# FitTalk: Benchmarking Interactive Virtual Try-On via Multi-Turn Conversations

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Virtual try-on (VTON) has advanced significantly due to generative models, but existing methods remain limited by static, single-pass pipelines that lack iterative refinement, resulting in fragmented textures, color mismatches, and inaccurate garment geometry. This static nature sharply contrasts with real-world fashion interactions, where users naturally engage in multi-turn dialogues to iteratively refine garment selection and fit. To bridge this gap, we introduce **FitTalk**, the first benchmark and framework explicitly designed for *multi-turn interactive virtual try-on*. FitTalk enables iterative garment refinement through natural language dialogue, closely mimicking realistic user-stylist interactions. Our primary contributions include: (1) constructing a large-scale dataset comprising 100,000 interactive try-on dialogues across 15 diverse garment and footwear categories, annotated with detailed refinement instructions and visual artifacts; (2) developing a unified multi-modal conditioning mechanism that integrates garment visuals, textual instructions, and prior dialogue contexts to support coherent, user-guided refinement; and (3) proposing an iterative multi-round training protocol that progressively enriches the model's refinement capabilities by leveraging high-quality, model-generated dialogue examples. Comprehensive evaluations demonstrate that FitTalk significantly outperforms traditional single-turn VTON baselines in both quantitative metrics and qualitative assessments via human preference studies. We also introduce FitMetric, an automated metric leveraging GPT-4o for efficiently evaluating multi-turn refinement quality in terms of garment alignment, color consistency, and identity preservation. By introducing iterative user interactions into the virtual try-on paradigm, FitTalk sets a new standard for interactive fashion generation. We release our dataset and implementation to facilitate future research in interactive, user-centric garment generation. Our project and more results are available at `https://mt-harden.github.io/FitTalk.github.io/`

## 1 Introduction

Virtual try-on (VTON) is rapidly reshaping the online fashion industry by allowing customers to digitally preview garments on personalized avatars, significantly enhancing shopping experiences and reducing returns. Despite substantial advancements driven by generative models, existing systems typically follow static, single-pass pipelines: users select a garment and receive a one-time synthesized image. If the generated result exhibits artifacts—such as fragmented textures, incorrect fit, or color mismatches—there is no mechanism to iteratively refine or correct it through user interaction. This rigid paradigm sharply contrasts with real-world shopping, where users frequently engage in iterative dialogues, requesting adjustments in fit, color, or style until satisfaction is achieved.

Although multi-turn interactions have gained attention in fashion retrieval and recommendation systems, prior works primarily focus on textual item selection rather than dynamic visual synthesis.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

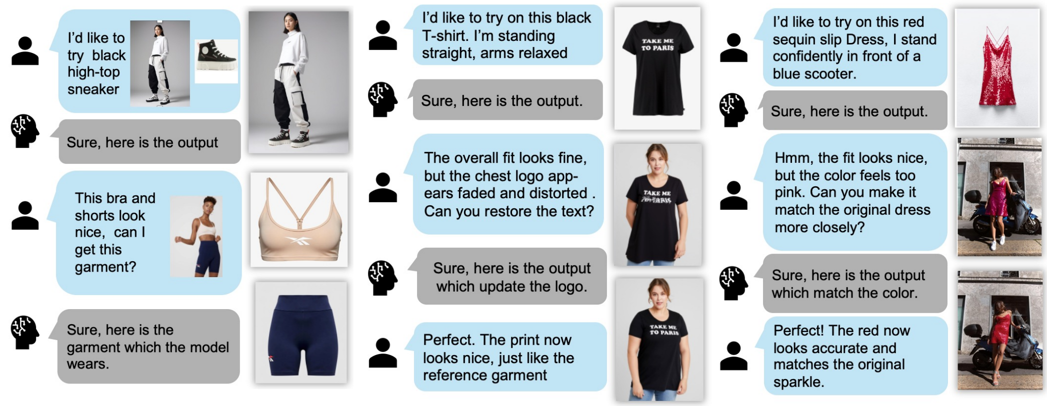


Figure 1: Demonstration of the proposed **FitTalk** system. Our method supports multi-round interactive virtual try-on and try-off for tops, bottoms, and shoes. Given a natural language prompt or an image garment, FitTalk allows users to iteratively refine generation results through dialogue.

Current VTON methods predominantly treat garment generation as a single-step process, failing to incorporate user feedback or preserve conversational context across interactions. This limitation is exacerbated by the absence of suitable multi-turn datasets and benchmarks that enable systematic research into interactive garment refinement.

