

Perceptual Evaluation of Garment Drape Intensity in AI-Generated Garment Images

A Pilot Study Using a Semi-Automated Evaluation Pipeline with CLIP

Noriaki Kuwahara^{1*}, Yuta Yamamoto², Dongeun Choi³, Kazunari Hirakoso⁴, Takashi Satou⁵

Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto 606-8585, Japan^{1, 2, 3}

Tamagawa University, 6-1-1 Tamagawa-Gakuen, Machida, Tokyo 194-8610, Japan⁴

Tokyo Denki University, 5 Senju-Asahi-cho, Adachi-ku, Tokyo 120-8551, Japan⁵

Abstract—Recent advances in generative artificial intelligence have enabled fine-grained manipulation of visual attributes in garment images, but consistency in automated feature space does not necessarily guarantee consistency in human perception. This study investigates the perceptual validity of garment drape intensity control in AI-generated garment images using a semi-automated Human-in-the-Loop evaluation pipeline. Five base garments were manipulated across five latent-space control levels and evaluated independently using Contrastive Language–Image Pre-training (CLIP)-based feature-space analysis and a two-alternative forced-choice (2AFC) experiment with 15 participants. The CLIP projection scores increased monotonically with the manipulation parameter for all five garment sequences. Human judgments showed a similar monotonic tendency at the aggregate level; however, garment-specific Bradley–Terry analysis revealed local reversals in the estimated perceptual ordering for three of the five base garments. These results demonstrate that aggregate agreement between automated and human evaluation can mask stimulus-specific perceptual discrepancies. The proposed framework provides a practical approach for combining automated screening with human perceptual validation when assessing fine-grained attributes in generative fashion systems.

Keywords—*Drape perception; generative AI; human-in-the-loop; CLIP; bradley–terry model; fashion image generation*

I. INTRODUCTION

Generative models have increasingly enabled fine-grained control of visual attributes in synthesized images. Diffusion-based and latent diffusion models have provided a flexible foundation for high-quality and controllable image generation [1], [2]. In the fashion domain, generative models are being applied to garment generation, editing, virtual try-on, and design support [3]. Beyond generating visually plausible garments, an important challenge is to determine whether controlled changes introduced by generative models correspond to visual differences that are consistently perceived by human observers.

This issue is particularly relevant for perceptual garment attributes. A generative model may produce systematic changes in its latent representation or in an automated feature space, but such changes do not necessarily imply equally systematic changes in human perception. Recent work on human-preference-based evaluation of generated images has also highlighted the importance of evaluating correspondence between automated representations and human judgments [4]. Drape provides a useful case for examining this problem

because its visual impression emerges from multiple garment-dependent characteristics, including folds, silhouette, and the apparent manner in which fabric hangs. The perceived strength of drape may therefore depend not only on the magnitude of the generative manipulation but also on the visual context of the individual garment.

Our previous study proposed an evaluation framework combining latent-space analysis, CLIP-based automated evaluation, and human perceptual judgments to assess the reliability of attribute control in diffusion-based fashion generation [5]. The present study extends this approach to garment drape as a textile-oriented perceptual attribute. Latent-space manipulation is employed because it enables systematic variation of a target attribute along a continuous control parameter while using the same base garment as a reference. This controlled setting is particularly useful for perceptual evaluation, where changes in the target attribute should be distinguished from unrelated variations in garment appearance. Rather than examining only whether an automated score changes monotonically with the manipulation parameter, we investigate whether the corresponding perceptual ordering is reproduced by human observers and whether this correspondence remains consistent across individual base garments. In this study, the term drape intensity is used operationally to denote the perceived strength of drape in AI-generated garment images and is not intended to represent a conventional physical property of textiles.

To address these questions, we employ a semi-automated Human-in-the-Loop evaluation pipeline integrating latent-space quality control, CLIP-based feature-space evaluation, and a two-alternative forced-choice (2AFC) perceptual experiment analyzed using the Bradley–Terry model. The main contributions of this study are threefold: 1) we operationalize drape intensity as a perceptual attribute for controlled evaluation of AI-generated garment images; 2) we apply and extend our previously proposed semi-automated evaluation framework to garment drape, with automated feature-space behavior and human perceptual judgments evaluated independently and subsequently compared; and 3) we demonstrate that aggregate agreement between automated and human evaluation can mask garment-specific perceptual discrepancies. Through this analysis, the study examines the role of human perceptual validation in assessing fine-grained attribute control in generative fashion systems.

*Corresponding author.

II. RELATED WORK

A. Textile Drape Measurement and Perceptual Evaluation

Drape has long been studied in textile and apparel engineering as a property arising from the interaction between fabric mechanical characteristics, gravity, and garment construction. Early quantitative approaches introduced drape coefficients and related geometric measures to characterize the deformation of fabric under its own weight [6]. Subsequent work extended such physical evaluation by examining static and dynamic drape behavior and its relationship to fabric characteristics [7].

In parallel with physical measurement, subjective and visual evaluations have been used to investigate how drape is perceived as part of garment appearance. Studies have examined the aesthetic impression of drape [8], the effects of material properties on gathered skirts and their visual evaluation [9], and the formation and visual characteristics of drape in flared-skirt models [10, 11]. Other studies have related the silhouettes of flared and gathered skirts to fabric mechanical, bending, and vibration properties [12, 13]. Together, these studies demonstrate that garment drape is not solely a physical material property but also a perceptual phenomenon influenced by garment structure and visual appearance.

