human evaluation.** In line with prior works [46, 47], we conduct a user study involving 14 participants with an average of 57 answers each, targeting the evaluation of localized control, rather than visual perception. We generate images ensuring that attributes appear only once in the entire outfit. We leverage a questionnaire to assess whether a specific attribute associated with the  $i$ -th garment is correctly localized in the desired garment of the image and if it leaks to other ones. We quantitatively evaluate considered models in terms of Precision ( $\uparrow$ ), Recall ( $\uparrow$ ), and F1 Score ( $\uparrow$ ) metrics with respect to localized conditioning. A high F1 Score indicates the model’s overall performance in correctly reflecting and localizing the attribute, showing its capacity to balance accurate placement with less attribute confusion. Additional details are available in the *Supp. Mat.*

### 4.3. Main Comparisons

**Baselines.** We compare LOTS with baselines and state-of-the-art sketch-to-image approaches. Specifically, for text-only approaches, we evaluate Stable Diffusion 1.5 (SD) [34] and Stable Diffusion XL (SDXL) [30], which generate images solely from a global text prompt, and GLIGEN [20], which enables localized textual conditioning. We then compare to sketch-to-image adapter methods, including SD-based ControlNet [46], SDXL-based T2I-Adapter [25] and SDXL-based IP-Adapter [45], which all incorporate both a global text prompt and a global sketch as inputs, offering enhanced spatial conditioning. Additionally, we examine compositional modifications of ControlNet (Multi-ControlNet) and T2I-Adapter (Multi-T2I-Adapter) that were modified to allow multiple local sketches conditioning alongside a single global text prompt. Furthermore, we evaluate the most recent localized control method, AnyControl [37], that allows for local sketches and a single global text description as inputs. Finally, we also assess the performance of the adapter-based approaches with fine-tuning on our Sketchy dataset.

**Experimental setup.** We adapt conditioning to different generative models based on their input requirements. For global descriptions, we concatenate all garment descriptions, while for models requiring a single image guide, we create a composite global sketch by merging individual gar-

ment sketches. When localized control is supported, we use garment-specific sketches and/or descriptions as input. For consistency, all images are generated at (512x512) resolution using each model’s default inference setup. Additionally, to ensure fairness, our global description  $T_g$  remains fixed across samples as “A picture of a model posing, high-quality, 4k”. Experiments exploring variations in global descriptions are detailed in the *Supp. Mat.*

**Implementation Details.** We adopt DINOv2 vits14 [29] as sketch encoder. As text encoder, we follow the findings in [30] and use a combination of OpenCLIP ViT-bigG [11] and CLIP ViT-L [31] by concatenating the penultimate text encoder outputs along the channel-axis. Additional implementation details are available in the *Supp. Mat.*

**Quantitative results.** Tab. 1a reports the quantitative evaluation across global quality, semantic alignment, and structural similarity metrics on the test split of Sketchy. Our method demonstrates state-of-the-art performance in GlobalCLIP, LocalCLIP, and VQAScore, while ranking third in SSIM, indicating overall superiority and strong alignment both semantically and structurally. Specifically, LOTS attains the second lowest FID, demonstrating a high perceptual fidelity. A similar pattern is observed in GlobalCLIP and LocalCLIP scores, with our method surpassing all baselines and state-of-the-art models (+3.4% and +1.2%), indicating strong semantic alignment and feature similarity with the ground truth. For compositional semantic alignment, T2I-Adapter and our method achieve the highest VQAScore, outperforming all alternatives in textual prompt following. While some models outperform LOTS in some metrics, their improvements come with trade-offs: the fine-tuned IP-Adapter achieves a lower FID but sacrifices both text and sketch guidance, as evidenced by its significantly lower Compositional Alignment metrics. Similarly, for structural similarity, Multi-T2I-Adapter achieves the highest SSIM, followed by IP-Adapter, while our method ranks third. As opposed to LOTS, however, both IP-Adapter and Multi-T2I-Adapter emphasize sketch-guidance over prompt adherence and image coherence, as evidenced by their subpar LocalCLIP, VQAScore, and FID scores, which figure among the lowest. In conclusion, these results highlight how, thanks to its novel pairing strategy, LOTS strikes an optimal balance between image quality and prompt adherence, surpassing prior approaches and setting the new state-of-the-art performance in the fashion localized sketch-to-image task.

