This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Text-Driven Fashion Image Editing with Compositional Concept Learning and Counterfactual Abduction

Shanshan Huang[1], Haoxuan Li[2], Chunyuan Zheng[3], Mingyuan Ge[1], Wei Gao[1], Lei Wang[1], Li Liu[1]*
[1]School of Big Data & Software Engineering, Chongqing University
[2]Center for Data Science, Peking University [3]School of Mathematical Sciences, Peking University
{shanshanhuang, dcsliuli}@cqu.edu.cn, {hxli, cyzheng}@stu.pku.edu.cn,
{leiwangtt, mingyuange}@stu.cqu.edu.cn, wgao2002@gmail.com

Abstract

Fashion image editing is a valuable tool for designers to convey their creative ideas by visualizing design concepts. With the recent advances in text editing methods, significant progress has been made in fashion image editing. However, they face two key challenges: spurious correlations in training data often induce changes in other areas when editing an area representing the intended editing concept, and these models typically lack the ability to edit multiple concepts simultaneously. To address the above challenges, we propose a novel Text-driven Fashion Image ediTing framework called T-FIT to mitigate the impact of spurious correlation by integrating counterfactual reasoning with compositional concept learning to precisely ensure compositional multi-concept fashion image editing relying solely on text descriptions. Specifically, T-FIT includes three key components. (i) Counterfactual abduction module, which learns an exogenous variable of the source image by a denoising U-Net model. (ii) Concept learning module, which identifies concepts in fashion image editing—such as clothing types and colors and projects a target concept into the space spanned from a series of textual prompts. (iii) Concept composition module, which enables simultaneous adjustments of multiple concepts by aggregating each concept's direction vector obtained from the concept learning module. Extensive experiments show that our method can achieve state-of-the-art performance on various fashion image editing tasks, including single-concept editing (e.g., sleeve length, clothing type) and multi-concept editing (e.g., color & sleeve length).

1. Introduction

Fashion image editing has emerged as a pivotal field within computer vision, driven by the increasing need for personalized and dynamic visual modifications in the fashion industry [1, 33]. The diversity and high-quality image generation capabilities of large-scale text-image diffusion models such as Stable Diffusion [23], DALL-E 2 [24], and Imagen [22] has inspired many subsequent efforts to leverage pre-trained large-scale models for text-driven fashion image editing (TFIE) [12, 19, 32]. A foundational aspect of this field is single-concept editing, where a specific concept [35, 37] like color, texture, or style of an item is altered while keeping the rest of the fashion image the same.

However, the existing methods still suffer from several limitations. First, these methods [2, 6, 41] often encounter challenges related to spurious correlation and unwanted associations learned by the pre-trained model (e.g., Stable Diffusion [23]) due to biases in training data. For example, several methods [28, 29] may erroneously associate a specific clothing type (e.g., *sleeve length*) with a specific cloth style (e.g., *garment length*), as shown in Fig. 1, resulting in unintended modifications outside the target concept. This is because the long sleeves always co-occur with long garments, inducing spurious correlations in the training data. Therefore, eliminating spurious correlations is critical to ensuring precise, isolated modifications. There are many methods to address spurious correlation [5, 11, 13, 31, 36, 38, 39]. For example, Li et al. [16] propose a novel machine unlearning method to reduce the instance weight of biased samples in e-commerce and Sanchez et al. [25] discuss important challenges present in healthcare applications such as processing high-dimensional and unstructured data, as well as temporal relationships. However, due to differences in data structure and final goal, previous methods could not be adapted to image editing scenarios.

Second, existing methods [32, 33] are mainly limited to editing the texture of fashion images, and it is difficult to achieve non-rigid editing, such as precisely adjusting the length of sleeves or skirts. As shown in Fig. 1, UltraEdit [41] and Turbo-edit [6] cannot achieve non-rigid editing, e.g., turning a long skirt into a mini skirt; altering short sleeves into long sleeves, which may attribute to the poor understanding of the concept from the prompt.

Third, existing methods [4, 14, 15] typically focus on editing a single concept in the source image. While single-

*Corresponding author

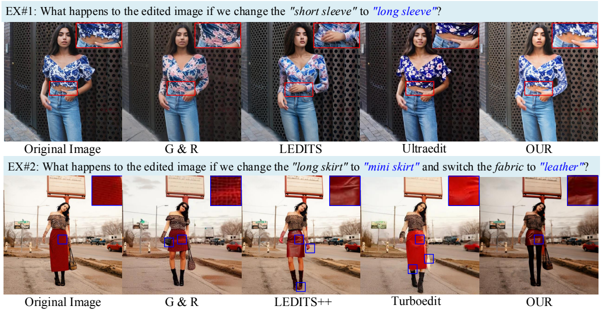
EX#1: What happens to the edited image if we change the "short sleeve" to "long sleeve"?
Original Image
G & R
LEDITS
Ultraedit
OUR
EX#2: What happens to the edited image if we change the "long skirt" to "mini skirt" and switch the fabric to "leather"?
Original Image
G & R
LEDITS++
Turboedit
OUR

Figure 1. Two examples to illustrate the spurious correlation and conceptual omission issues in fashion image editing. Unlike other methods, our method alleviates the issue of over-editing that exists in fashion image editing due to spurious correlation (e.g., sleeve length and garment length, mini skirt and plaid style), as well as the confusion or lack of concepts when editing multiple concepts.

concept methods enable the modification of multiple concepts sequentially, i.e., changing the concepts one by one, they present certain drawbacks. Specifically, these methods either introduce concept omission due to the difficulty of identifying each concept from the prompt; or they lead to unexpected changes due to interactions between different concepts, which is still caused by spurious correlation [10]. For example, modifications made to later concepts may affect previous modifications. From Fig. 1, we can observe that G & R [28] incorrectly adds a "leather bag" when editing the skirt fabric to *leather* and LEDITS++ [2] causes other unrelated attributes changes when changing the *length* and *fabric* of the skirt, such as *the color of the gloves*.