To address these gaps, we propose **FitTalk**, the first benchmark and framework explicitly designed for multi-turn interactive virtual try-on. FitTalk redefines virtual try-on as an iterative, dialogue-driven refinement process, empowering users to progressively enhance garment generation through natural language interactions. As illustrated in Figure 1, FitTalk allows users to initially select or describe a garment and subsequently refine visual outputs across multiple turns of dialogue, supporting diverse categories such as tops, bottoms, and shoes. Unlike static, single-turn systems, FitTalk enables detailed, conversational adjustments—ranging from minor color corrections to significant style modifications—closely mimicking realistic fitting-room interactions.

To support this novel task, we introduce a large-scale multi-turn dataset containing 100,000 realistic fashion dialogues annotated with corresponding garment visuals and refinement instructions across 15 clothing and footwear categories. Each dialogue explicitly simulates practical refinement scenarios, capturing common visual artifacts encountered during garment synthesis. Moreover, we develop a unified multi-modal conditioning mechanism that seamlessly integrates visual garment features, textual user instructions, and dialogue context, ensuring coherent generation throughout iterative refinements. To further enhance the model's refinement capabilities, we propose an iterative multi-round training strategy, progressively incorporating high-quality, model-generated dialogue examples. Through comprehensive experiments, we demonstrate that FitTalk significantly outperforms traditional single-turn VTON systems across quantitative metrics and qualitative evaluations. Additionally, we introduce **FitMetric**, an automated evaluation metric leveraging GPT-4o, designed specifically for assessing multi-turn refinement quality in terms of garment alignment, color consistency, and identity preservation.

In summary, our key contributions are as follows:

- **FitTalk Dataset**: The first large-scale benchmark dataset for interactive virtual try-on, comprising 100,000 multi-turn fashion dialogues annotated with precise garment visuals, detailed artifact descriptions, and iterative refinement instructions.

- **Interactive Multi-turn Framework**: A unified multi-modal conditioning mechanism and iterative multi-round training protocol designed explicitly to support and enhance conversational refinement in virtual try-on tasks.

- **Comprehensive Evaluation**: Extensive benchmarks demonstrating that interactive multi-turn refinement significantly improves garment synthesis quality, user satisfaction, and alignment with user intent compared to traditional single-turn methods. Our dataset and implementation will be released to support future research.

Table 1: **Comparison with existing virtual try-on datasets.** FitTalk is the only benchmark to offer multi-turn refinement supervision, garment/image captions, and broad clothing category coverage.

| Dataset | Samples | Garment | Garment Caption | Image Caption | Refine Data | Multi-Category |
|---|---|---|---|---|---|---|
| DeepFashionMM [16] | 44K | × | × | ✓ | × | × |
| VITON-HD [7] | 16K | ✓ | × | × | × | × |
| DressCode [12] | 50K | ✓ | × | × | × | ✓ |
| CVDD [23] | 0.5K | ✓ | × | × | × | × |
| FitTalk (Ours) | **100K** | ✓ | ✓ | ✓ | ✓ | ✓ |

# 2 Related Work

## 2.1 Virtual Try-On Systems

Image-based virtual try-on (VTON) aims to digitally dress a person with garments from other images. Early approaches were mostly GAN-based [41, 45, 40, 27, 3, 12, 39, 14, 22, 38], typically following a two-stage pipeline: first warping the clothing to fit the target pose, and then rendering it onto the person. Notable works such as CP-VTON [35] employed Thin-Plate Spline (TPS) transformations for garment warping, followed by refinement networks to preserve texture details. While these methods improved garment-body alignment, they often struggled with occlusion and high-frequency texture preservation due to GAN limitations [36, 13, 43, 15, 20, 7].

More recently, diffusion-based methods [17, 32, 33, 31, 42, 21, 5] have demonstrated superior image quality and robustness. TryOnDiffusion [46] introduced a dual-UNet framework that enables implicit garment warping through cross-attention. Other approaches [32, 17, 5, 9, 41, 15, 20, 3, 29] incorporate garment features via learned embeddings or combine warping with inpainting for fine-grained control. Despite these improvements, nearly all existing diffusion-based VTON models operate in a static, single-turn setting, offering no support for iterative refinement or user-in-the-loop corrections.