**Results with human evaluation.** In this evaluation, we aim to measure whether an attribute is correctly localized in a garment of the generated image. In Tab. 1b, we present the results of our human study for attribute localization and confusion across different models. Our LOTS achieves the highest F1 score among all models, indicating superior per-

Model	Conditioning Visual/Textual	Global Quality		Compositional Alignment		
		FID ( $\downarrow$ )	GlobalCLIP ( $\uparrow$ )	LocalCLIP ( $\uparrow$ )	VQAScore ( $\uparrow$ )	SSIM ( $\uparrow$ )
SD [34]	-G	1.11	.603	.745	.719	.663
SDXL [30]	-G	1.77	.529	.701	.660	.544
GLIGEN [20]	-L	0.93	.568	.704	.395	.614
ControlNet [46]	G/G	1.08	.622	.789	.733	.674
Multi-ControlNet [46]	L/G	1.10	.615	.780	.730	.672
IP-Adapter [45]	G/G	2.80	.537	.682	.611	.715
T2I-Adapter [25]	G/G	2.16	.534	.705	.635	.482
Multi-T2I-Adapter [25]	L/G	1.14	.583	.766	.697	<b>.723</b>
AnyControl [37]	L/G	0.99	.602	.777	.712	.544
GLIGEN [20]	-L	1.70	.564	.713	.419	.514
ControlNet [46]	G/G	0.80	<u>.645</u>	<u>.801</u>	.717	.574
Multi-ControlNet [46]	L/G	0.84	.638	.799	.720	.572
IP-Adapter [45]	G/G	<b>.69</b>	.621	.787	.714	.631
T2I-Adapter [25]	G/G	1.03	.570	.753	<b>.749</b>	.612
Multi-T2I-Adapter [25]	L/G	1.11	.559	.744	<u>.734</u>	.605
LOTS (Ours)	L/L	<u>.79</u>	<u>.679</u>	<u>.813</u>	<u>.749</u>	.678

(a) Comparisons between LOTS and state-of-the-art sketch-to-image approaches. In the Conditioning column, L and G indicate whether the model accepts Local or Global inputs as Visual or Textual conditioning. We divide the table into three sections: zero-shot approaches, fine-tuned approaches on Sketchy, and our approach LOTS. We highlight the best performance in bold and underline the second best.

Table 1. On the left, the quantitative metrics on the test split of Sketchy. On the right, are metrics obtained through our user study.

formance in accurately localizing attributes on the desired items while minimizing unintended leakage. Furthermore, it performs the best in Precision, demonstrating its effectiveness in preventing attribute confusion. In contrast, models such as SDXL [30] and Multi-T2I-Adapter [25], while strong in Recall, exhibit lower Precision scores, suggesting that attributes may inadvertently leak to unintended items, as shown in Fig. 4 (first row, the striped trousers). On the other hand, T2I-Adapter [25], ControlNet [46], and AnyControl [37] show relatively lower performance in both Precision and Recall. Models that underwent fine-tuning also improved performance, with Multi-T2I-Adapter [25] attaining the second-highest F1 score following LOTS. Finally, we measure the statistical relevance of these results and record a high Krippendorff’s  $\alpha = 0.81$ , testifying to a high inter-annotator agreement.

**Qualitative results.** We qualitatively analyze the attribute localization ability of different models, evaluating whether models accurately associate attributes with the intended garment, ensuring that details such as patterns, silhouettes, and structural elements are correctly rendered in their designated locations. As shown in Fig. 4 (top), the input description specifies three main items associated with their attributes and garments. In particular, a plain single-breasted blazer jacket, a floral top, and striped pants. In this case, a proper generation should be able to faithfully follow the description, placing the floral pattern on the top while keeping the jacket’s plain pattern. As shown, LOTS is the only approach that effectively generates a plain jacket with a floral top and striped pants. In contrast, multiple baselines and state-of-the-art models exhibit errors. For instance, Multi-T2I-Adapter successfully follows the sketch, but fails to capture the floral pattern of the top. On the other hand, both

Model	Attribute Localization		
	Precision ( $\uparrow$ )	Recall ( $\uparrow$ )	F1 ( $\uparrow$ )
SDXL [30]	.636	<b>.754</b>	<u>.690</u>
ControlNet [46]	.596	.449	.512
Multi-ControlNet [46]	.487	.365	.418
IP-Adapter [45]	.625	.139	.227
T2I-Adapter [25]	.409	.170	.240
Multi-T2I-Adapter [25]	.370	.270	.312
AnyControl [37]	.281	.134	.182
ControlNet [46]	<u>.667</u>	.516	.582
Multi-ControlNet [46]	.541	.417	.471
IP-Adapter [45]	.559	.384	.455
T2I-Adapter [25]	.463	.397	.427
Multi-T2I-Adapter [25]	.551	.692	.614
LOTS (Ours)	<b>.813</b>	<u>.650</u>	<u>.722</u>

(b) Results of qualitative user study of attribute localization and confusion conducted between LOTS and other models. We highlight the best results for each metric in bold and underline the second best.