To bridge this gap, we develop a Text-driven Fashion Image ediTing framework (T-FIT) that relies solely on textual prompts, enabling users to perform both single- and multi-concept editing on a single fashion image. Specifically, T-FIT adopts Stable Diffusion [23] and introduces a counterfactual abduction module to address the spurious correlation and concept learning module with concept composition to achieve precise single- and multi-concept editing as shown in Fig. 2. Our contributions are as follows:

- We reveal the issue of spurious correlation and concept omission in fashion image editing and propose T-FIT, a text-driven framework to address these issues.
- We propose counterfactual abduction (CA) to capture visual content and a concept learning module (CLM) to identify and disentangle concepts, mitigating issues of spurious correlation and non-rigid editing.
- We propose a concept composition strategy that enables simultaneous, fine-grained editing across multiple concepts by adjusting the weight of each learned concept to address the concept omission problem.
- Extensive empirical evaluation, in terms of both automated and human assessment metrics, qualitatively and quantitatively demonstrates that our T-FIT approach significantly enhances the quality of fashion image editing.

2. Methodology

2.1. Problem Formulation

We first formulate the image editing problem as follows: given a source fashion image I_S, a corresponding text prompt P that describes the contents of I_S, a set of concept prompts $P_{C^i} = \{p_0^i, p_1^i, \ldots, p_{N-1}^i\}$ used to construct a concept C^i (e.g., *clothing type, color, fabric*) that the user wishes to edit, and a target text prompt P' describing the final editing goal, our objective is to generate a new fashion image I_T by editing I_S in alignment with P' while preserving other areas that are unrelated to the concept C^i.

2.2. Preliminary

In the context of fashion image editing, *concepts* $C = \{C^1, \ldots, C^i, \ldots, C^m\}$ are a distinct and human-interpretable feature or attribute within an image, encompassing elements like *sleeve length, fabric type, color*, and *cloth type*. Each concept C^i contains mutually selected attributes, where each attribute embodies certain shared properties. For instance, the attributes "long sleeves" and "short sleeves" collectively represent the concept of "*sleeve length*", while "woman" and "man" represent the concept of "*gender*". The concept representations $s(C)$ can be derived from the intermediate representation of some text-to-image generative models (e.g., Stable Diffusion) pre-trained over annotated datasets that link visual attributes to textual descriptions [27, 35]. Here, $s : \mathbb{C} \to \mathbb{R}^d$ is the concept representation function, where $\mathbb{C}$ is the concept space containing all possible concepts, and $\mathbb{R}^d$ represents the representation space with dimension d. Following [27, 35], we focus on the compositional concepts, as defined in Definition 1.

Definition 1 (Compositional Concepts)**.** *For concepts* $C^i, C^j \in \mathbb{C}$*, the concept representation* $s : \mathbb{C} \to \mathbb{R}^d$ *is considered compositional if there exist positive weights* $w_i, w_j \in \mathbb{R}^+$ *such that:*

$$s(C^i \cup C^j) = w_i s(C^i) + w_j s(C^j).$$

This definition indicates that the representation of the composed concepts corresponds to the weighted sum of the individual concept representations in the representation space. Following this definition, this study aims to develop a framework for learning and composing conceptual representations from pre-trained text-to-image models (i.e., Stable Diffusion) to achieve flexible and targeted fashion image editing based on these compositional concepts.

2.3. Counterfactual Abduction for Image Editing

To address spurious correlation issues inherent in existing image editing methods, we introduce a counterfactual abduction approach that employs an abduction loss to infer unknown exogenous variables U for each image. These variables capture the additional visual content of source fashion

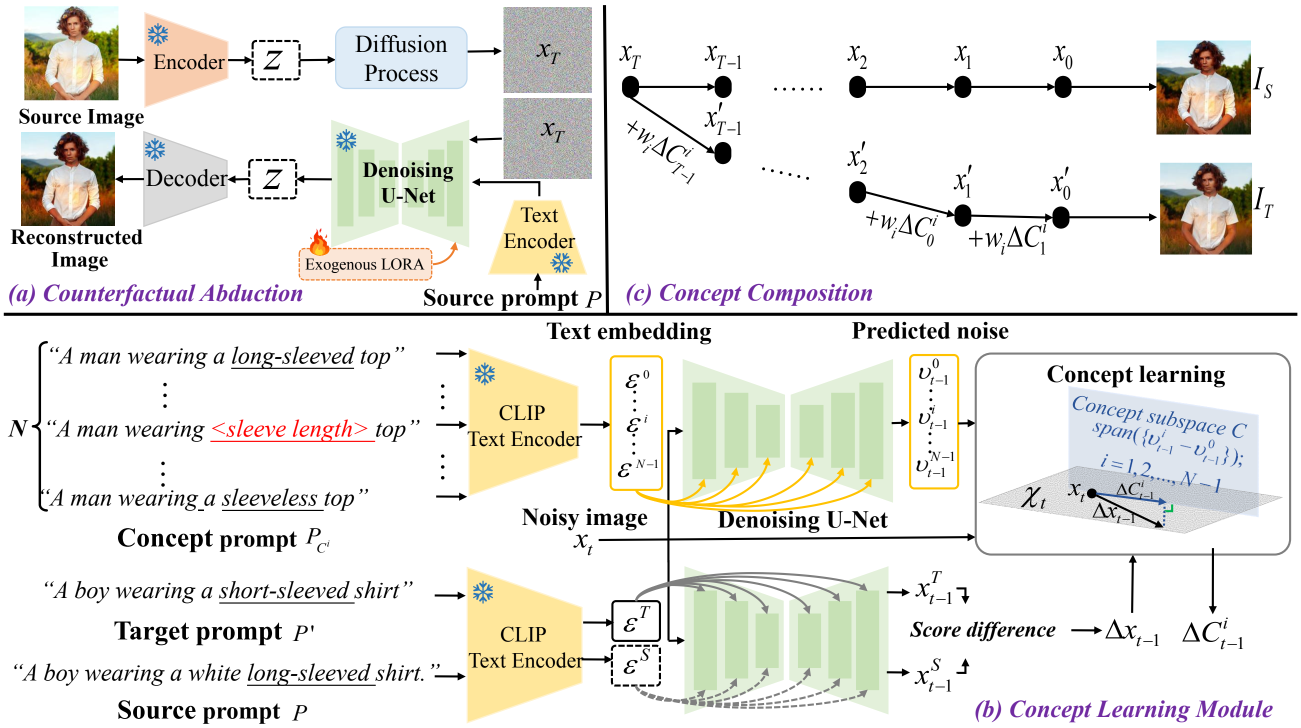

Figure 2. Overview of our T-FIT for fashion image editing. We divide the entire editing process into two stages. In the first stage, we perform counterfactual abduction to infer exogenous variables U that capture the core visual contents of the source fashion image. In the second stage, we employ a concept learning module to discover the target concept from representation space and design a concept composition strategy for combining multiple aspects of the fashion image, such as fabric, color, or pattern, while maintaining consistency with the original image.