Another limitation lies in the data itself. Public benchmarks such as VITON-HD [7] and DressCode [12] offer high-quality person-garment pairs but only support one-shot try-on, without dialogue traces or user-centric feedback. As summarized in Table 1, our benchmark **FitTalk** addresses these gaps by introducing a large-scale, multi-turn dataset annotated with user instructions, artifact types, and iterative try-on images. It spans diverse categories such as tops, pants, dresses, and shoes, enabling systematic study of interactive refinement under realistic conditions.

## 2.2 Multi-Turn Interactive Editing

Beyond virtual try-on, multi-turn interaction [1, 2, 5] has been explored in other vision-language tasks. Sequential AttnGAN [6] pioneered multi-step image generation conditioned on dialogue-style prompts. ChatEdit [11] proposed a benchmark for editing facial images across dialogue turns, emphasizing cumulative reasoning.

More recent works focus on instruction-following editing. InstructPix2Pix [4] fine-tuned diffusion models to follow free-form edit prompts in a single-step setup. DiffEdit [10] introduced masked guidance by comparing prompt denoising trajectories. DialogGen [28] links LLMs and diffusion for conversational generation. However, these models are not designed for identity-preserving or pose-aligned tasks like virtual try-on. They lack garment consistency and cannot handle progressive updates without semantic drift.

**FitTalk bridges these gaps** by enabling multi-turn, user-driven garment refinement through natural language, grounded in both visual references and dialogue history. It turns static try-on into a collaborative, iterative process—closer to real-world fashion interactions.

# 3 Method

## 3.1 Problem Formulation

We formulate interactive virtual try-on as a multi-turn conditional image generation task guided by natural language dialogue. The goal is to enable users to iteratively refine try-on results through conversational feedback, mimicking real-world fitting-room behavior. Given an initial person image

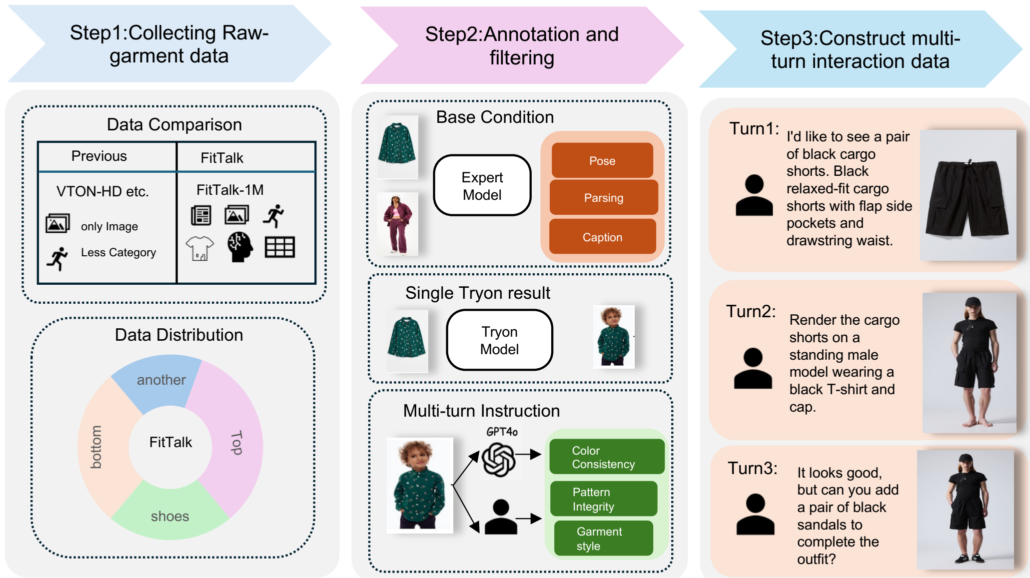


Figure 2: **Overview of FitTalk dataset construction.** We collect person and garment images from both open and proprietary sources. A mask-based generator synthesizes initial try-on results with intentional mask perturbations to simulate user-facing errors. These are filtered using LLMs and human annotators based on predefined artifact types. Multi-turn refinement dialogues are generated via GPT-4o templates and paired with corresponding try-on images.

$I_p$—typically rendered in a clothing-agnostic form where original garments are masked or neutralized—and a sequence of $T$ user utterances $U = [u_1, u_2, \ldots, u_T]$, the system generates a sequence of try-on images $[I_1, I_2, \ldots, I_T]$. Each utterance $u_t$ represents the user's request or feedback at turn $t$, and each image $I_t$ should reflect all garment modifications expressed up to that point. At $t$=1, the user provides either a garment image or a natural language description of the desired item (e.g., "a red velvet dress with short sleeves"). The system synthesizes an initial try-on result $I_1$. In subsequent turns ($t > 1$), the user may issue refinement instructions such as "make the color darker" or "shorten the sleeves," and the system responds with an updated try-on image $I_t$ that applies the requested changes while preserving the person's identity, pose, and all previously accepted modifications.