IP-Adapter and T2I-Adapter wrongly localize the “striped” attribute to the jacket, failing to maintain semantic alignment with the prompt. These trends are not limited to isolated cases but are consistently observed across multiple examples, as denoted by the results in Tab. 1a and Tab. 1b. More qualitatives are available in *Supp. Mat.*

#### 4.4. Ablation Analysis

In this section, we present an ablation analysis of the key components used in LOTS, and evaluate the impact of different design choices, such as the image encoder, diffusion guidance strategy, and the number of pooling tokens used in the Pair-Former module.

**Sketch Encoder.** In the first section of Tab. 2, we analyze different encoders for our sketch guidance. Training a dedicated sketch encoder end-to-end simultaneously to LOTS (Trained) is ineffective, resulting in the lowest SSIM recorded (.615). We hypothesize this is due to the pre-trained text features dominating the initial training phase, causing back-propagation to neglect the sketch encoder optimization. Thus, to enhance sketch guidance, we test different frozen pre-trained image encoders, namely CLIP [31], BLIP2 [18], and DINOv2 [29]. DINOv2 (LOTS) results in the highest SSIM (.678), with both CLIP and BLIP2 achieving subpar SSIM performance (.623 and .630 respectively).

**Diffusion Guidance.** The second section of Tab. 2 showcases different diffusion guidance strategies. In the first experiment (No Pooling), we compose our conditioning sequence  $P$  by concatenation of all the  $N$  paired features coming from the projectors without any pooling operation, i.e.,  $P = [h_1^S; h_1^T; \dots; h_N^S; h_N^T]$ . While this approach keeps all the encoder information available for guidance,

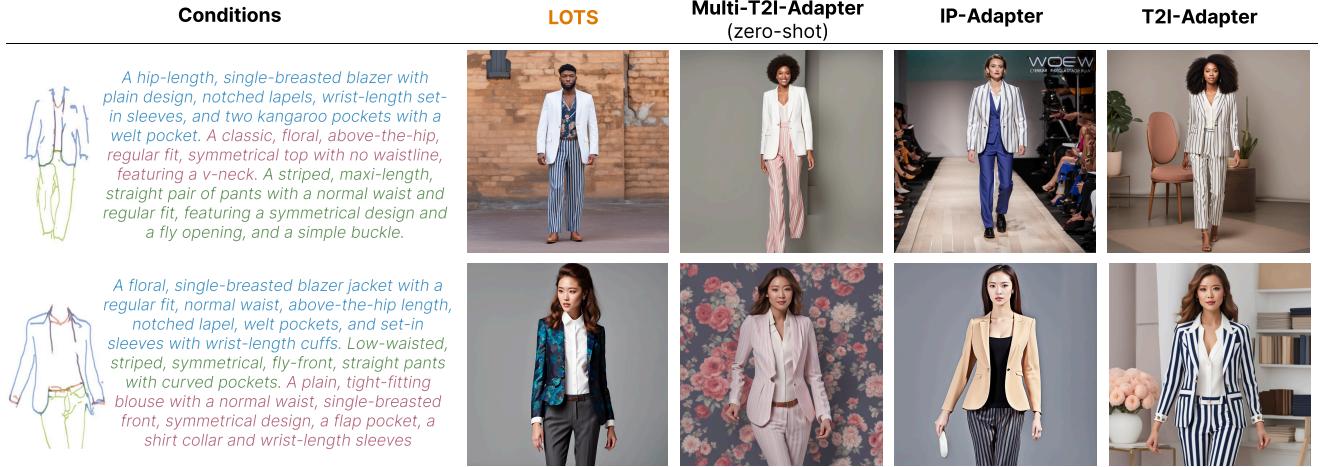


Figure 4. Qualitative results of LOTS in comparison with Multi-T2I-Adapter [25], IP-Adapter [45], and T2I-adapter [25]. Given paired localized text-sketch pairs as conditioning inputs, LOTS can better reflect fine-detailed attributes in the intended local region of the generated images, effectively mitigating attribute confusion.

the model ignores the sketch information, focusing only on textual conditioning, as testified by the high VQAScore (.724) and low SSIM (.623). We hypothesize this is due to the large number of uninformative tokens coming from the image encoder, *i.e.*, the tokens where no sketch is present, pushing the model to ignore their guidance and focus only on the feature-rich textual features. Averaging all the pairs into a single unified representation  $P = \frac{1}{N} \sum_{i=1}^N [h_i^S; h_i^T]$ , improves SSIM, but is unsatisfactory in textual adherence with a low .676 VQAScore. By compressing the pairs’ features and deferring the pair-merging operation to the diffusion process as presented in Sec. 3.1, LOTS achieves the best performance in both compositional semantic alignment (.749 VQAScore) and sketch-guidance (.678 SSIM).