images, aiming to reduce uncertainty in fashion image editing and prevent unintended changes in the structure or identity of the source image. In other words, by incorporating U, our method ensures that the edited fashion image I_T integrates the influence of P' while retaining other unrelated visual content from the source image I_S unchanged, thus addressing spurious correlation introduced in pre-trained models from the causal perspective. Specifically, we train the abduction U by optimizing the following Gaussian noise regression, a process analogous to training the reversed diffusion steps:

$$\arg\min_{U} \mathbb{E}_{(t,\epsilon)}||\epsilon - \epsilon_U(x_t, t, \varepsilon_{src})||_2^2, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is the Gaussian noise, $t \in [0, T]$ represents a sampled time step, ϵ_U represents a new denoising model trained on a pre-trained denoising U-Net, wherein U is treated as a trainable parameter. All other model parameters, aside from those of the denoising U-Net, remain fixed during training. The text embedding $\varepsilon_{src} = \psi(P)$ is obtained by a pre-trained CLIP text encoder $\psi(\cdot)$, and the noisy input at time t can be obtained by $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$. Here, $x_0 = I_S$ denotes the original image and α_t is a predefined variance schedule [20, 34].

Following [26], we parameterize U as the UNet LoRA in ϵ_U, LoRA structure is incorporated across all attention layers, convolution layers, and feed-forward layers, which can be expressed by $W' = W + \beta U_A \cdot U_B$. Here, $W \in \mathbb{R}^{d\times d}$ denotes the original weight matrix, $U_A \in \mathbb{R}^{d\times r}$ and $U_B \in \mathbb{R}^{r\times d}$ are low-rank matrices. Wang et al. [30] observed that the larger the time step in the diffusion model, the better the editable and the lower the fidelity. For this reason, the annealing parameter $\beta = \frac{1-\kappa}{T^2}(t-T)^2 + \kappa$ is proposed to improve the editability. In our experiment, the rank r is set to 512, $\kappa \in [0, 1]$ is a constant.

2.4. Compositional Concept Learning for Editing

Unfortunately, the abduction of U is inherently ill-posed [26], often leading to overfitting to the specific P and x. This overfitting reduces the editability of the generative model $G(\cdot, U)$ and may cause challenges like concept omission or confusion, where distinctions between concepts may blur, or certain concepts might overshadow others. Therefore, $G(\cdot, U)$ struggles to generate images that adhere to a new prompt P'. To address this, we introduce a concept learning module that analyzes the representation space to identify composable concepts. Additionally, we propose a multi-concept composition strategy for flexible fashion image editing, enabling single and multi-concept continuous edits. For example, our method can simultaneously modify both the "*sleeve length*" (e.g., changing from long sleeves to short sleeves) and the "*color*" of the clothing (e.g., transitioning from red to blue). The intensity of these edits can also be adjusted, such as making the blue lighter or darker.

Concept Learning Module. Given a target concept C^i, we aim to learn an interpretable direction vector ΔC_t^i at time t. The learned direction vector allows us to control the generation process by varying the latent x_t in the desired directions ΔC_t^i with weight w_i, which is considered as $s_{new}^i = s_{src} + w_i \Delta C_t^i$. Here, $s_{src} = \epsilon_U(x_t, t, \varepsilon_{src})$, $\varepsilon_{src} = \psi(P)$ is derived from pre-trained CLIP text encoder $\psi(\cdot)$, and $\epsilon_U(x_t, t, \varepsilon_{src})$ is trained by counterfactual abduction and is fixed here. Specifically, we first define a set of text prompts $P_{C^i} = \{p_0^i, p_1^i, ..., p_{N_i-1}^i\}$ designed to elicit distinct distributions for a target concept C^i while maintaining a consistent distribution for another concept C^j. Without loss of the generality, we define p_0^i as the source prompt without redundant concepts. In essence, each prompt captures a unique representation of C^i while keeping the representation of C^j consistent. For instance, when editing the concept of "*fabric*," we can formulate a series of prompts like "a lace [fabric 0] dress," "a silk [fabric 1] dress," and "a [fabric $N_i - 1$] dress." These prompts are intended to generate variations in the *fabric* concept while keeping other aspects, like *dress type*.

Then, to identify the direction vector for concept C^i, we first define its representation space as $\mathcal{R}_{C_t^i} := span(\{\upsilon_t^j - \upsilon_t^0 : j = 1, .., N_i - 1\})$. By leveraging the span of conceptual space, we enable the model to edit not just predefined attributes, but also a broader range of related attributes. For example, by defining a structured representation space with "*blue*," "*green*," and "*yellow*," our model can generalize to other colors in the spectrum, such as "*red*". Here, $\upsilon_t^j = \epsilon_U(x_t, t, \varepsilon_j)$, each $\varepsilon_j = \psi(p_j^i)$ is derived from the pre-trained CLIP text encoder $\psi(\cdot)$. Next, to compute the projection matrix M_i for the C^i-space within this representation space, we start by constructing matrix S^i as

$$S_t^i := [\upsilon_t^1 - \upsilon_t^0, ..., \upsilon_t^{N_i-1} - \upsilon_t^0]. \tag{2}$$

In addition, to align changes in the representation space with the semantic structure of the concept space, we first apply singular value decomposition (SVD) on S_t^i, i.e., $SVD(S_t^i) = V_t^i \Sigma_t^i {U_t^i}^T$, to identify the principal orthogonal semantic directions. We then select the top K components of V_t^i to capture the most significant semantic variations (for simplicity, we denote this matrix as V_t^i), and construct the projection matrix $M_t^i = V_t^i {V_t^i}^T$, which projects the concept vector into $\mathcal{R}_{C_t^i}$ and ensures that the changes in the representation space are corrected and aligned with the key semantic directions.