We formally express the iterative generation process as:

$$I_t = G(I_{t-1}, u_t, \mathcal{H}_{t-1}), \quad t = 1, 2, \ldots, T,$$

where $I_0 \doteq I_p$ denotes the initial state, $u_t$ is the current instruction, and $\mathcal{H}_{t-1}$ represents the dialogue history up to turn $t{-}1$.

The core challenge lies in maintaining visual coherence across turns—ensuring that edits are localized to the relevant garment regions while unrelated content such as facial identity, body shape, or background remains unchanged. To address this, we employ a unified multi-modal conditioning mechanism that integrates textual instructions, visual context, and historical outputs for coherent and controllable image synthesis throughout the dialogue loop.

### 3.2 Dataset Construction

To support multi-turn interactive try-on, we construct **FitTalk**, a large-scale dataset capturing realistic refinement dialogues grounded in diverse fashion scenarios. The pipeline is illustrated in Figure 2.

**Data Collection.** We source person and garment images from both public datasets (e.g., VITON-HD, DressCode) and proprietary web collections to ensure garment diversity. The dataset covers a wide range of garment types, including jackets, coats, dresses, and various footwear such as sneakers and high heels.

**Initial Generation with Controlled Perturbations.** For each person–garment pair, a mask-guided try-on generator synthesizes an initial try-on result. To simulate common generation artifacts, we

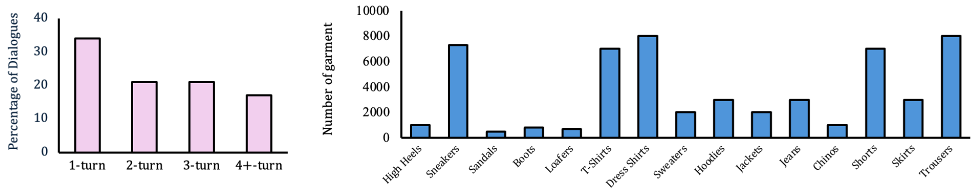


Figure 3: **FitTalk dataset statistics.** Left: Distribution of dialogue lengths in FitTalk, showing the proportion of 1-turn, 2-turn, 3-turn, and 4+-turn conversations. Right: Distribution of garment and footwear categories across the dataset, demonstrating broad and balanced category coverage.

apply controlled perturbations to garment masks (e.g., resizing, shifting), inducing issues such as over-extended regions, color spill, or incomplete rendering.

**Artifact Annotation and Filtering.** The synthesized results are evaluated in a two-stage filtering pipeline. First, a large language model (LLM) identifies potential issues and classifies them into predefined categories: color inconsistency, garment style deviation, pattern integrity loss, or limb distortion. Human annotators then verify the classification and approve instances suitable for dialogue-based refinement.

**Dialogue Synthesis.** Given the filtered images, we use GPT-4o with templated prompts to generate multi-turn user–system interactions. The initial turn includes a garment selection request (image or description), followed by refinement turns that address the identified issues. Each dialogue turn is aligned with a corresponding try-on image and refinement instruction, forming coherent, step-wise supervision for iterative synthesis.

**Dataset Summary.** Overall, **FitTalk** contains over 100,000 multi-turn dialogue trajectories spanning 15 diverse garment and footwear categories. Each dialogue consists of 1–4+ turns, with each turn aligned to a generated try-on image and a structured refinement instruction. The left side of Figure 3 shows the proportion of dialogues by their total turn count: 64% of all conversations contain at least two turns The right side of the figure shows the sample count distribution across categories, highlighting balanced coverage across tops, bottoms, outerwear, and shoes. This diversity ensures that FitTalk supports learning robust multi-turn generation across varied garment types.This pipeline enables construction of high-quality refinement dialogues with diverse artifacts and instructions.

### 3.3 FitTalk Architecture

Our generative model $G$ is implemented as a multi-stream Multi-Modal Diffusion Transformer (MMDiT) architecture designed explicitly for interactive multi-turn refinement. At each conversation turn, the model receives three input streams: the previous generated person image ($I_{t-1}$), a garment reference image ($I_g$), and the user's current textual instruction ($u_t$), integrated with historical context ($\mathcal{H}_{t-1}$).