**Pair-Former.** We first evaluate the impact of our Pair-Former by replacing it with standard cross-attention layers (Cross-attn), resulting in lower performance across all metrics. We believe this is due to cross-attention layers limiting interactions to cross-modality only, while Pair-Former’s self-attention enables both intra-/cross- modality interactions between text and sketch. Additionally, to validate the feature compression performed by our Pair-Former, we augment the number of learning pooling tokens and find that, interestingly enough, a higher number of pooling tokens (64 vs. 32) lowers the performance of LOTS by .08 in VQAScore. We hypothesize this is due to both the increased number of trainable parameters and the data redundancy introduced by the larger number of tokens, causing issues similar to the No Pooling ablation experiment.

## 5. Conclusions

In conclusion, we tackled the challenging problem of *image generation with localized sketch-text pairs*, featuring a

Component	Choice	Compositional Alignment		
		LocalCLIP ↑	VQAScore ↑	SSIM ↑
Sketch Encoder	Trained	.762	.689	.615
	CLIP	.802	.685	.623
	BLIP2	.797	.678	.630
Diffusion Guidance	No Pooling	.813	.724	.623
	Mean Pooling	<b>.819</b>	.676	.659
Pair-Former	Cross-attn	.804	.681	.626
	64 Tokens	.804	.669	.628
LOTS		.813	<b>.749</b>	<b>.678</b>

Table 2. Ablation over different components of LOTS

more realistic scenario in the fashion design process. We proposed a novel method *LOTS*, introducing a dynamic conditioning strategy that mitigates attribute confusion between input sketch-text pairs through modular processing of pairs and defers multi-condition fusion to the downstream denoising stage. We then integrate our approach with Stable Diffusion via an adapter-based method to effectively aggregate localized information. Moreover, we introduced *Sketchy*, a new dataset with high-quality fashion images associated with localized sketch-text pairs, to support model training and evaluation. LOTS achieves state-of-the-art performance on Sketchy via both quantifiable metrics and human evaluation, in terms of text-image alignment and visual grounding of local attributes. Future works will explore more appropriate evaluation techniques for localized text-image alignment with fine details. Furthermore, LOTS currently struggles with generating non-visually grounded attributes (*e.g.*, “summer-vibe”), likely due to CLIP’s limited sensitivity to abstract concepts [38]. As such, an interesting future direction would be to include more powerful textual encodings to include such abstract concepts.

## 6. Acknowledgment

This study was supported by LoCa AI, funded by Fondazione CariVerona (Bando Ricerca e Sviluppo 2022/23), PNRR FAIR - Future AI Research (PE00000013) and Italiadomani (PNRR, M4C2, Investimento 3.3), funded by NextGeneration EU. This study was also carried out within the PNRR research activities of the consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23062022, ECS\_00000043). This manuscript reflects only the Authors' views and opinions. Neither the European Union nor the European Commission can be considered responsible for them. We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 as BSC, Spain. Finally, we acknowledge HUMATICS, a SYS-DAT Group company, for their valuable contribution to this research.

## References

- [1] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *ICCV*, 2023. 3
- [2] Dina Bashkirova, Jose Lezama, Kihyuk Sohn, Kate Saenko, and Irfan Essa. Masksketch: Unpaired structure-guided masked image generation. In *CVPR*, 2023. 5
- [3] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Adaptively-realistic image generation from stroke and sketch with diffusion model. In *WACV*, 2023. 5
- [4] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *ICCV*, 2019. 2
- [5] Vudit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair diffusion: A comprehensive multimodal object-level image editor. In *CVPR*, 2024. 3, 5, 6
- [6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5
- [7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint*, 2022. 2
- [8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2
- [9] 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 *NeurIPS*, 2023. 3
- [10] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023. 3
- [11] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 6
- [12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2
- [13] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *ECCV*, 2020. 2, 5
- [14] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023. 3, 5
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014. 1
- [16] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *CVPR*, 2023. 2
- [17] Subhadeep Koley, Ayan Kumar Bhunia, Deependhu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's all about your sketch: Democratizing sketch control in diffusion models. In *CVPR*, 2024. 3
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 4, 7
- [19] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. In *WACV*, 2019. 5, 2
- [20] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Glichen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 3, 6, 7
- [21] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024. 6
- [22] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *ECCV*, 2018. 2
- [23] Denis Lukovnikov and Asja Fischer. Layout-to-image generation with localized descriptions using controlnet with cross-attention control. *arXiv preprint*, 2024. 5
- [24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2, 3
- [25] 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 *AAAI*, 2024. 1, 2, 3, 6, 7, 8, 5
- [26] Pouyan Navard, Amin Karimi Monsefi, Mengxi Zhou, Wei-Lun Chao, Alper Yilmaz, and Rajiv Ramnath. Knobgen:

- Controlling the sophistication of artwork in sketch-based diffusion models. *arXiv preprint*, 2024. 3
- [27] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2
- [28] Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. In *ICML*, 2024. 3
- [29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 6, 7
- [30] 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. In *ICLR*, 2024. 6, 7, 1, 5
- [31] 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 *ICML*, 2021. 5, 6, 7
- [32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. 2
- [33] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 2
- [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 6, 7, 1
- [35] 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. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2
- [36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2
- [37] Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In *ECCV*, 2024. 2, 3, 5, 6, 7
- [38] Davide Talon, Federico Girella, Ziyue Liu, Marco Cristani, and Yiming Wang. Seeing the abstract: Translating the abstract language for vision language models. In *CVPR*, 2025. 8
- [39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaee, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, 2023. 5
- [40] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. 1
- [41] Andrey Vojnov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH*, 2023. 2, 3
- [42] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint*, 2022. 2
- [43] Xi Wang, Hongzhen Li, Heng Fang, Yichen Peng, Haoran Xie, Xi Yang, and Chuntao Li. Lineart: A knowledge-guided training-free high-quality appearance transfer for design drawing with diffusion model. In *CVPR*, 2025. 3, 6
- [44] Zhifeng Xie, Hao Li, Huiming Ding, Mengtian Li, Xinhan Di, and Ying Cao. Hierafashdiff: Hierarchical fashion design with multi-stage diffusion models. In *AAAI*, 2025. 3
- [45] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ipadapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, 2023. 1, 2, 3, 4, 6, 7, 8, 5
- [46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 6, 7, 5
- [47] 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. In *NeurIPS*, 2024. 3, 6

# LOTS of Fashion!

## Multi-Conditioning for Image Generation via Sketch-Text Pairing

### Supplementary Material

#### A. Overview

In this supplementary material, we supply more details regarding the implementation of our proposed method LOTS (Sec. B). Then, we describe in more detail our subjective evaluation protocol with human evaluation (Sec. C). In Sec. D we explore the impact of different global descriptions during inference. Sec. E provides additional details on the creation of the Sketchy dataset. Finally, we report more qualitative results to provide visual evidence of how our proposed method can improve image generation with a better visual grounding of localized textual attributes (Sec. F).

#### B. More details on LOTS implementation

**Implementation Details.** In LOTS, the image and text projectors are linear layers followed by a layer normalization operation, while the Pair-Former is implemented as a self-attention transformer with two self-attention blocks. As diffusion backbone, we employ Stable Diffusion XL [30], using a generic global description as a prompt to contextualize the generation in the fashion domain while ensuring that all garment-level information comes from our adapter model. All the Stable Diffusion family models and weights, as well as cross-attention blocks and training utilities, were gathered from [40]. During training, we freeze all model weights, except for the Image and Text projectors, our Pair-Former, and the additional cross-attention blocks. LOTS and all the fine-tuned approaches are trained on the train split of Sketchy.

We train LOTS following the standard Stable Diffusion procedure [34], while all other fine-tuned adapters are trained using their official implementations and default parameters. For LOTS, we use Adam [15] optimizer, learning rate of  $1e^{-5}$ , and a total batch size of 32, while other approaches are trained using their default hyper-parameters.

#### C. Details on human evaluation

We design an attribute-focused questionnaire that focuses on the objective attribute existence, rather than subjective user preferences. This is done to avoid introducing bias from human perception, which can be influenced by complex factors, e.g., image quality, that are irrelevant to localized attribute grounding. Specifically, we aim to evaluate whether an attribute is correctly localized in the generated image and whether it appears in unintended regions.

In the study, images are generated using models from Sec. 4.3 with the same sketch and textual descriptions,

and are then refined using the Stable Diffusion XL Refiner<sup>1</sup> model to avoid bias from the overall image quality. Notably, we enrich each garment description with a pattern attribute during generation (e.g., striped, dotted) so that each garment is assigned a unique pattern, i.e. the same pattern can't appear on more garments in the same picture. The rationale behind this decision is twofold: i) patterns are one of the few attributes that can be applied to any kind of garment without limitations, as opposed, for instance, to sleeve length or neckline shape; ii) by being uniquely assigned and easily identifiable, user can detect attribute confusion with relative ease.