Finally, to derive the editing direction, we calculate the score difference between source representation s_{src} and target representation s_{tar} as $\Delta s_t = s_{tar} - s_{src}$. We then project Δs_t onto the spanned concept space using M_t^i, yielding $\Delta C_t^i = \Delta s_t M_t^i$. This projection ensures the editing operation reflects the relevant semantic changes in the concept space. The modified score can be achieved by

$$s = s_\emptyset + \lambda(s_{new}^i - s_\emptyset), \tag{3}$$

where λ is guidance scale, $s_\emptyset$ denotes the unconditional score, is given by $s_\emptyset = \epsilon_U(x_t, t, \varepsilon_\emptyset)$ with $\varepsilon_\emptyset = \psi(\text{“ ”})$. In practice, we noticed that a few prompts are sufficient to identify the concept direction, which is attributed to the generalization ability of the span space $\mathcal{R}_{C_t^i}$.

Algorithm 1 Compositional Concept Learning

```
Require: Diffusion model $\epsilon_U(x_t, t, \varepsilon)$, pre-trained CLIP text encoder $\psi(\cdot)$, guidance scale $\lambda$, concept weight $\{w_1, ..., w_m\}$, covariance matrix $\sigma_t^2 I$, empty prompt “ ”, source prompt $P$, target prompt $P'$, prompts for build concept space $P_{C^i} = \{p_0^i, \ldots, p_j^i, \ldots, p_{N_i-1}^i\}$
Initialize sample $x_t \sim \mathcal{N}(0, I)$
$\varepsilon_\emptyset, \varepsilon_{src}, \varepsilon_{tar} = \psi(\text{“ ”}), \psi(P), \psi(P')$ # Text embedding
$\varepsilon_j = \psi(p_j^i)$ # Concept text embedding
for $t = T, ..., 1$ do
    $s_\emptyset \leftarrow \epsilon_U(x_t, t, \varepsilon_\emptyset)$ # Unconditional score
    $s_{src}, s_{tar} \leftarrow \epsilon_U(x_t, t, \varepsilon_{src}), \epsilon_U(x_t, t, \varepsilon_{tar})$ # Conditional score
    $\Delta s_t = s_{tar} - s_{src}$ # Score difference
    for $i = 1, 2, ..., m$ do # m concepts
        $S_t^i \leftarrow \big[\epsilon_U(x_t, t, \varepsilon_1) - \epsilon_U(x_t, t, \varepsilon_0), \epsilon_U(x_t, t, \varepsilon_2) - \epsilon_U(x_t, t, \varepsilon_0), \ldots, \epsilon_U(x_t, t, \varepsilon_{N_i-1}) - \epsilon_U(x_t, t, \varepsilon_0)\big]$
        Compute the top-$k$ left singular vectors $V_i$ via: $SVD(S_t^i) = V_t^i \Sigma_t^i {U_t^i}^T$
        $M_t^i = V_t^i (V_t^i)^T$ # Projection matrix
        $\Delta C_t^i = M_t^i \Delta s_t$ # Editing direction
    end for
    $s \leftarrow s_\emptyset + \lambda((s_{src} + w_i \Delta C_t^i) - s_\emptyset)$ # For single-concept editing
    $s \leftarrow s_\emptyset + \lambda \sum_{i=1}^m ((s_{src} + w_i \Delta C_t^i) - s_\emptyset)$ # For multi-concept editing
    $x_{t-1} \sim \mathcal{N}(x_t - s, \sigma_t^2 I)$
end for
```

Concept Composition Strategy. Now, our method enables single-concept editing for fashion images, but many scenarios demand multi-concept compositional editing. A straightforward way to achieve this is to create a multi-concept space using a set of concept prompts. While feasible, this approach limits fine-grained control over individual concepts, such as adjusting shades of color in specific items. Alternatively, sequential single-concept edits can be applied, but this accumulates errors and is time-consuming.

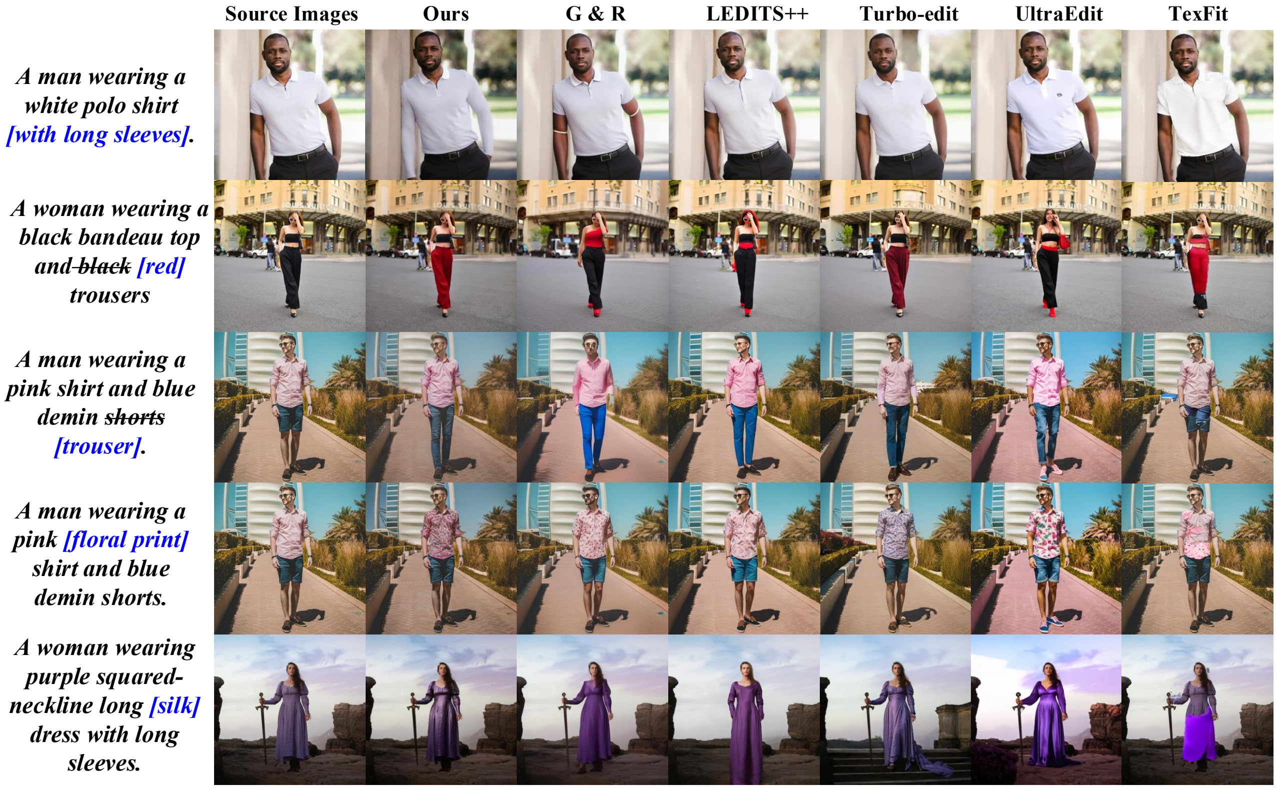
Source Images
Ours
G & R
LEDITS++
Turbo-edit
UltraEdit
TexFit
A man wearing a white polo shirt [with long sleeves].
A woman wearing a black bandeau top and ~~black~~ [red] trousers
A man wearing a pink shirt and blue demin ~~shorts~~ [trouser].
A man wearing a pink [floral print] shirt and blue demin shorts.
A woman wearing purple squared-neckline long [silk] dress with long sleeves.