Each input is encoded into a dedicated token sequence within a unified generation process. The Generation Stream represents the target try-on image to be synthesized and is initialized from diffusion latent embeddings. The Visual Reference Stream encodes visual contexts, including the previous turn's generated image ($I_{t-1}$) and the current garment reference image ($I_g$), which together provide spatially structured visual guidance. Finally, the Instruction Stream captures the user's textual prompt ($u_t$), using a pre-trained language encoder to model user intent and semantic conditions. These three streams are fused via a shared transformer backbone to enable multimodal conditioning across visual and textual domains.

We fuse these token sequences into a unified multi-modal sequence and feed it into a shared Diffusion Transformer backbone. Through multi-head cross-modal attention, tokens dynamically interact across visual and textual modalities, effectively grounding linguistic instructions into specific spatial locations. This flexible fusion enables precise localization of edits and maintains visual consistency (e.g., identity, pose, background) across multiple conversational refinements.

Crucially, historical dialogue context ($\mathcal{H}_{t-1}$) is incorporated by concatenating past textual instructions and visual embeddings from previous turns, ensuring coherent and progressive refinement throughout the dialogue interaction.


Figure 4: Qualitative comparison with GPT-4o on multi-turn try-on tasks. Given a garment and sequential user instructions, our FitTalk system better preserves pattern details, restores correct color tones, and avoids garment distortion compared to both single-turn baselines and GPT-4o.

### 3.4 Iterative Multi-round Training Protocol

To effectively handle iterative refinement dialogues, we adopt a progressive multi-round training strategy. Our goal is to gradually expose the model to increasingly complex dialogue interactions and refinement instructions.

**Round 0: Multi-task Initialization.** We first pre-train the model with multi-task learning on diverse conditional patterns, randomly sampling tasks at each training step. Tasks include (a) caption-to-garment generation, (b) caption-to-avatar synthesis, (c) standard single-turn try-on (garment image with caption), (d) image-only try-on/try-off, and (e) single-step refinement from previous outputs with corrective instructions. We randomly drop or replace input modalities (image or text) with textual descriptions or noise with probability $p_{\text{drop}} = 0.2$ to encourage model robustness under incomplete input scenarios. At this stage, we optimize only the inserted Low-Rank Adaptation (LoRA) modules, keeping the core diffusion model parameters frozen.

**Iterative Dialogue Harvesting (Rounds 1 to R).** After the initial training epoch, we iteratively enrich our training corpus by harvesting high-quality multi-turn dialogues generated by the current model. Specifically, at each epoch $k$, we use the trained model $G^{(k)}$ to synthesize dialogues on new person–garment pairs from a held-out set. Generated dialogues are automatically evaluated by FitMetric based on garment alignment, color consistency, and identity preservation scores. Only dialogues surpassing stringent quality thresholds are manually verified and retained as high-quality examples $\mathcal{F}^{(k)}$. We progressively expand the training set for epoch $k+1$ by merging these harvested dialogues with the initial seed dataset:

$$\mathcal{D}^{(k+1)} = \mathcal{D}^{(0)} \cup (\mathcal{F}^{(0)} \cup \cdots \cup \mathcal{F}^{(k)})$$

During each subsequent round, we maintain an $80\%/20\%$ ratio of seed-to-harvested dialogues per mini-batch. This curriculum learning approach ensures model exposure to gradually more challenging and realistic refinement scenarios. Empirically, we observe convergence in validation performance after three iterative harvest-and-train cycles ($R = 3$), indicating the effectiveness of our iterative multi-round training protocol.

# 4 Experiment

## 4.1 Experimental Setup

We design our experiments to assess the effectiveness of **FitTalk** in addressing the key challenges of virtual try-on. Specifically, we evaluate: (i) the typical failure patterns of existing single-turn methods, (ii) the ability of FitTalk to correct these artifacts through multi-turn refinement, and (iii) comparative performance across standard image quality metrics and human preference scores.

All models are trained and evaluated using 8× NVIDIA A100 (80GB) GPUs. We adopt Flux [26] as the base diffusion model, extending its architecture with multi-modal conditioning and iterative generation. Each training round requires approximately three days to complete. All images are processed at a resolution of 1024×768 during training and evaluation. Additionally, we provide a 1536-resolution version of the FitTalk dataset to support high-fidelity generation and future research at larger scales.