In practice, we instruct the user to answer a pair of templated questions, ‘‘Consider the garment <class1>: is it <attribute>?’’, and ‘‘Consider the garment <class2>: is it <attribute>?’’, where <class> is the garment class, and the <attribute> is an attribute that is visually noticeable, e.g., ‘‘check’’ or ‘‘striped’’. Since only one item per image can contain the sampled attribute, if it fails to represent an item or attribute or binds it to an unrelated garment, the responses will reflect this misalignment.

To measure this quantitatively, we use *Recall*, *Precision*, and *F1 scores* as metrics. Recall is defined as the fraction of times that a specified attribute is correctly applied to the intended clothing garment. In other words, Recall measures how often the model successfully localizes the desired attribute on the correct garment. Formally, Recall is computed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

where:

- TP (True Positives): The attribute is reflected correctly at the intended garment in the generated image.
- FN (False Negatives): The attribute should have been reflected, but is missing.

Similarly, Precision is defined as the extent to which the generated attribute appears exclusively on the intended item without being mistakenly applied to other objects:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

where a False positive (FP) is defined as when an attribute appears on unintended garments indicating attribute confusion.

<sup>1</sup><https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0>



Figure 5. Effects of different global descriptions on the generation of LOTS. By changing the global description, we are able to customize general aspects such as the background and style of the model and the outfit (see text).

A high recall value indicates that attributes are generated where they are supposed to, whereas a high precision score indicates that attributes are not leaking to irrelevant objects. F1 Score offers a balanced view of both precision and recall, taking into account both false positives and false negatives:

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

## D. Impact of global description on generation

To illustrate the impact of our global conditioning  $T_g$ , we present in Fig. 5 three examples where the local sketch and text remain fixed, while the global text varies across three different descriptions. Notably, the description “A goth model” results in an image with extra stylistic details, such as additional bracelets, earrings, pale skin, and red-tinted hair, while maintaining the overall properties of the garments.

## E. Data Recipe for Sketchy

In this section, we add specific details regarding the construction of Sketchy. Please note that our pre-computed data and data curation scripts are available on our project website. First, we organize the garments in Fashionpedia’s annotations [13] to create our localized hierarchical structure (Sec. E.1). Then, we use an LLM to automatically generate a natural language description of the whole-body items starting from their attributes (Sec. E.2). Finally, we generate a sketch for every item that appears in each image (Sec. E.3).

### E.1. Local garments organization

As stated in the main document, Fashionpedia contains annotations for both whole-body items, *e.g.*, shirts, and

garment-parts, *e.g.*, sleeves. Despite this, Fashionpedia does not explicitly link garment-parts with their respective whole-body item, *e.g.*, the shirt and sleeve annotations are not linked to each other. This can become problematic when multiple suitable whole-body items appear in a single image for a given garment-part: if we have a “sleeve” annotation, and both a “shirt” and “jacket” in the same image, it is not clear which item the sleeve belongs to. At the same time, garment-part annotations contain fine-grained attributes that we would like to associate with the whole-body item, *e.g.*, the “long” attribute for the sleeves does not appear in the whole-body shirt annotation.

To amend this, we use the segmentation masks provided by Fashionpedia to find associations between garment-parts and whole-body items. Specifically, for images containing garment parts, we calculate the overlapping area between each garment-part mask and the masks of whole-body items. The garment-part’s attributes are then assigned to the whole-body category with the largest overlapping area. In extreme cases where no overlap exists, we assign the garment attribute to the whole-body category that is most frequently associated with this specific garment part across the dataset, *i.e.*, based on co-occurrence statistics.

## E.2. Partial Descriptions

Fashionpedia annotations contain a list of attributes describing the properties of the item depicted (for both whole-body and garment parts). While these lists of attributes could be used directly to condition a text-to-image model, we found the interaction to be unrealistic: in a real-world case, we believe the user would rather use natural language descriptions to build a coherent sentence describing the contents of the image. For this reason, for each whole-body annotation we use an off-the-shelf Llama 3.1 8B-Instruct<sup>2</sup> model to generate a natural language description starting from the attribute list. The used prompt is reported in Fig. 6. To further guide the model in the generation, we provide some in-context-learning examples, describing both the input structure and corresponding expected output, as depicted in Fig. 7.