Figure 3. Qualitative comparison with the state-of-art methods in **single** concept editing.

To overcome these issues, we propose a controllable multi-concept editing strategy. This strategy enables precise, simultaneous adjustments of multiple concepts by weighting each concept's direction vector, which is shown as

$$s_m = s_\emptyset + \lambda \sum_{i=1}^{m} (s_{new}^i - s_\emptyset), \tag{4}$$

where $s_{new}^i = s_{src} + w_i \Delta C_t^i$, we can control the manipulation strength of each concept. The whole algorithm is shown in Algorithm 1.

3. Experiments

3.1. Experimental Setups

Implementation Details. In our work, all experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory. Following the prior work [3, 4, 26], we employ the official pre-trained Stable Diffusion v2.1-base model as our foundational model, downsampling all images to a resolution of 512 × 512 pixels for consistency across experiments. For the counterfactual abduction phase, We fine-tune the model on the single image using LoRA with a rank of 512 for 100 epochs, employing the Adam optimizer with a learning rate of 1e-4. To conserve memory, we adopt the mixed precision [17] and gradient accumulation strategies, setting both the accumulation step count and batch size to 1. During the inference phase, we utilize the DDIM sampling with 80 steps and a classifier-free guidance scale of $\lambda = 4$. Additionally, we explore various values for the concept weight w_i within the range [1, 5] to control the editing strength, where i represents the i-th edited concept.

Dataset. To comprehensively evaluate the performance of our T-FIT in fashion image editing, we follow [14, 26, 40] by collecting a diverse set of fashion images from Unsplash [1], including both full-body and half-body images. The selected images comprise various clothing types, such as long dresses, skirts, T-shirts, shorts, pants, and strapless tops. Our evaluation encompasses a diverse set of editing concepts, which include specific attributes (e.g., gender, color, patterns, fabric texture) as well as broader elements such as clothing type, garment length, and sleeve style. Each image in the experiment is annotated with N_i+2 prompts, consisting of one prompt for the original image and N_i+1 prompts for target images.

Baselines. We conduct comparisons with existing representative diffusion-based image editing methods, including LEDITS [29], LEDITS++ [2], TexFit [32], Guide-and-Rescale (G & R) [28], UltraEdit [41], and Turboedit [6].

Evaluation Metrics. To comprehensively evaluate the performance of our T-FIT, we conduct assessments from both

[1] https://unsplash.com/

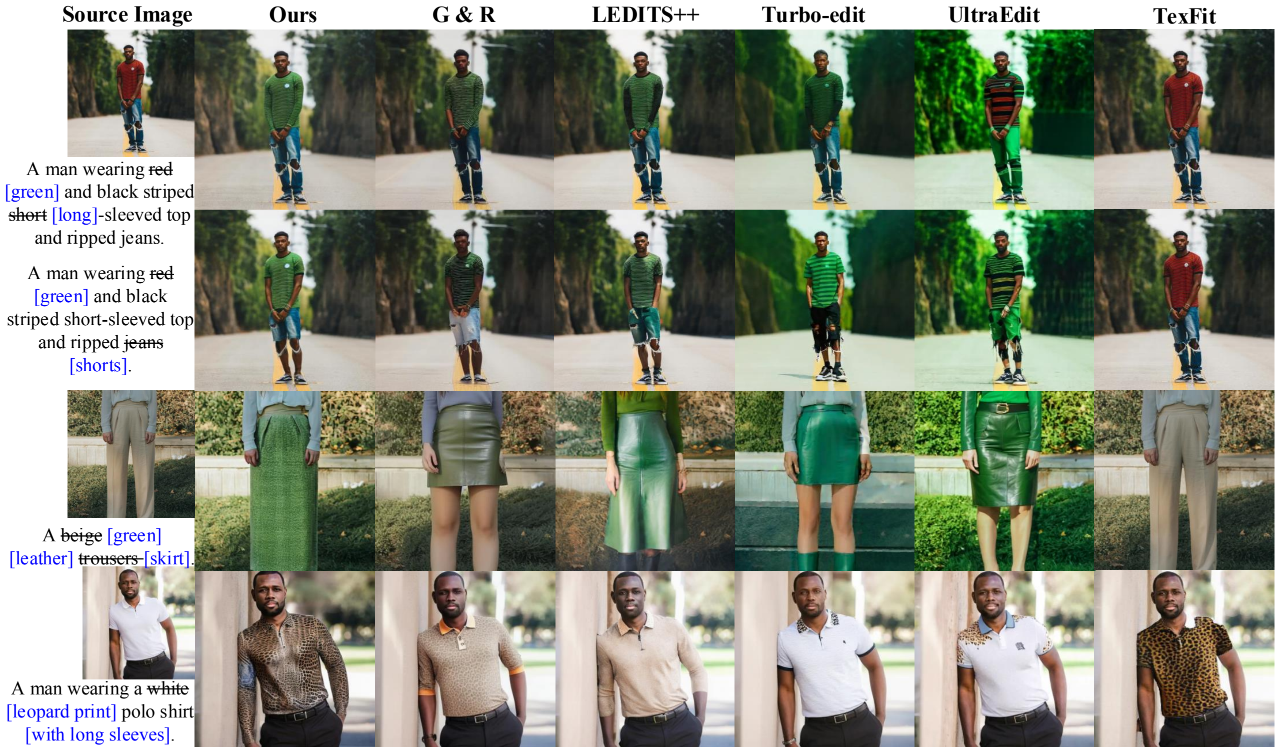
Source Image
Ours
G & R
LEDITS++
Turbo-edit
UltraEdit
TexFit
A man wearing ~~red~~ [green] and black striped ~~short~~ [long]-sleeved top and ripped jeans.
A man wearing ~~red~~ [green] and black striped short-sleeved top and ripped ~~jeans~~ [shorts].
A ~~beige~~ [green] [leather] ~~trousers~~ [skirt].
A man wearing a ~~white~~ [leopard print] polo shirt [with long sleeves].