## 4.2 Single-turn Error Analysis

Before introducing multi-turn interaction, we analyze failure patterns in state-of-the-art single-pass try-on systems. We collect 3,000 outputs from four representative pipelines—two open-source diffusion-based models and two commercial APIs (Meitu[1], HuiWa[2])—and annotate each image based on a refined taxonomy of visual artifacts, summarized from extensive audit sheets.

Table 2: **Distribution of Failure Types in Single-turn Try-on.**

| **Error Type** | **Count** |
|---|---|
| Colour inconsistency | 900 |
| Style deviation | 540 |
| Unreasonable generation | 570 |
| Garment hallucination | 450 |
| Pose misalignment | 271 |
| Material failure | 269 |

As shown in Figure 2, the most prevalent issues are **color inconsistency** (e.g., hue mismatches or tone shifts) and **style deviation** (e.g., inaccurate silhouette or length), indicating limited understanding of fine-grained garment attributes. **Unreasonable generations** such as missing limbs or distorted body parts remain common, often caused by mask errors or diffusion instability. **Garment hallucination** refers to undesired duplication or insertion of clothing elements, while **material failures** include pattern blurring and texture breakdown. **Pose misalignment** frequently occurs around occluded limbs and non-frontal poses, where garment warping is misapplied.

Together, these findings reveal a pattern: most failures are localized, interpretable, and amenable to explicit user correction. This motivates the interactive refinement paradigm of FitTalk, which enables users to iteratively resolve such issues through targeted natural language feedback.

## 4.3 Evaluation

We adopt both standard image generation metrics and a specialized multi-turn evaluation protocol to assess the performance of FitTalk.

**Standard Metrics.** Following prior work, we evaluate try-on image quality using four widely-used metrics. Structural Similarity Index (SSIM) [37] and Learned Perceptual Image Patch Similarity (LPIPS) [44] quantify low-level structural similarity and perceptual closeness to ground truth references. To measure distribution-level realism, we report Fréchet Inception Distance (FID) [30] and Kernel Inception Distance (KID) [34]. Higher SSIM and lower FID, KID, LPIPS scores indicate better visual quality.Additionally ,We also conducted a user study with 100 participants. For multi-turn tryon task (e.g., color change, fit adjustment), participants were shown paired results from FitTalk and GPT4o and asked to select the one that better matched the instruction.

**FitMetric.** To evaluate iterative refinement performance, we introduce **FitMetric**, a multi-turn evaluation protocol capturing three key dimensions: garment alignment, color consistency, and identity preservation. For each turn, GPT-4o serves as an automatic evaluator and scores the try-on image on a scale from 0 to 4 in each dimension. These raw scores are normalized to [0,1] and

[1] `https://www.designkit.com/`
[2] `https://www.ihuiwa.com/`

Table 4: Quantitative comparisons across methods. Our method achieves the best performance under both standard and multi-turn-specific metrics.

| Method | SSIM ↑ | FID ↓ | LPIPS ↓ | KID ↓ | FitMetric ↑ |
|---|---|---|---|---|---|
| PF-AFN [15] | 0.885 | 9.616 | 0.087 | 3.85 | 0.78 |
| FS-VTON [20] | 0.881 | 9.735 | 0.091 | 3.69 | 0.71 |
| SDAFN [3] | 0.881 | 9.497 | 0.092 | 2.73 | 0.63 |
| GP-VTON [38] | 0.893 | 9.405 | 0.079 | 0.88 | 0.80 |
| DCI-VTON [17] | 0.868 | 9.166 | 0.096 | 1.10 | 0.76 |
| StableVTON [24] | 0.866 | 8.992 | 0.079 | 1.03 | 0.83 |
| Any2AnyTryOn [19] | 0.852 | 9.98 | 0.117 | 3.50 | 0.83 |
| FitDiT [23] | 0.838 | 8.18 | 0.096 | 1.10 | 0.84 |
| PromptDresser [25] | 0.846 | 8.53 | 0.104 | 0.89 | 0.82 |
| **FitTalk(Ours)** | **0.897** | **8.18** | **0.07** | **0.88** | **0.9** |

averaged across turns to produce a final multi-turn quality score. This enables scalable and reliable evaluation without exhaustive manual labeling.

### 4.4 Do Multi-turn Dialogues Improve Performance?

To evaluate the impact of interactive refinement, we compare our method against two sets of baselines: (i) state-of-the-art single-turn virtual try-on models, and (ii) a GPT-4o-based instruction-following editing pipeline. All methods are tested under the same garment-person pairs and user instructions on the VITON-HD [8], DressCode [12], and FitTalk-Test datasets.