The present study differs from this conventional line of textile research in that no physical fabric is measured. Instead, we examine the perceived drape of synthetically manipulated garment images generated by a diffusion model. Accordingly, drape intensity is used here as an operational perceptual term rather than as a substitute for conventional physical measures such as the drape coefficient. This distinction allows established knowledge of garment drape perception to be connected with the emerging problem of perceptual validation in generative fashion imagery.

B. Generative AI and Attribute Control in Fashion

Recent progress in image generation has shifted attention from the synthesis of realistic images toward controllable generation and editing. Diffusion models provide a flexible foundation for this transition by progressively reconstructing images from noisy representations [1]. Latent diffusion further reduces the computational burden of this process by performing generation in a compressed representation space, enabling practical high-resolution synthesis and editing [2]. Additional conditioning mechanisms have subsequently been introduced to provide more explicit control over generated content. For example, ControlNet incorporates structural conditions into diffusion-based generation [14], while instruction-guided approaches enable existing images to be modified according to natural-language descriptions [15].

Another line of research has investigated whether interpretable visual changes can be produced by manipulating internal representations of generative models. Studies based on GANs demonstrated that variations in attributes can be associated with directions or subspaces within learned latent representations. InterFaceGAN identified semantic boundaries that support controlled manipulation of facial attributes [16], whereas GANSpace derived interpretable controls from principal directions in intermediate latent representations [17].

StyleCLIP further demonstrated that language-based semantic information can be connected to latent editing operations [18]. Although these approaches differ in architecture and manipulation strategy, they collectively show that learned representations can support continuous and interpretable changes in generated images.

Such controllability is particularly relevant to fashion, where the usefulness of generative systems depends not only on image realism but also on the ability to preserve garment identity while modifying selected characteristics. Recent reviews describe applications of generative AI to fashion design, garment synthesis, virtual try-on, and image editing, while also identifying fine-grained controllability and evaluation as continuing challenges [3]. However, successful manipulation in a generative representation does not by itself establish that the intended attribute change is visually interpreted in the same way by human observers. For perceptual garment attributes such as drape, this distinction motivates an evaluation framework that considers not only the behavior of the generative representation but also the resulting human perceptual response.

C. Automated Evaluation and Human Perceptual Alignment

The increasing capability of generative models has created a corresponding need for scalable methods to evaluate generated images. Vision-language models provide one approach by representing visual and linguistic information within related feature spaces. CLIP, trained through contrastive learning on large-scale image-text pairs, has become a widely used representation for measuring semantic relationships between images and textual concepts [19]. Other approaches have moved beyond general image-text similarity toward models trained to approximate human preferences for generated images. Examples include ImageReward [20] and Human Preference Score v2 [4], which use human preference data to improve the correspondence between automated evaluation and subjective judgments.

The existence of such preference-oriented metrics also reflects an important limitation of purely automated evaluation: a systematic change in a learned representation is not necessarily equivalent to a systematic change in human perception. Deep perceptual metrics such as LPIPS have demonstrated that learned visual representations can capture aspects of perceptual similarity more effectively than simple pixel-level measures [21]. Nevertheless, perceptual judgments remain dependent on the task, the visual attribute under consideration, and the context in which that attribute appears. Consequently, automated metrics are useful for characterizing generated-image behavior but cannot automatically be treated as substitutes for task-specific human validation.

Human paired-comparison methods provide a complementary means of evaluating subtle perceptual differences without requiring participants to assign absolute numerical values to visual attributes. In a two-alternative forced-choice (2AFC) experiment, observers select one of two stimuli according to a specified perceptual criterion, allowing comparative judgments to be collected under a relatively simple task design [22]. Models for paired-comparison data, such as the Bradley-Terry model, can then transform these outcomes into latent worth estimates representing relative perceptual strength

[23]. This combination is particularly useful when the experimental question concerns the ordering of a continuously manipulated attribute rather than the assignment of an absolute physical quantity.

In our previous work, we combined latent-space diagnostics, CLIP-based evaluation, and human paired-comparison judgments to investigate perceptual reliability across multiple controlled attributes in diffusion-based fashion generation [5]. That study showed that agreement among the different evaluation layers varied across attributes, motivating closer examination of cases in which automated and perceptual assessments do not coincide. The present study takes a more focused approach by treating garment drape as a textile-oriented perceptual attribute and by examining correspondence at the level of individual base garments. This shift from multi-attribute aggregate validation to garment-specific analysis enables us to investigate whether apparently consistent automated trajectories can conceal stimulus-dependent differences in human perceptual ordering.

III. METHOD

A. Overview of the Semi-Automated Evaluation Pipeline

To evaluate whether latent-space attribute manipulation produces perceptually meaningful changes in garment drape, we developed a semi-automated evaluation pipeline that integrates AI-based stimulus generation, human perceptual evaluation, and automated feature-space evaluation into a unified experimental workflow. Rather than relying solely on either human judgment or automated metrics, the proposed pipeline enables systematic comparison between the two evaluation tracks, thereby providing a reproducible framework for assessing the perceptual validity of latent attribute manipulation.

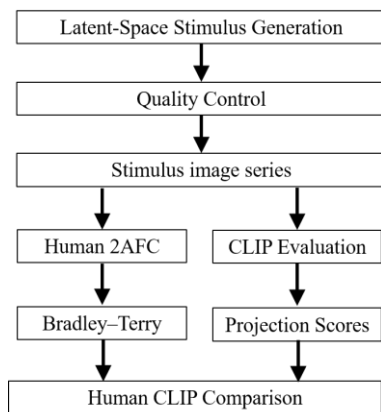


Fig. 1. Overview of the semi-automated evaluation pipeline. Human perceptual evaluation and CLIP-based feature-space evaluation are conducted as parallel tracks and subsequently compared to assess perceptual consistency.