Figure 4. Qualitative comparison with the state-of-art methods in **multiple** concept editing.

objective and subjective perspectives. Objectively, we follow the previous works [3, 8, 9] and use metrics in CLIP embedding space: CLIP image similarity (CLIP-I) [21] to measure identity preservation via calculate the cosine similarity between the edited image and the original image, and CLIP text-image direction similarity (CLIP-D) [7] to evaluate the correspondence between image changes and text changes. Subjectively, we qualitatively evaluate the fidelity of the edited images and their consistency with the target prompts, supplemented by human preference studies involving recruited volunteers from different disciplines.

3.2. Qualitative Evaluation

We show some qualitative experimental results in Fig. 3- 4, from our experiments, we observe the following:

For single-concept fashion image editing, 1) several methods, such as LEDITS++ [2], introduce unintended changes to non-target areas, possibly due to spurious correlation embedded in the pre-trained models from their training datasets. These correlations cause inaccurate concept associations, leading to changes outside the intended editing area. For instance, in Fig. 3, line 2, LEDITS++ unexpectedly replaces the shoelaces with pink ones. Similarly, in Fig. 3, line 5, UltraEdit [41] transforms a squared neckline into a V-neckline, Turbo-edit [6] alters the dress length. 2) G & R [28], and UltraEdit often confuse different objects. For instance, when attempting to change the color of pants to red (Fig. 3, line 2), these methods inadvertently modify unrelated elements, such as the color of the shirt or hair. 3) TexFit's effectiveness is significantly hampered by its reliance on the accuracy of the recognized mask, which restricts its applicability in fashion image editing. As shown in the last column of Fig. 3, TexFit [32] either alters only localized information in the target editing area or excessively expands the intended modification area. This results in unnatural edits that disrupt the overall coherence of the image.

For multi-concept fashion image editing, 1) UltraEdit [41], TexFit [32] often overlooks certain concepts in multiple-concept editing scenarios. For example, in Fig. 4, line 1, UltraEdit remains in top with short sleeves rather than long sleeves; in Fig. 4, line 2, the trousers edit is ignored. 2) G & R [28], TurboEdit [6], and LEDITS++ [2], demonstrate commendable performance in adhering to user instructions. However, the edited images can exhibit issues such as alterations to human poses or discontinuities in the edited results. For example, when modifying sleeve lengths, the transitions in the sleeve areas may appear inconsistent or fragmented (Fig. 4, line 1). These problems highlight the challenges in maintaining coherence and integrity in the edited images, despite following the provided directives closely. 3) Several methods such as Turbo-edit [6] and UltraEdit [41] may over-edit, making it difficult to retain the details of the original image. For example, when modifying pants into a skirt (Fig. 4, line 3), the edited image may change the length of the skirt as well as other aspects such as the color and fabric of the top along with the skirt. This phenomenon may be attributed to the spurious correlation that were learned by the pre-trained model from the training dataset. Compared to the existing methods, our method excels in fashion image editing, and it can directly control precise multi-concepts from text prompts without additional input, such as masks or poses. The edited image retains non-target areas, closely following the target text while maintaining the consistency of the image. In addition, our method allows continuous editing adjustments by adjusting the concept weight (Fig. 5), which provides fine control over the editing strength. This results in high-quality editing that is superior to the baseline model, especially for multi-concept fashion image editing.

3.3. Quantitative Evaluation

We quantitatively evaluate our T-FIT against baselines using both automatic metrics and human evaluations.

Automatic Metric Comparisons. We summarize the experimental findings in Tab. 1. Our method achieves state-of-the-art results for single- and multi-concept fashion image editing in CLIP-D, which demonstrates a significant advantage in terms of the accuracy and fidelity of our editing results in following the text description. In terms of the CLIP-I metric, our method demonstrates superior performance compared to most competing approaches. Although it slightly underperforms Turbo-edit [6] and TexFit [32] on CLIP-I, this is due to specific limitations in these methods during multi-concept fashion image editing. Turbo-edit often omits certain concepts, resulting in edited images that closely resemble the source, thus boosting its CLIP-I score. TexFit, on the other hand, struggles with precise edit masking, often restricting changes to small areas and failing in non-rigid edits (e.g., changing long sleeves to shorts or pants to shorts), leaving much of the image unaltered. Overall, our method shows superior performance on objective metrics for both multi- and single-concept editing tasks.

Human Preference Study. In this section, we quantitatively evaluate our method with an extensive human perceptual evaluation study. We first collect a diverse set of 30 fashion images and their corresponding text descriptions, covering a range of clothing types (e.g., long sleeves, short sleeves, dresses, pants, shorts), patterns (e.g., solid, stripes, prints), and materials (e.g., cotton, denim, chiffon). Then, we invite 36 volunteers to participate in the evaluation. Each volunteer views the original fashion image, the original description text, the target description text, and the edited images generated by 6 different methods, including ours. The volunteers rate the edited images based on two criteria: (1) how well the edits align with the concepts specified in the target text and (2) the consistency of the non-edited areas with the original image. Ratings range from 1 to 8, with 1 representing the best performance and 8 the poorest. We calculate the average ranks provided by each participant, and the final results are displayed in Fig. 5. Our findings indicate that our method surpasses baseline methods on both criteria, demonstrating enhanced fidelity in executing intended edits while maintaining consistency across unaltered image areas.

Table 1. Quantitative comparisons. The best results are bolded and the second-best results are underlined, respectively.