**Single-turn Comparison.** We first examine single-step generation quality. As shown in Table 4, our FitTalk-single outperforms strong baselines such as DCI-VTON [18], StableVTON[24], and PromptDresser [25] across FID, LPIPS, and SSIM. In particular, the improvement in FitMetric demonstrates better alignment with user intent, even without iterative feedback.

**Multi-turn Comparison.** To measure the benefits of iterative dialogue, we compare our full FitTalk pipeline with GPT-4o, a powerful general-purpose instruction-following vision-language model. Both systems are evaluated using the same multi-turn user inputs and garment references. Table 3 highlights our superior performance on all key metrics. Additionally, Figure 4 provides qualitative comparisons showing that FitTalk produces more consistent, accurate, and visually aligned refinements, including pattern matching, color harmonization, and geometric correction.

Table 3: **Multi-turn Evaluation: FitTalk vs. GPT-4o.** Results on the FitTalk-Test subset using identical instructions.

| Method | FID ↓ | FitMetric ↑ | Human Study |
|---|---|---|---|
| GPT-4o | 10.38 | 0.84 | 0.35 |
| FitTalk | **8.18** | **0.9** | **0.65** |

### 4.5 Ablation Study

We further analyze key components contributing to the effectiveness of our FitTalk framework. Specifically, we investigate the benefits of multi-turn refinement, the impact of iterative multi-round training, and the influence of dataset scale and diversity on model performance. We systematically quantify these effects through controlled experiments, summarized in Fig. 5.

**Impact of Multi-turn Refinement.** We first evaluate whether multi-turn interactions help improve try-on results. As shown in Fig. 5(a), FitMetric and FID significantly increases from single-turn (Turn-1) to multi-turn (Turn>2) dialogues. This result clearly indicates the practical effectiveness of interactive refinement in aligning the final generation closer to user expectations.

**Effectiveness of Iterative Multi-round Training.** Next, we measure the effectiveness of our iterative multi-round training strategy. Fig. 5(b) demonstrates a clear improvement in both FitMetric and FID when comparing single-round training (Round-1) with iterative multi-round training (Round-2 and Round-3). This confirms that integrating selectively harvested refinement dialogues from previous training rounds effectively enhances the model's generation quality and consistency.

**Influence of Dataset Scale and Diversity.** Lastly, we examine the impact of dataset size and category diversity. Fig. 5(c) compares models trained on the VITON-HD dataset versus our larger-scale and more diverse FitTalk dataset across two metrics (FID, FitMetric). Training on the expanded FitTalk dataset leads to consistent and significant performance gains across all metrics, validating that increasing dataset diversity and scale plays a critical role in boosting the overall generation fidelity and generalization capabilities.

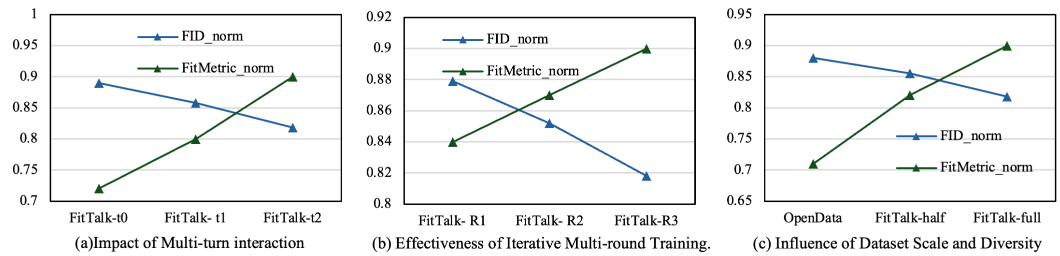


Figure 5: Ablation Study with Normalized Scores. (a) Normalized comparison of FID, LPIPS, and FitMetric across single-turn and multi-turn settings. (b) Effects of iterative multi-round training (R1 to R3) on normalized performance. (c) Impact of dataset scale and diversity using normalized results on OpenData, FitTalk-half, and FitTalk-full. All scores are normalized to [0,1] for fair visualization.

Overall, these results systematically demonstrate the individual contributions of key factors in the proposed FitTalk system, emphasizing the practical necessity of iterative feedback, progressive training, and data diversity for interactive virtual try-on tasks.

## 5 Conclusion

We introduced **FitTalk**, the first benchmark and framework explicitly designed for **multi-turn interactive virtual try-on**. By redefining try-on as an iterative, dialogue-driven process, FitTalk enables users to refine garment appearance through natural language interactions—closely emulating real-world fitting-room dynamics.