As illustrated in Fig. 1, the pipeline consists of five functional stages: 1) stimulus generation by latent-space manipulation, 2) quality control of generated stimuli, 3) human perceptual evaluation using a web-based two-alternative forced-choice (2AFC) task, 4) CLIP-based automated feature-space evaluation, and 5) comparative analysis between Bradley-Terry perceptual scores and CLIP projection scores. Human evaluation and automated evaluation were performed independently and compared only during the final analysis stage.

B. Stimulus Generation

Stimulus images were generated using a Stable Diffusion-based inpainting pipeline with the Realistic Vision V5.1 Inpainting checkpoint (Realistic_Vision_V5.1-inpainting.safetensors). Generation was performed at a resolution of 512×512 pixels using 35 inference steps, a guidance scale of 7.5, and a denoising strength of 0.7. An initial mannequin image and a garment-region mask were used for inpainting. A randomly initialized seed was used for each generation and recorded in the corresponding generation metadata to enable traceability of the generated stimuli.

Five female garment images were manually selected from the generated candidates and used as base stimuli (Dress A–E). The selection was intended to provide visually distinct garment examples for this pilot evaluation; no formal probabilistic or stratified sampling procedure was applied. Accordingly, these five garments should be regarded as an exploratory stimulus set rather than as a representative sample of a broader garment population. The base condition was generated using the prompt “moderate drape, gentle folds clearly visible, slightly gathered at the waist.” For each selected base garment, weak- and strong-drape endpoint images were subsequently generated using the same inpainting framework. The minimum-drape condition was specified using the prompt “studio product photo of a mannequin torso wearing a dress with no visible drape, smooth fitted surface, minimal folds, matte fabric, clean background,” whereas the maximum-drape condition was specified as “studio product photo of a mannequin torso wearing a dress with strong drape, deep flowing folds, gathered at the waist, matte fabric, clean background.” These descriptions served as operational anchors for the drape manipulation rather than as established physical definitions of minimum and maximum fabric drape.

Each generated endpoint image was encoded using the variational autoencoder (VAE) associated with the diffusion pipeline, and the resulting latent representation was stored after application of the model scaling factor. Thus, the present procedure did not employ a separate latent inversion algorithm such as DDIM inversion; instead, the generated images were re-encoded into the model's VAE latent space.

For each base garment, the stored minimum- and maximum-drape latent representations were used as the endpoints of the manipulation trajectory. Intermediate latent representations were generated by spherical linear interpolation (SLERP) between these two endpoints. Five equally spaced manipulation levels, $\alpha = -1.5, -0.75, 0.0, 0.75, \text{ and } 1.5$, were used. The α values were mapped linearly onto the interpolation interval between the minimum- and maximum-drape latent representations, with $\alpha = -1.5$ and $\alpha = 1.5$ corresponding to the two endpoints and $\alpha = 0$ representing the midpoint of the interpolation trajectory. The interpolated latent representations were decoded using the VAE to produce the final 512×512 stimulus images. The resulting stimulus set comprised 25 images (5 base garments $\times$ 5 α levels). Representative images at $\alpha = -1.5, 0.0$, and 1.5 are presented in Fig. 2.

Before the final stimulus sequences were used for perceptual evaluation, the latent configuration was subjected to an automated quality-control procedure based on our previous evaluation framework [5]. For each base garment, the difference

between the distances from the base latent representation to the minimum and maximum endpoints (Δ) and the relative centrality of the base representation (c) along the minimum-to-maximum direction were calculated. The mean values of Δ and c across the five base garments were then used for quality-control classification. A configuration was classified as PASS when $|\text{mean } \Delta| < 8.0$ and $\text{mean } c > 0.3$, and as WEAK-PASS when $|\text{mean } \Delta| < 12.0$ and $\text{mean } c > 0.15$. Configurations not satisfying these conditions were classified as FAIL. These threshold values were empirically determined through iterative testing during the development of our previous evaluation framework [5] and were retained in the present study as practical quality-control criteria rather than theoretically derived universal thresholds.

Prior to deployment, all generated stimulus images were visually inspected by the authors. No images were excluded from the final stimulus set because all generated images passed this visual quality-control procedure without requiring regeneration.

C. Human Perceptual Evaluation

Fifteen participants completed the experiment. Because this study was conducted as a preliminary evaluation of the proposed pipeline, demographic information such as age, gender, and occupation was not collected. The experiment comprised ten primary comparison trials: two α intervals (-1.5 vs. 0.0 and 0.0 vs. 1.5) for each of the five base garments. One additional replicate trial of the Dress B comparison between $\alpha = 0.0$ and $\alpha = 1.5$ was included as a limited within-participant consistency check, resulting in 11 trials per participant. This replicate was intended only to provide a basic check of repeated responses for one comparison and was not designed as a comprehensive assessment of reliability across garments or comparison difficulty. Trial order was randomized, and the left-right presentation of each image pair was randomized. The experiment therefore yielded a total of 165 paired-comparison records.

D. CLIP-Based Feature-Space Evaluation

Following our previously proposed CLIP-based evaluation framework [5], automated feature-space evaluation was performed independently for each base garment using CLIP image embeddings. The evaluation was implemented using the OpenCLIP library with OpenAI-pretrained weights. CLIP was originally pretrained on large-scale image-text pairs; however, no text prompt was directly used to define the drape-related evaluation direction in the present analysis. Instead, as described below, the direction was defined separately for each base garment from the CLIP image embeddings of its minimum- and maximum- α endpoint images.

For each base garment, five generated images corresponding to the manipulation levels $\alpha = -1.5, -0.75, 0.0, 0.75$, and 1.5 were encoded using the CLIP image encoder. Each resulting image embedding was L2-normalized before subsequent analysis.

