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`[1] 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:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

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:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

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

[1] https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

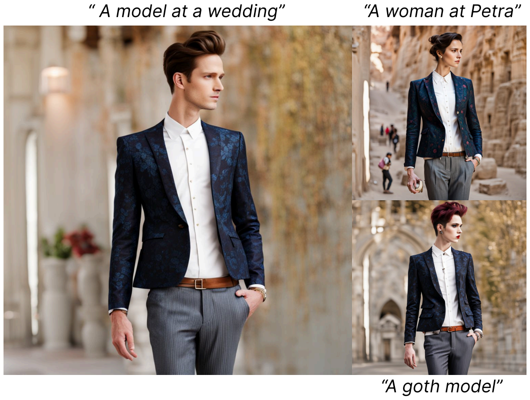


Figure 5. Effects of different global descriptions on the generation of LOTS. By changing the global description, we are able to customize general aspects such as the background and style of the model and the outfit (see text).

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:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

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`[2] 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]. Photo-sketching 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

[2] 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.

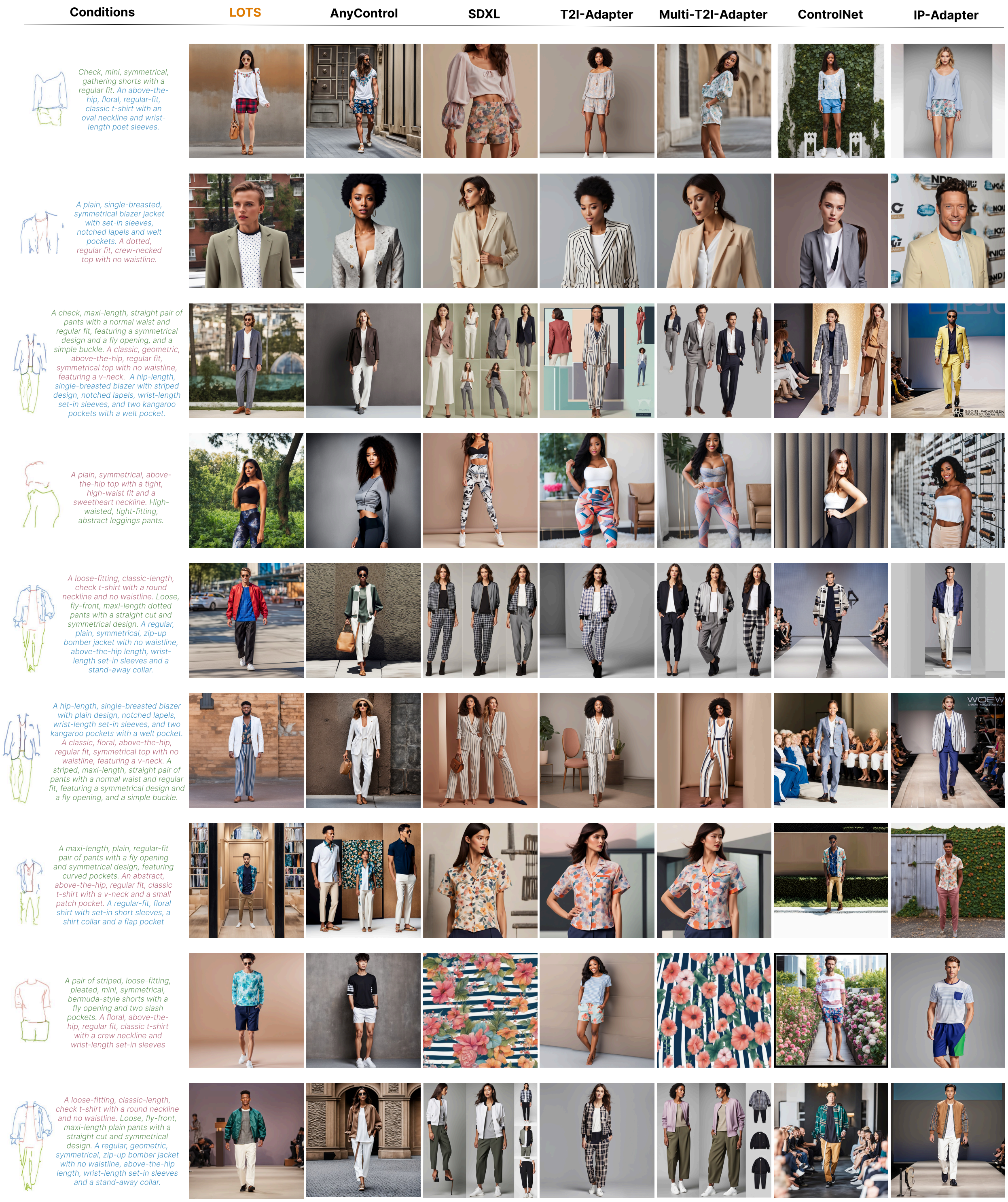
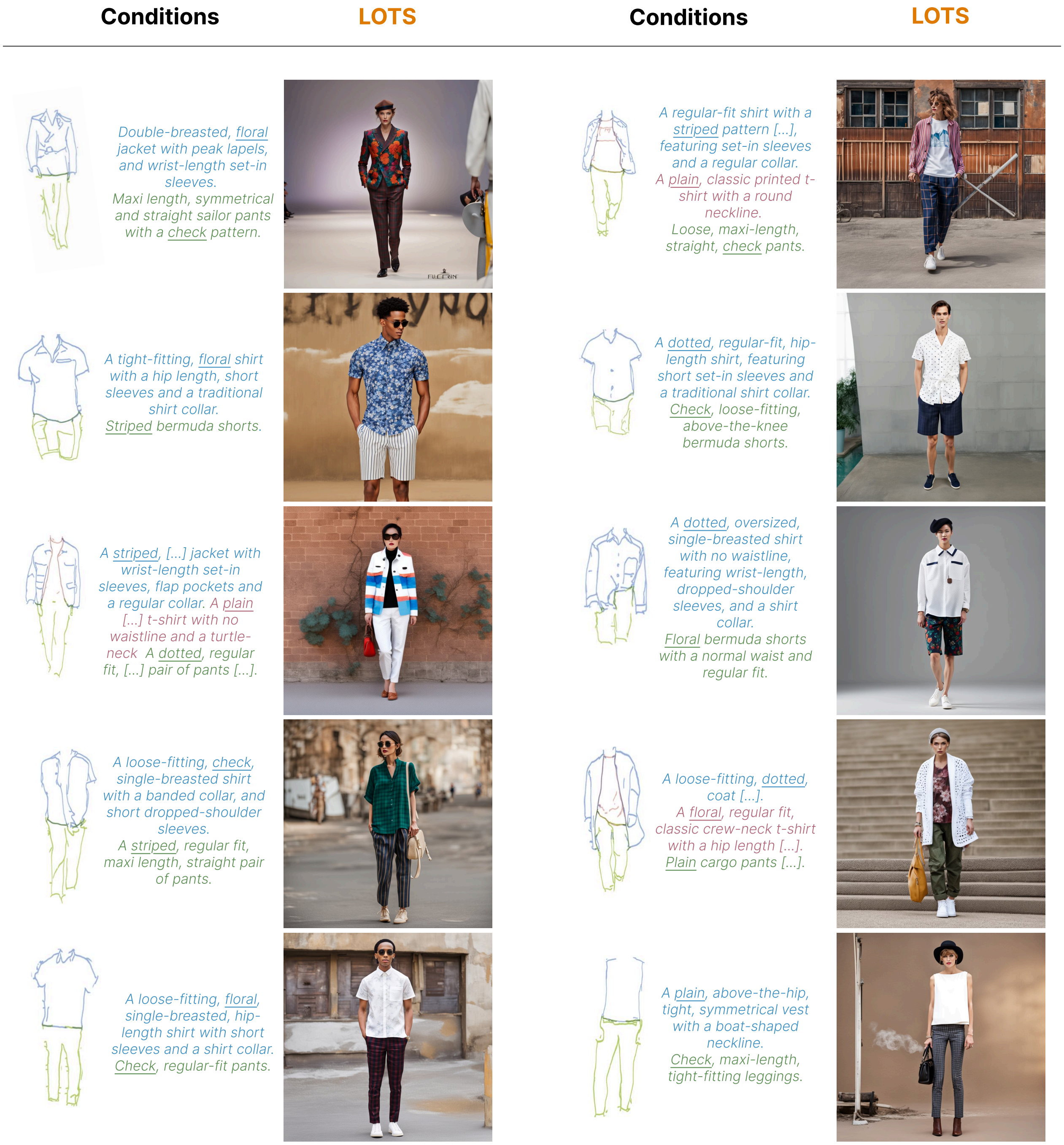

Conditions
LOTS
AnyControl
SDXL
T2I-Adapter
Multi-T2I-Adapter
ControlNet
IP-Adapter
Check, mini, symmetrical, gathering shorts with a regular fit. An above-the-hip, floral, regular-fit, classic t-shirt with an oval neckline and wrist-length poet sleeves.
A plain, single-breasted, symmetrical blazer jacket with set-in sleeves, notched lapels and welt pockets. A dotted, regular fit, crew-necked 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, wrist-length set-in sleeves, and two kangaroo pockets with a welt pocket.
A plain, symmetrical, above-the-hip top with a tight, 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. Loose, fly-front, maxi-length dotted pants with a straight cut and symmetrical design. A regular, plain, symmetrical, zip-up bomber 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 a fly opening and symmetrical design, featuring curved pockets. An abstract, above-the-hip, regular fit, classic t-shirt with a v-neck and a small patch pocket. A regular-fit, 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 floral, 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, fly-front, maxi-length plain pants with a straight cut and symmetrical design. A regular, geometric, symmetrical, zip-up bomber 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.



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.