	Single concept		Multiple concepts	
Method	**CLIP-D**	**CLIP-I**	**CLIP-D**	**CLIP-I**
OUR	**0.1433**	0.9160	**0.1152**	0.9259
G & R	0.0905	0.9116	0.0539	0.9079
Turbo-edit	0.0821	0.9101	0.0967	0.9355
UltraEdit	0.0771	0.9056	0.0533	0.8974
LEDITS++	0.0954	0.9160	0.0584	0.9115
Texfit	0.0346	**0.9500**	0.0463	**0.9562**

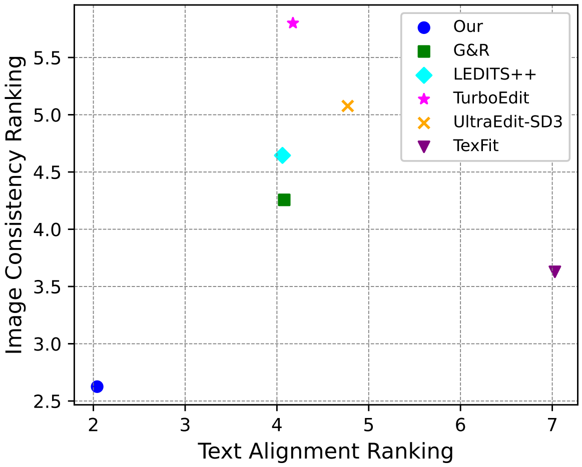

Figure 5. **Human preference study.** Our method outperforms baseline models in both text- and image alignment, demonstrating significant advantages in fashion image editing.

3.4. Ablation Study

Next, we turn to an ablation study where we analyze the effect of the different components inherent in our approach. Specifically, we consider: (1) the effect of substituting the CA with null-text inversion [18], (2) the effect of removing CLM, (3) the effects of varying annealing parameter κ, and (4) the influence of different weights on the editing results of concepts. The visual results are provided in Fig. 6-8.

From these figures, it is evident that when T-FIT replaces the CA with the null-text inversion module, the edited images exhibit a noticeable spurious correlation. For instance, in Fig. 7, when the concept of "*gender*" is applied, the collar of the clothing incorrectly changes from a "*stand-up collar*" to a "*round collar*". On the other hand, T-FIT is equipped

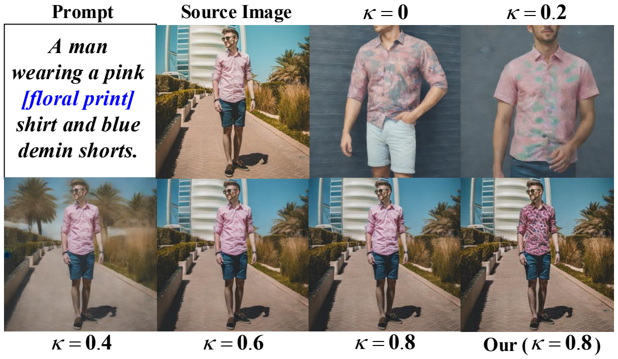

Figure 6. Qualitative ablation results. The last image shows the result of T-FIT, and the other images show the result of T-FIT without CLM and adjusting annealing parameter κ. Removing CLM significantly reduces the model's editability. Adjusting κ still makes balancing editability and fidelity challenging.

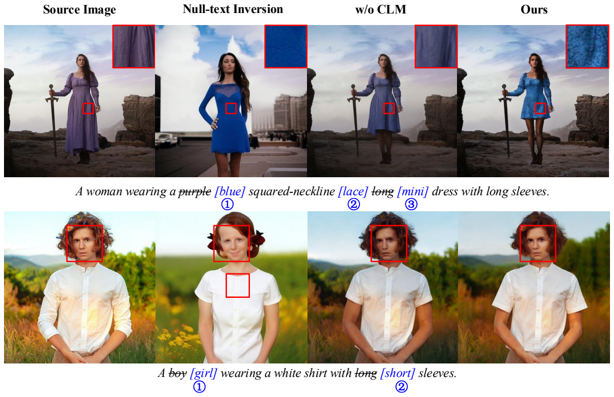
Source Image
Null-text Inversion
w/o CLM
Ours
A woman wearing a ~~purple~~ [blue] squared-neckline [lace] ~~long~~ [mini] dress with long sleeves.
①
②
③
A ~~boy~~ [girl] wearing a white shirt with ~~long~~ [short] sleeves.
①
②

Figure 7. Qualitative ablation results. With CA, our model captures essential visual content from the original image, mitigating the impact of spurious correlation in pre-trained data. CLM ensures that no editing concepts are overlooked during modifications.

solely with CA, it effectively preserves the visual details of the source image and mitigates issues arising from spurious correlation. However, this setup offers limited flexibility for editing, even if we adopt varying annealing parameters κ (see Fig. 6). This suggests that while CA effectively addresses the spurious correlation issues arising from the pre-trained model, it is limited in its editability. When CLM is added, T-FIT's editability improves. This is attributable to the CLM's capability to identify disentangled concepts within the representation space, which facilitates more flexible and precise editing. Furthermore, as illustrated in Fig. 8, T-FIT addresses the issues of concept omission and confusion through the multi-concept composite strategy, while also providing fine control over the intensity of each concept by adjusting individual concept weights. In our work, the concept weights can be manually adjusted in the range [1-5], with higher values indicating greater editing intensity on that concept and vice versa. This offers greater control and flexibility for fashion image editing of our method.

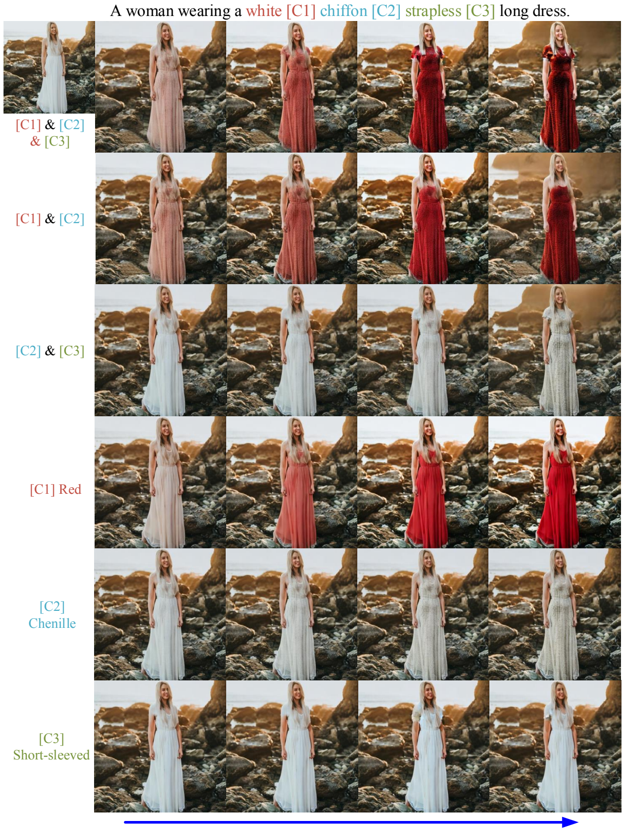

Figure 8. Experimental comparison of varying concept weights (from 1 to 5) on single- and multi-concept fashion image editing. [C1], [C2], and [C3] mean different editing concepts, respectively.