The drape-related direction was defined separately for each base garment using the normalized CLIP image embeddings of the two endpoint stimuli. The minimum endpoint corresponds to $\alpha = -1.5$, and the maximum endpoint corresponds to $\alpha = 1.5$. Let

$v_{b,min}$ and $v_{b,max}$ denote the normalized CLIP image embeddings of these two endpoint images, respectively. The normalized drape direction d_b was defined as follows:

$$d_b = \frac{(v_{b,max} - v_{b,min})}{\|v_{b,max} - v_{b,min}\|} \quad (1)$$

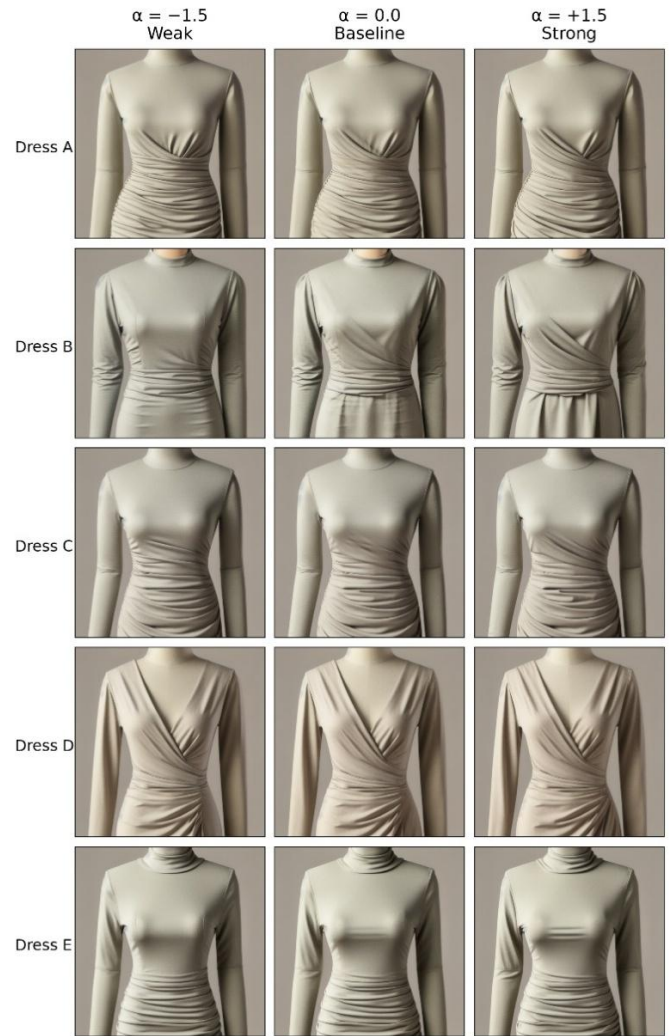


Fig. 2. Representative stimulus sequences for the five base garments (Dress A–E) at three drape manipulation levels: $\alpha = -1.5$ (weak), 0.0 (baseline), and $+1.5$ (strong).

The midpoint between the minimum and maximum endpoint embeddings was then calculated. Let $v_{b,mid}$ denote this midpoint:

$$v_{b,mid} = \frac{v_{b,max} + v_{b,min}}{2} \quad (2)$$

For each generated image, its relative position along the drape-related direction was quantified using a CLIP projection score. Let $v_{b,\alpha}$ denote the normalized CLIP image embedding of the image generated from base garment b at manipulation level α . The CLIP projection score was calculated as follows:

$$S_{CLIP(b,\alpha)} = (v_{b,\alpha} - v_{b,mid})^T d_b \quad (3)$$

The CLIP projection score represents the relative position of each generated image along the feature-space direction connecting the minimum- and maximum- α images of the same base garment. Importantly, this analysis does not assume that CLIP contains a predefined category corresponding to drape intensity. Instead, the drape-related direction is operationally defined for each base garment from the CLIP image embeddings of its two endpoint stimuli. Accordingly, the CLIP analysis should be interpreted as an internal feature-space consistency check of whether the intermediate stimuli progress systematically between the predefined endpoints, rather than as an independent semantic validation that CLIP recognizes drape intensity.

For each base garment, monotonic consistency between α and the CLIP projection score was evaluated using Spearman's rank correlation coefficient ρ . A positive value of ρ indicates that the CLIP projection score tends to increase as α increases.

In addition, trajectory straightness was calculated to quantify how directly the five image embeddings progressed from the minimum- α endpoint to the maximum- α endpoint in CLIP feature space. The straightness score was calculated as follows:

$$S_{str} = \frac{\|v_{b,max} - v_{b,min}\|}{\sum_{i=1}^{n-1} \|v_{b,i+1} - v_{b,i}\|} \quad (4)$$

A straightness score closer to 1 indicates that the sequence of image embeddings follows a more direct trajectory between the two endpoint embeddings.

E. Comparative Analysis

Human paired-comparison data were analyzed using the Bradley-Terry (BT) model to estimate perceptual worth scores β . The probability $P(i > j)$ that stimulus i (with coefficient α_i) is perceived as having stronger drape than stimulus j (with coefficient α_j) is defined as:

$$P(i > j) = \frac{\exp(\beta_i)}{\exp(\beta_i) + \exp(\beta_j)} \quad (5)$$

These perceptual worth scores were subsequently compared with CLIP projection scores S_{CLIP} to evaluate the agreement and potential divergence between human visual perception and automated feature-space evaluation across the manipulated drape levels.