## E.3. Partial Sketches

To generate the localized sketches we rely on our hierarchical annotation structure and an off-the-shelf Image-to-Sketch model, Photo-sketching [19]. Photosketching takes an image as input and generates the corresponding human-like sketch. However, our goal is to generate localized sketches, *i.e.*, sketches of single items inside the image, and not the entire scene. To do so, we first obtain the segmentation mask (from the Fashionpedia annotations) of every whole-body item inside the input image. Then, for every item, we crop the image around the entire mask and

<sup>2</sup><https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

```

# SYSTEM PROMPT
"You are a fashion expert. Describe the clothing item concisely based on
the information provided, strictly within 70 tokens.
You need to prioritize keeping the information that may influence
the appearance the most, while trying to describe the image
as informatively as possible.
Within each whole-body item, there could be sub-items, each may have
its own descriptive attributes.

Here is the structure of the clothing item information provided:
The item information is a structured dictionary including the following keys:
(1) "category": A string indicating the main item category
(e.g., "Coat", "Trousers").
(2) "top_level": A list [] of attributes that provide a general description
of the main item (e.g., ["long", "wool"]).
(3) "sub_level": A list [] of dictionaries {} where each dictionary describes
a specific part of the main item. Each dictionary contains:
- The part's name as the key (e.g., "Collar", "Pockets").
- A list [] of attributes as the value that
describes the details of that part (e.g., ["wide", "deep", "large"]).

Please provide a cohesive description of the item, incorporating all
the details provided for both the main item and its sub-items.
Ensure the description maintains clarity and preserves the hierarchical
relationship between main items and sub-items.
Refrain from giving any personal opinion.
You must reply in the format of a Python dictionary {desc: description}."
```

Figure 6. System prompt for generating whole-body garment descriptions from attributes.

resize the image, effectively “zooming in” on the item of interest. The Photo-sketching model is then used to generate the sketch of the cropped image. Finally, we remove unwanted background information by masking the generated sketch with the item segmentation mask.

Thanks to our zooming-in operation, we found that Photo-sketching was able to focus on more fine-grained details, such as pockets and patterns, that would have otherwise been lost by generating a sketch of the entire image. At the same time, our final masking operation removes unwanted information coming from neighboring items and background objects. We found that this allows the generative model downstream (LOTS) to focus on the item shape during training, avoiding information corruption coming from unrelated items being present in the sketch.

## F. More qualitative results

Figure 8 shows more qualitative results of LOTS in comparison with baselines and competitors as described in Sec. 4. Given paired localized text-sketch as conditioning inputs,

LOTS can more correctly reflect fine-grained attributes in the intended local region in the generated images, effectively mitigating attribute confusion, a common problem when only a global description is provided as the textual condition. We observe two main failure cases of state-of-the-art approaches. On the one hand, models tend to confuse attributes, conveying them on main whole-body garments (see striped and checked outfits in SDXL, T2I-Adapter, and Multi-T2I-Adapter). This is in line with the intuition that the model focuses on the most peculiar attribute in the single global description. On the other hand, generation can completely fail to follow the conditioning sketch, e.g., three out of seven generations for Multi-T2I-Adapter, indicating that explicit merge of the spatial conditioning, before generation, could possibly corrupt the guiding information. In contrast, with the modularized processing of pairs and the diffusion pair guidance, which overcomes the pooling of conditioning localized information, LOTS allows for effective and fine-grained conditioning at both semantic (textual descriptions) and spatial level (sketches). More results in Fig. 9.

```

# IN-CONTEXT SAMPLE 1
# input structure
"{
    "category": "coat",
    "top_level": ["long", "wool"],
    "sub_level": [
        {"Collar": ["wide"]},
        {"Pockets": ["deep"]},
        {"Buttons": ["large"]}
    ]
}"
# output
"{desc: A long wool coat with a wide collar, deep pockets and large buttons}"

# IN-CONTEXT SAMPLE 2
# input structure
"{
    "category": "trousers",
    "top_level": ["slim-fit"],
    "sub_level": [
        {"Stitching": ["subtle"]},
        {"Leg": ["tapered"]}
    ]
}"
# output
"{desc: Slim-fit trousers with subtle stitching and a tapered leg}"

# IN-CONTEXT SAMPLE 3
# input structure
"{
    "category": "shirt",
    "top_level": ["cotton"],
    "sub_level": []
}"
# output
"{desc: A cotton shirt}"

# IN-CONTEXT SAMPLE 4
# input structure
"{
    "category": "shoe",
    "top_level": [],
    "sub_level": []
}"
# output
"{desc: A pair of shoes}"

```

Figure 7. In-context-learning samples appended to the system prompt for every input annotation for generating whole-body garment descriptions from attributes.