Our contributions include: (i) a large-scale dataset of 100,000 multi-turn dialogues across 15+ garment and footwear categories, annotated with visual artifacts and refinement instructions; (ii) a unified multi-modal conditioning architecture that fuses garment images, user instructions, and dialogue history; and (iii) an iterative multi-round training strategy that incrementally improves refinement ability via model-generated dialogue harvesting.

Comprehensive experiments demonstrate that FitTalk substantially outperforms existing single-turn VTON systems across standard image quality metrics and human preference studies. By enabling controllable, user-guided refinement, FitTalk sets a strong foundation for future research in interactive fashion generation.

**Social Impact.** FitTalk promotes more efficient and sustainable fashion workflows by reducing reliance on physical prototyping and enabling intuitive user interaction. During dataset construction, we took precautions to avoid identity exposure, culturally sensitive attire, and inappropriate visual content. The dataset emphasizes diversity in gender, pose, and garment type to support inclusive and fair modeling. We encourage future applications of virtual try-on to further consider fairness, representation, and downstream social impacts.

**Limitations.** While FitTalk enables robust multi-turn garment refinement, it currently lacks support for fine-grained compositional edits—such as localized texture manipulation or conditional logic (e.g., "only add a zipper if the jacket is leather"). This is due to the absence of dense attribute annotations and fine-grained grounding in current training data. We are actively addressing this through ongoing annotation expansion. Additionally, the current dataset focuses on frontal poses and common apparel; future work may explore broader body types, view angles, and real-time interactive systems. =

## References

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

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

[3] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, 2022.

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2023.

[5] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.

[6] Bowen Cheng, Xiaojiang Liu, Lin Li, and Chang Liu. Sequential attention gan for interactive image editing via dialogue. In *ICCV*, 2019.

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

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021.

[9] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024.

[10] Guillaume Couairon, Jakob Verbeek, and Patrick Perez. Diffedit: Diffusion-based semantic image editing with mask generation. *arXiv preprint arXiv:2210.11427*, 2022.

[11] Ling Cui, Zhaoxin Chen, Pengfei Liu, and Xiaodong Zhang. Chatedit: Multi-turn interactive editing via conversational dialogue. *arXiv preprint arXiv:2302.03767*, 2023.

[12] Morelli Davide, Fincato Matteo, Cornia Marcella, Landi Federico, Cesari Fabio, and Cucchiara Rita. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[13] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[14] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[15] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[16] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019.

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

[18] Xiaodan Gou, Yangyang Liu, Yadan Wu, Jianfeng Liu, Yong Yang, and Hengtao Wang. Diffusion-based conditional inpainting for virtual try-on. *arXiv preprint arXiv:2310.15489*, 2023.

[19] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks, 2025.

[20] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[21] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

[22] Zaiyu Huang, Hanhui Li, Zhenyu Xie, Michael Kampffmeyer, Xiaodan Liang, et al. Towards hard-pose virtual try-on via 3d-aware global correspondence learning. *Advances in Neural Information Processing Systems*, 2022.

[23] Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on, 2024.

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

[25] Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask, 2024.

[26] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[27] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, 2022.

[28] Ming Li, Haotian Liu, Yikang Shen, Jian Tang, Mingxuan Wang, and Jian Yin. Dialoggen: Dialogue-driven image generation via large-language models. *arXiv preprint arXiv:2305.07325*, 2023.

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

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

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

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

[33] 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. *Advances in Neural Information Processing Systems*, 2022.

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

[35] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, and Liang Lin. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018.

[36] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

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

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

[39] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, Xin Dong, Feida Zhu, and Xiaodan Liang. Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on. *arXiv preprint arXiv:2207.13475*, 2022.

[40] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems*, 2021.

[41] Zhenyu Xie, Xujie Zhang, Fuwei Zhao, Haoye Dong, Michael C Kampffmeyer, Haonan Yan, and Xiaodan Liang. Was-vton: Warping architecture search for virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

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

[43] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

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

[45] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[46] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

# A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline (see above), or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

# NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

   Answer: [Yes]

   Justification: Yes ,it’s accurately

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: I have discussed in conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes I have provided it.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: all of result can be reproduce.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: I have provided it

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: I have provided it

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes I do

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: Yes I do

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes I do

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes I do.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: yes I do

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer:[Yes]

    Justification: Yes I do

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.