IV. RESULTS

A. Stimulus Quality and CLIP-Based Feature-Space Evaluation

The latent configuration satisfied the predefined PASS criteria described in Section III-B. The measured mean Δ and mean c values were -0.791 and 0.537 , respectively, both comfortably satisfying the predefined PASS criteria ($|\text{mean } \Delta| < 8.0$ and mean $c > 0.3$).

To examine whether the aggregate PASS classification concealed weaker configurations for individual garments, the same quality-control metrics were also examined separately for each base garment. Table I summarizes the garment-specific Δ and c values. All five garments had $|\Delta| < 8.0$ and $c > 0.3$, with Δ ranging from -2.578 to 2.234 and c ranging from 0.388 to 0.596 . Thus, all individual garments satisfied the corresponding

numerical conditions used in the predefined aggregate QC criteria.

Visual inspection of the final stimulus set identified no images requiring regeneration or exclusion.

TABLE I. GARMENT-SPECIFIC LATENT-SPACE QUALITY-CONTROL METRICS

Dress	Δ	c
A	-1.803	0.596
B	-2.578	0.564
C	-0.649	0.544
D	-1.160	0.591
E	2.234	0.388
Mean	-0.791	0.537

Fig. 3 shows the CLIP projection scores along the base-specific directions defined by the endpoint image embeddings for the five base garments. For all five garments, the projection score increased monotonically as α increased from -1.5 to 1.5 . Spearman's rank correlation coefficient between α and the projection score was $\rho = 1.0$ for all five image sequences.

Thus, all five generated image sequences exhibited monotonic trajectories along their respective base-specific CLIP projection directions over the evaluated α range.

B. Human Perceptual Evaluation

The 2AFC experiment yielded 165 paired-comparison records from 15 participants. The replicated Dress B comparison between $\alpha = 0.0$ and $\alpha = 1.5$ yielded consistent responses for 14 of the 15 participants (93.3%). Because this consistency check was based on a single replicated comparison, the value should not be interpreted as an estimate of overall within-participant reliability across the full experiment.

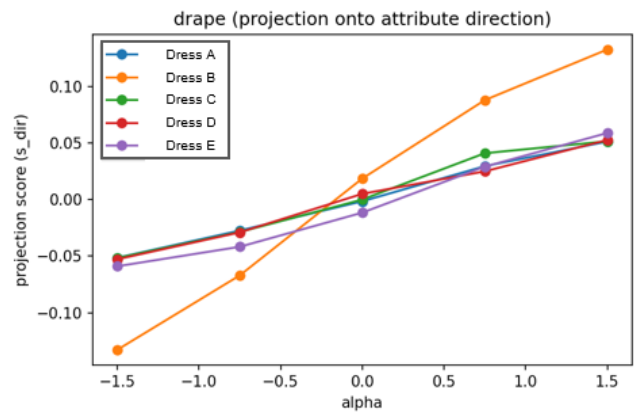


Fig. 3. CLIP-based projection scores along the drape attribute direction for five base images.

Fig. 4 shows the proportion of trials in which the higher- α stimulus was selected as having a stronger drape impression for each base garment and comparison interval. Across all base garments, the higher- α stimulus was selected in 42 of 75 comparisons (0.560) for $\alpha = -1.5$ versus 0.0 and in 77 of 90 comparisons (0.856) for $\alpha = 0.0$ versus 1.5 .

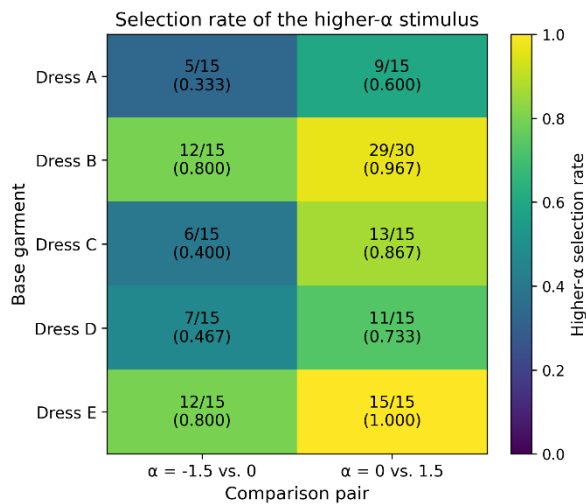


Fig. 4. Higher- α stimulus selection rates for each base garment and comparison interval in the 2AFC experiment. Each cell shows the number of higher- α selections over the number of comparisons, with the corresponding selection rate in parentheses.

The selection rates differed among the five base garments. For $\alpha = -1.5$ versus 0.0, the higher- α selection rates were 0.333, 0.800, 0.400, 0.467, and 0.800 for Dresses A–E, respectively. For $\alpha = 0.0$ versus 1.5, the corresponding rates were 0.600, 0.967, 0.867, 0.733, and 1.000. The neutral-to-high comparison for Dress B contained the additional replicate trial and therefore comprised 30 responses rather than 15.

The aggregated paired-comparison data were analyzed using the Bradley–Terry model described in Section III-E. The normalized BT worth values were 0.102, 0.130, and 0.768 for $\alpha = -1.5$, 0.0, and 1.5, respectively, showing an aggregate monotonic ordering with α .

C. Garment-Specific Bradley–Terry Analysis

To examine variation in perceptual ordering among the individual base garments, the Bradley–Terry model was fitted separately to the paired-comparison data for each garment using the same estimation procedure as in the aggregate analysis. Table II summarizes the resulting normalized BT worth scores for the three α levels evaluated in the human perceptual experiment.