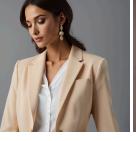
Conditions	LOTS	AnyControl	SDXL	T2I-Adapter	Multi-T2I-Adapter	ControlNet	IP-Adapter
 Check, mini, symmetrical, gathering shorts with a regular fit. An above-the-hip, floral, regular-fit, classic t-shirt with an oval neckline and wrist-length poet sleeves.							
 A plain, single-breasted, symmetrical blazer jacket with set-in sleeves, notched lapels, and two welt pockets. A dotted, regular-fit top with no waistline.							
 A check, maxi-length, straight pair of pants with a normal waist and regular fit, featuring a symmetrical design and a fly opening, and a simple buckle. A classic, geometric, above-the-hip, regular-fit, symmetrical top with no waistline, featuring a v-neck. A hip-length, single-breasted blazer with striped design, notched lapels, and two kangaroo pockets with a welt pocket.							
 A plain, symmetrical, above-the-hip, regular-fit, high-waist fit and a sweetheart neckline. High-waisted, tight-fitting, abstract leggings pants.							
 A loose-fitting, classic-length, check t-shirt with a round neckline and no waistline, and a loose, tie-front, maxi-length dotted pants with a straight cut and symmetrical design. A regular, plain, symmetrical, zip-up button jacket with no waistline, above-the-hip length, wrist-length set-in sleeves and a stand-away collar.							
 A hip-length, single-breasted blazer with plain design, notched lapels, wrist-length set-in sleeves, and two kangaroo pockets with a welt pocket. A classic, floral, above-the-hip, regular-fit, symmetrical top with no waistline, featuring a v-neck. A striped, maxi-length, straight pair of pants with a normal waist and regular fit, featuring a symmetrical design and a fly opening, and a simple buckle.							
 A maxi-length, plain, regular-fit pair of pants with no waistline and symmetrical design, featuring curved pockets. An abstract, above-the-hip, regular-fit, classic t-shirt with a v-neck and a small pattern. A plain, short-sleeved, floral shirt with set-in short sleeves, a shirt collar and a flap pocket.							
 A pair of striped, loose-fitting, pleated, mini, symmetrical, bermuda-style shorts with a fly opening and two slash pockets. A plain, above-the-hip, regular-fit, classic t-shirt with a crew neckline and wrist-length set-in sleeves.							
 A loose-fitting, classic-length, check t-shirt with a round neckline and no waistline. Loose, tie-front, maxi-length pants with a straight cut and symmetrical design. A regular, geometric, symmetrical, zip-up button jacket with no waistline, above-the-hip length, wrist-length set-in sleeves and a stand-away collar.							

Figure 8. More qualitative results of LOTS in comparison with baselines and competitors as described in Sec. 4. In this table, SDXL [30], AnyControl [37] are zero-shot approaches, T2I-Adapter [25], Multi-T2I-Adapter [25], ControlNet [46], IP-Adapter [45] are fine-tuned versions. Given paired localized text-sketch are conditioning inputs, LOTS can more correctly reflect fine-detailed attributes in the intended local region in the generated images, effectively mitigating attribute confusion, a common problem when only a global description is provided as the textual condition.

## Conditions

## LOTS

## Conditions

## LOTS



*Double-breasted, floral jacket with peak lapels, and wrist-length set-in sleeves.*  
*Maxi length, symmetrical and straight sailor pants with a check pattern.*



*A tight-fitting, floral shirt with a hip length, short sleeves and a traditional shirt collar.*  
*Striped bermuda shorts.*



*A striped, [...] jacket with wrist-length set-in sleeves, flap pockets and a regular collar. A plain [...] t-shirt with no waistline and a turtle-neck. A dotted, regular fit, [...] pair of pants [...].*



*A loose-fitting, check, single-breasted shirt with a banded collar, and short dropped-shoulder sleeves.*  
*A striped, regular fit, maxi length, straight pair of pants.*



*A loose-fitting, floral, single-breasted, hip-length shirt with short sleeves and a shirt collar.*  
*Check, regular-fit pants.*



*A regular-fit shirt with a striped pattern [...], featuring set-in sleeves and a regular collar.*  
*A plain, classic printed t-shirt with a round neckline.*  
*Loose, maxi-length, straight, check pants.*



*A dotted, regular-fit, hip-length shirt, featuring short set-in sleeves and a traditional shirt collar.*  
*Check, loose-fitting, above-the-knee bermuda shorts.*



*A dotted, oversized, single-breasted shirt with no waistline, featuring wrist-length, dropped-shoulder sleeves, and a shirt collar.*  
*Floral bermuda shorts with a normal waist and regular fit.*



*A loose-fitting, dotted, coat [...].*  
*A floral, regular fit, classic crew-neck t-shirt with a hip length [...]. Plain cargo pants [...].*



*A plain, above-the-hip, tight, symmetrical vest with a boat-shaped neckline.*  
*Check, maxi-length, tight-fitting leggings.*



Figure 9. More qualitative results of LOTS. We underline the textual pattern attributes to facilitate visual inspection.