4. Conclusion

In this paper, we introduce T-FIT, a novel framework for fashion image editing. Our approach, which uses a single fashion image and simple text description, supports both single- and multi-concept image editing tasks. By incorporating counterfactual abduction and a concept learning module, T-FIT effectively mitigates issues of spurious correlation in the pre-trained model, enabling efficient, flexible, and accurate text-driven image editing. Furthermore, T-FIT allows fine-grained editing of complex fashion concepts by adjusting concept weights. Extensive experiments confirm the effectiveness of T-FIT and its components, demonstrating significant advantages over several existing methods. Our future work will focus on enhancing editing accuracy by incorporating visual information (e.g., style images) to learn richer visual concept representations.

5. Acknowledgment.

This work was supported by grants from the National Key R & D Program of China (grant no. 2022YFB3303302), and the National Natural Science Foundation of China (grant nos. 62377040, 62477004, 62207007, 623B2002).

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. 1

[2] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. LEDITS++: Limitless image editing using text-to-image models. In *CVPR*, 2024. 1, 2, 5, 6

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 5, 6

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 1, 5

[5] Zhikang Chen, Min Zhang, Sen Cui, Haoxuan Li, Gang Niu, Mingming Gong, Changshui Zhang, and Kun Zhang. Neural collapse inspired feature alignment for out-of-distribution generalization. In *NeurIPS*, 2024. 1

[6] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. *arXiv:2408.00735*, 2024. 1, 5, 6, 7

[7] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):1–13, 2022. 6

[8] Julia Guerrero-Viu, Milos Hasan, Arthur Roullier, Midhun Harikumar, Yiwei Hu, Paul Guerrero, Diego Gutierrez, Belen Masia, and Valentin Deschaintre. Texsliders: Diffusion-based texture editing in clip space. In *SIGGRAPH*, 2024. 6

[9] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *CVPR*, 2024. 6

[10] Mingzhen Huang, Jialing Cai, Shan Jia, Vishnu Suresh Lokhande, and Siwei Lyu. Multiedits: Simultaneous multi-aspect editing with text-to-image diffusion models. In *NeurIPS*, 2024. 2

[11] Shanshan Huang, Haoxuan Li, Qingsong Li, Chunyuan Zheng, and Li Liu. Pareto invariant representation learning for multimedia recommendation. In *ACM MM*, 2023. 1

[12] Shanshan Huang, Qingsong Li, Jun Liao, Shu Wang, Li Liu, and Lian Li. Controllable image synthesis methods, applications and challenges: a comprehensive survey. *Artificial Intelligence Review*, 57(12):336, 2024. 1

[13] Shanshan Huang, Lei Wang, Jun Liao, and Li Liu. Multi-attentional causal intervention networks for medical image diagnosis. *Knowledge-Based Systems*, 299:111993, 2024. 1

[14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 1, 5

[15] Gwanhyeong Koo, Sunjae Yoon, Ji Woo Hong, and Chang D. Yoo. Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. In *ECCV*, 2024. 1

[16] Meng Li and Haochen Sui. Causal recommendation via machine unlearning with a few unbiased data. In *AAAI Workshop on AICT*, 2025. 1

[17] Paulius Micikevicius, Sharan Narang, Jonah Alben, Greg Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *ICLR*, 2018. 5

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 7

[19] Martin Pernuš, Clinton Fookes, Vitomir Štruc, and Simon Dobrišek. FICE: Text-conditioned fashion-image editing with guided gan inversion. *Pattern Recognition*, 158:111022, 2025. 1

[20] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 1

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2

[24] 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. *NeurIPS*, 2022. 1

[25] Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022. 1

[26] Xue Song, Jiequan Cui, Hanwang Zhang, Jingjing Chen, Richang Hong, and Yu-Gang Jiang. Doubly abductive counterfactual inference for text-based image editing. In *CVPR*, 2024. 3, 5

[27] Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Towards compositionality in concept learning. In *ICML*, 2024. 2

[28] Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitry Vetrov, and Aibek Alanov. Guide-and-Rescale: Self-guidance mechanism for effective tuning-free real image editing. In *ECCV*, 2024. 1, 2, 5, 6

[29] Linoy Tsaban and Apolinário Passos. LEDITS: Real image editing with ddpm inversion and semantic guidance. *arXiv:2307.00522*, 2023. 1, 5

[30] Luozhou Wang, Shuai Yang, Shu Liu, and Ying-cong Chen. Not all steps are created equal: Selective diffusion distillation for image manipulation. In *ICCV*, 2023. 3

[31] Lei Wang, Shanshan Huang, Shu Wang, Jun Liao, Tingpeng Li, and Li Liu. A survey of causal discovery based on functional causal model. *Engineering Applications of Artificial Intelligence*, 133:108258, 2024. 1

[32] Tongxin Wang and Mang Ye. TexFit: Text-driven fashion image editing with diffusion models. In *AAAI*, 2024. 1, 5, 6, 7

[33] Xiaolong Wang, Zhi-Qi Cheng, Jue Wang, and Xiaojiang Peng. DPDEdit: Detail-preserved diffusion models for multimodal fashion image editing. *arXiv:2409.01086*, 2024. 1

[34] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In *ICML*, 2023. 3

[35] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *NeurIPS*, 2024. 1, 2

[36] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. In *IJCAI*, 2022. Survey Track. 1

[37] Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. In *NeurIPS*, 2024. 1

[38] Yuntian Wu, Yuntian Yang, Jiabao Sean Xiao, Chuan Zhou, Haochen Sui, and Haoxuan Li. Invariant spatiotemporal representation learning for cross-patient seizure classification. In *NeurIPS Workshop on NeuroAI*, 2024. 1

[39] Min Zhang, Haoxuan Li, Fei Wu, and Kun Kuang. Metacoco: A new few-shot classification benchmark with spurious correlation. In *ICLR*, 2024. 1

[40] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, 2023. 5

[41] Haozhe Zhao, Xiaojian (Shawn) Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. In *NeurIPS*, 2024. 1, 5, 6