TABLE II. GARMENT-SPECIFIC BRADLEY–TERRY WORTH SCORES

Dress	$\alpha = -1.5$	$\alpha = 0.0$	$\alpha = 1.5$
A	0.444	0.222	0.333
B	0.008	0.033	0.959
C	0.167	0.111	0.722
D	0.234	0.204	0.562
E	<0.001	<0.001	>0.999

For Dress E, the BT fit approached the boundary because the $\alpha = 1.5$ stimulus was selected in all comparisons against $\alpha = 0.0$; therefore, these values should be interpreted as indicating an extremely strong ordering toward $\alpha = 1.5$ rather than as stable finite estimates.

The garment-specific BT worth scores differed substantially among the five base garments. Dress B exhibited a monotonic increase in BT worth across the three evaluated α levels, with values of 0.008, 0.033, and 0.959 for $\alpha = -1.5$, 0.0, and 1.5,

respectively. Dress E also showed a strong ordering toward the highest α level, with estimated worth values below 0.001 for $\alpha = -1.5$ and 0.0 and above 0.999 for $\alpha = 1.5$. As noted in Table II, the $\alpha = 1.5$ stimulus for Dress E was selected in all comparisons against $\alpha = 0.0$, resulting in a boundary-like BT estimate.

In contrast, the estimated BT worth scores for Dresses A, C, and D showed local reversals across the three evaluated α levels. For Dress A, the normalized worth values were 0.444, 0.222, and 0.333 for $\alpha = -1.5$, 0.0, and 1.5, respectively. For Dress C, the corresponding values were 0.167, 0.111, and 0.722, and for Dress D they were 0.234, 0.204, and 0.562. Thus, $\alpha = 1.5$ had the highest estimated worth for Dresses C and D, whereas the point estimate at $\alpha = -1.5$ was slightly higher than that at $\alpha = 0.0$. Dress A showed a different estimated ordering, with the highest worth at $\alpha = -1.5$, followed by $\alpha = 1.5$ and $\alpha = 0.0$. Because these garment-specific BT worth values were estimated from a limited number of binary comparisons, the observed local reversals should be interpreted as descriptive, exploratory patterns rather than as statistically established violations of monotonicity.

V. DISCUSSION

A. Agreement and Divergence Between Automated and Human Evaluation

The results revealed both overall agreement and garment-specific divergence between the CLIP-based and human perceptual evaluations. In the automated evaluation, all five garment sequences exhibited monotonic CLIP projection scores across the five α levels, with Spearman's $\rho = 1.0$ for every base garment. Human evaluation also showed monotonic ordering at the aggregate level across the three α levels included in the human experiment, with normalized Bradley–Terry worth increasing from 0.102 at $\alpha = -1.5$ to 0.130 at $\alpha = 0.0$ and 0.768 at $\alpha = 1.5$.

However, this aggregate agreement did not extend uniformly to the individual garments. The garment-specific BT analysis showed a clear monotonic pattern for Dress B and a strong ordering toward the highest α level for Dress E, whereas the estimated BT worth scores for Dresses A, C, and D contained local reversals across the three α levels included in the human experiment. Given the limited number of binary comparisons, these reversals should be regarded as exploratory garment-specific patterns rather than statistically established non-monotonic effects. In particular, the low-to-neutral comparison was less consistent than the neutral-to-high comparison: the higher- α stimulus was selected in 56.0% of the former comparisons and 85.6% of the latter comparisons.

These findings highlight a distinction between feature-space consistency and perceptual consistency. The CLIP-based analysis characterizes how generated images progress along a direction defined by the endpoint images in the model's feature space. Because the minimum- and maximum- α images themselves define this direction, their relative ordering is inherent in the evaluation procedure, and the resulting perfect rank correlations should not be interpreted as independent evidence that CLIP recognizes drape intensity. The informative result of this analysis is instead whether the intermediate images progress systematically and monotonically between the predefined endpoints. Such a monotonic feature-space trajectory,

however, does not necessarily imply that human observers perceive the corresponding images as an equally ordered progression of drape intensity. The garment-specific human analysis was therefore important because the aggregate results alone would have suggested substantially stronger agreement between automated and human evaluation.

B. Garment-Dependent Perceptual Effects

Although latent-space manipulation was used to vary drape with each base garment serving as a common reference, the differences among Dresses A–E suggest that the perceptual effect of the manipulation may depend on the visual context of the base garment. As illustrated by the stimulus examples in Fig. 2, the same range of α values produced visually different manifestations of drape across the five garments. Dresses B and E showed human perceptual ordering broadly consistent with increasing α , whereas the estimated BT worth scores for Dresses A, C, and D showed local reversals in at least part of the evaluated range. These patterns are descriptive and should be interpreted cautiously given the limited number of comparisons in this pilot study.

The present experiment does not establish which visual characteristics caused these garment-dependent differences or whether the manipulated perceptual impression can be attributed exclusively to drape. The operational endpoint descriptions intentionally referred to visually associated characteristics such as surface smoothness, fold depth, and gathering, and the resulting manipulation may therefore have covaried with silhouette, apparent fit, fold density, garment structure, or other visual properties. Participants were asked only to select which image exhibited a stronger sense of drape and did not independently rate these potentially correlated attributes. Accordingly, the present results should be interpreted as evaluating the operational perceptual construct of drape intensity defined in this study, rather than an isolated visual dimension that has been experimentally separated from all related garment attributes. Determining which visual cues contribute independently to perceived drape will require controlled attribute-specific evaluations in future work.

Importantly, perceptually ambiguous cases should be distinguished from generation failures. During quality control, stimuli were screened for clear generation failures and structural defects, but images were not excluded merely because the perceptual ordering of drape might be difficult to judge. Such cases are informative for the present research question because the purpose of the perceptual evaluation is precisely to determine whether technically controlled image changes produce consistent human judgments. Removing perceptually ambiguous stimuli in advance could therefore conceal cases in which automated control and human perception diverge.

C. Implications for Human-in-the-Loop Evaluation

The observed discrepancy illustrates the complementary roles of automated and human evaluation in generative fashion research. Automated feature-space analysis provides an efficient means of screening large numbers of generated image sequences and identifying systematic trajectories. In the present experiment, CLIP-based evaluation indicated monotonic progression for all five garment sequences. Human evaluation,

in contrast, identified garment-specific deviations that were not apparent from the automated results alone.

This complementary relationship is central to the proposed semi-automated evaluation pipeline. The role of automation is not to replace human perceptual judgment but to reduce the effort required to generate, organize, screen, and quantitatively characterize controlled stimuli. Human evaluation can then be concentrated on determining whether technically consistent manipulations correspond to the intended perceptual effect. The pipeline therefore provides a practical mechanism for detecting cases in which model-space consistency does not translate directly into perceptual consistency.

The results also suggest a possible feedback loop for improving automated evaluation models. Garment–attribute combinations for which automated trajectories and human perceptual ordering disagree can be identified systematically and accumulated as perceptually informative cases. Such human-annotated discrepancies could potentially be used in future work to evaluate or adapt vision-language models for fashion- and textile-specific perceptual attributes. Rather than assuming that a general-purpose representation such as CLIP fully captures these attributes, the Human-in-the-Loop framework provides empirical data for determining where additional domain adaptation may be beneficial.

D. Limitations and Future Directions

Several limitations should be considered when interpreting these findings. First, this pilot study included only five AI-generated female garments and 15 participants. Because the base garments were manually selected rather than sampled according to predefined objective criteria, the present findings should not be interpreted as estimating the prevalence of such perceptual discrepancies across garment populations. The garment-specific patterns and local reversals in the estimated BT worth scores observed in the present analysis should therefore be interpreted as exploratory findings rather than as stable estimates of garment-dependent perceptual ordering. The observed garment-specific patterns cannot yet be generalized to other garment categories, designs, materials, or populations. Participant demographic information was not collected, which also prevents examination of whether perceptual judgments differed according to age, gender, fashion expertise, or other participant characteristics. In addition, the operational drape manipulation was not experimentally separated from potentially correlated visual attributes such as silhouette, apparent fit, fold density, gathering, or garment structure. Independent attribute ratings or controlled manipulations will be needed to determine the extent to which these visual factors contribute separately to perceived drape intensity.

In addition, each garment– α condition was represented by a single generated stimulus. Although the generation parameters and random seeds were recorded to ensure traceability, the present design does not distinguish systematic effects of the latent manipulation from variation associated with individual stochastic generations. The observed feature-space trajectories and perceptual judgments should therefore be interpreted as applying to the specific stimulus realizations examined in this pilot study. Future work should generate multiple independent realizations for each garment– α condition and evaluate the

robustness of both automated feature-space trajectories and human perceptual ordering across generation seeds.

The present Bradley–Terry analyses also treat the paired-comparison responses as aggregated outcomes and do not explicitly model participant-level heterogeneity or within-participant dependence across repeated judgments. Given the small number of participants and the sparse comparison structure of this pilot experiment, the garment-specific BT estimates were used primarily as descriptive summaries of perceptual ordering rather than as population-level inferential estimates. Future studies with larger samples and denser repeated-measures designs should employ hierarchical Bradley–Terry models or mixed-effects logistic regression to jointly account for participant- and garment-level variation.

Second, although five α levels were generated and evaluated using CLIP, the human 2AFC experiment evaluated only two intervals: $\alpha = -1.5$ versus 0.0 and $\alpha = 0.0$ versus 1.5 . The present results therefore do not establish the detailed functional relationship between α and perceived drape intensity. In particular, the difference in selection rates between the two evaluated intervals should not be interpreted as evidence of a general nonlinear psychophysical response. The sparse comparison design also limits the stability of garment-specific Bradley–Terry estimates, particularly in cases of highly consistent choices that produce boundary-like fits. Future experiments should include a denser set of pairwise comparisons across the manipulation range.

Third, the automated evaluation used base-specific CLIP directions defined from the minimum- and maximum- α endpoint images. This procedure is useful as an internal feature-space consistency check for characterizing the relative trajectory of each generated sequence, but it does not provide an independent semantic validation of drape intensity or establish that CLIP contains an intrinsic or general representation of this attribute. Future work should therefore evaluate intermediate stimuli using independently defined directions, such as text-derived, externally learned, or held-out directions, and investigate whether fashion- or textile-specific representations provide closer correspondence with human perceptual judgments.

Finally, the present study examined only drape intensity as an operational perceptual attribute. Extending the same experimental framework to additional garment attributes, such as material appearance, glossiness, fit, or design-related variations, would help determine whether the observed automated–human discrepancies are specific to drape or represent a more general characteristic of fine-grained attribute control in generative fashion systems. Larger and more diverse perceptual datasets would also enable systematic investigation of domain-specific adaptation or additional training of vision-language models using human perceptual annotations.

In addition, the within-participant consistency check was based on only one replicated comparison (Dress B, $\alpha = 0.0$ vs. 1.5), which showed a relatively decisive selection pattern. The observed 93.3% agreement therefore cannot be generalized to more ambiguous comparisons or interpreted as an overall measure of response reliability.

VI. CONCLUSION

This study investigated the perceptual validity of latent-space control of garment drape intensity in AI-generated garment images using a semi-automated Human-in-the-Loop evaluation pipeline. Five base garments were manipulated across five α levels and evaluated independently using CLIP-based feature-space analysis and a human 2AFC experiment with Bradley–Terry scaling.

The CLIP-based evaluation showed monotonic trajectories for all five garment sequences, whereas human evaluation revealed a more complex pattern. Although the aggregate Bradley–Terry scores increased monotonically with α , garment-specific analysis revealed local reversals in the estimated perceptual ordering for three of the five base garments. Given the limited number of binary comparisons, these garment-specific reversals should be interpreted as exploratory patterns rather than as statistically established non-monotonic effects. These findings demonstrate that aggregate agreement between automated and human evaluation can mask stimulus-specific perceptual discrepancies and that feature-space consistency does not necessarily imply perceptual consistency.

The proposed framework provides a practical means of combining automated screening with human perceptual validation for fine-grained attributes in generative fashion systems. Future work will extend the evaluation to larger and more diverse garment and participant sets, additional perceptual attributes, and denser manipulation levels. The Human-in-the-Loop data obtained through such evaluations may also support the development or adaptation of automated models that better reflect fashion- and textile-specific human perception.

DECLARATION ON GENERATIVE AI

Generative AI tools, including ChatGPT (OpenAI) and Gemini (Google), were used during manuscript preparation to assist with language refinement, organization, and improvement of clarity. All AI-assisted content was critically reviewed, revised, and verified by the authors. The scientific ideas, experimental design, data analysis, interpretation of the results, and conclusions presented in this study were developed and validated by the human authors, who take full responsibility for the content of the manuscript.

REFERENCES

- [1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.
- [2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10684–10695, 2022.
- [3] W. Shi, W. Wong, and X. Zou, “Generative AI in Fashion: Overview,” *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 4, Art. no. 74, 2025, doi: 10.1145/3718098.
- [4] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, “Human Preference Score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis,” *arXiv preprint arXiv:2306.09341*, 2023, doi: 10.48550/arXiv.2306.09341.
- [5] N. Kuwahara, S. Kawanami, T. Sato, and D. Choi, “Evaluating perceptual reliability of latent attribute control in diffusion-based fashion generation,” *International Journal of Advanced Computer Science and*

- Applications, vol. 17, no. 4, pp. 231–240, 2026, doi: 10.14569/IJACSA.2026.0170424.
- [6] C. C. Chu, C. L. Cummings, and N. A. Teixeira, “Mechanics of elastic performance of textile materials: Part V: A study of the factors affecting the drape of fabrics—The development of a drape meter,” *Textile Research Journal*, vol. 20, no. 8, pp. 539–548, 1950, doi: 10.1177/004051755002000802.
- [7] M. Matsudaira and M. Yang, “Features of conventional static and new dynamic drape coefficients of woven silk fabrics,” *Textile Research Journal*, vol. 73, no. 3, pp. 250–255, 2003, doi: 10.1177/004051750307300309.
- [8] I. Uchiyama, T. Asai, and H. Shigeno, “Quantitative consideration on the beauty of drape,” *J. Japan Research Association for Textile End-Uses*, vol. 15, no. 5, pp. 170–174, 1974, doi: 10.11419/senshoshi1960.15.170. (in Japanese)
- [9] K. Tsuji, K. Ito, and N. Kato, “Effect of fabric properties on gathered skirts and their visual evaluation,” *Journal of Home Economics of Japan*, vol. 32, no. 6, pp. 463–471, 1981, doi: 10.11428/jhej1951.32.463. (in Japanese)
- [10] N. Suda and M. Ohira, “On the drapeability of fabrics: Visual evaluation of drapeability,” *J. Japan Research Association for Textile End-Uses*, vol. 14, no. 4, pp. 122–126, 1973, doi: 10.11419/senshoshi1960.14.122. (in Japanese)
- [11] N. Suda and M. Ohira, “On the drapeability of fabrics: Analysis of drape formation mechanism using a flared-skirt-shaped model,” *J. Japan Research Association for Textile End-Uses*, vol. 20, no. 4, pp. 148–152, 1979, doi: 10.11419/senshoshi1960.20.148. (in Japanese)
- [12] M. Miki, M. Ayada, and M. Niwa, “Effects of the mechanical properties of fabrics and the number of gore panels have on the silhouettes of flared skirts,” *Journal of Home Economics of Japan*, vol. 46, no. 7, pp. 671–682, 1995, doi: 10.11428/jhej1987.46.671. (in Japanese)
- [13] K. Izumi and M. Niwa, “Relation between the silhouette of gathered skirts in wear and bending, vibration properties of fabrics,” *Journal of Home Economics of Japan*, vol. 34, no. 2, pp. 96–104, 1983, doi: 10.11428/jhej1951.34.96. (in Japanese)
- [14] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3836–3845, 2023.
- [15] T. Brooks, A. Holynski, and A. A. Efros, “InstructPix2Pix: Learning to follow image editing instructions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.
- [16] Y. Shen, G. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of GANs for semantic face editing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9240–9249, 2020.
- [17] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “GANSpace: Discovering interpretable GAN controls,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9841–9850, 2020.
- [18] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, “StyleCLIP: Text-driven manipulation of StyleGAN imagery,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 2085–2094, 2021, doi: 10.1109/ICCV48922.2021.00210.
- [19] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 8748–8763, 2021.
- [20] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, F. Tang, and Y. Dong, “ImageReward: Learning and evaluating human preferences for text-to-image generation,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.
- [21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 586–595, 2018.
- [22] M. Jogan and A. A. Stocker, “A new two-alternative forced choice method,” *Journal of Vision*, vol. 14, no. 3, 2014.
- [23] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. The method of paired comparisons,” *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, 1952.