Evaluating Perceptual Reliability of Latent Attribute Control in Diffusion-Based Fashion Generation

Noriaki Kuwahara[1], Shintaro Kawanami[2], Takashi Sato[3], Dongeun Choi[4]*
Kyoto Institute of Technology, Kyoto, Japan[1, 2, 4]
Tokyo Denki University, Tokyo, Japan[3]

***Abstract*—Although diffusion-based image generation models enable high-quality synthesis of fashion images, the reliable control of perceptual attributes in these models remains poorly understood. Current evaluation approaches primarily rely on semantic similarity metrics, such as CLIP scores, which may not accurately reflect human perceptual judgments. This study proposes a three-layer evaluation framework linking latent space geometry, semantic embedding space, and human perception. First, latent attribute directions are validated using geometric quality-control metrics measuring linearity and centrality. Second, semantic consistency is examined through directional projection in CLIP embedding space. Third, a two-alternative forced-choice experiment is conducted with 37 participants, and perceptual strength is estimated using a Bradley-Terry preference model. Experiments cover gender and garment conditions for four fashion attributes: fit, lightness, glossiness, and pattern scale. Results reveal that fit exhibits strong cross-layer alignment, while pattern scale shows semantic and perceptual ambiguity. The findings highlight that perceptual reliability in controllable generation is attribute-dependent and that semantic metrics alone cannot fully replace human evaluation.**

***Keywords*—Diffusion models; controllable generation; latent space analysis; human preference modeling; perceptual reliability; fashion image generation**

I. Introduction

Diffusion-based image generation models demonstrated remarkable visual quality and are widely used in text-to-image synthesis [1], [2]. In fashion applications, these models enable controllable manipulation of garment appearance through latent-space operations. However, reliable control of perceptual attributes such as fit, lightness, glossiness, and pattern scale has not been sufficiently validated.

Current evaluation methods primarily rely on semantic similarity metrics, including CLIP-based scores or reward models trained on human preferences [3], [4]. These approaches measure alignment between generated images and textual prompts, but do not explicitly assess whether latent attribute manipulations are geometrically stable or perceptually consistent. As a result, latent representation behavior, semantic embedding responses, and human perceptual discernment are often treated interchangeably.

Previous studies show that semantic attributes may correspond to specific directional changes in latent space [5], [6]. Despite these findings, the geometric validity of these directions and their perceptual reliability within diffusion models have not been systematically analyzed. Furthermore, while human preference modeling is increasingly employed in generative AI evaluation [4], [7], its relationship to latent space geometry and semantic embedding behavior is still unclear.

To address this gap, we propose a three-layer evaluation framework that links latent space geometry, semantic embedding space, and human perception. Our method employs a latent direction quality-control mechanism to validate the linearity and centrality of attribute directions prior to semantic and human evaluation. Experiments under gender and garment conditions measured performance across four perceptual attributes. Cross-layer consistency analysis shows that the perceptual reliability of controllable diffusion generation is attribute-dependent and cannot be fully predicted by semantic similarity metrics alone.

The contributions of this study are:

- A three-layer evaluation framework separating latent geometry, semantic consistency, and human perceptual reliability.
- A latent direction quality control approach for validating geometric stability.
- An empirical analysis revealing attribute-dependent alignment across evaluation layers.

II. Related Work

A. Controllable Diffusion Models

Diffusion-based generative models are widely known for producing high-fidelity images [1], [2]. Their iterative denoising framework supports stable training dynamics and allows for flexible conditioning. While many diffusion models focus on text-conditioned generation, some have introduced additional control signals into their architectures to guide structural or semantic properties. For example, ControlNet introduced extended diffusion models with conditional control branches to improve structural consistency under user constraints [8]. More recent work has further explored efficient and scalable virtual try-on systems using knowledge distillation techniques, improving both realism and computational efficiency [9]. In fashion-related applications, diffusion-based virtual try-on systems have demonstrated realistic garment transfer while preserving the human form. Most of these systems focus on achieving realistic results and maintaining structural coherence. However, limited attention has been paid to systematically evaluating the controllability of perceptual attributes.

*Corresponding author

Although diffusion models allow manipulation of latent spaces, the geometric properties of attribute directions in diffusion latent spaces have not been extensively examined. Unlike deterministic GAN-based architectures, diffusion models involve stochastic denoising processes. This presents challenges in maintaining stable latent attribute directions. As a result, controllability depends on explicit validation of the directions of geometric attributes rather than on assuming their stability.

B. Latent Space Geometry and Directional Editing

Latent space interpretability has been widely studied in GAN frameworks. StyleGAN demonstrated that semantically meaningful transformations can emerge along structured latent representations [10]. However, prior work has also challenged common assumptions in disentangled representation learning, indicating that such latent structures may not be as robust or universally interpretable as often assumed [11]. InterFaceGAN further showed that certain attributes correspond to approximately linear directions in latent space, enabling semantic editing through vector operations [5]. Subsequent studies emphasized the importance of geometric alignment and stability in defining controllable directions [6], [12].

Representation learning theory suggests that meaningful variation predominantly lies in structured subspaces rather than arbitrary directions [13]. While most analyses focus on GAN architectures that employ deterministic latent-variable mappings, diffusion models differ significantly due to their latent characteristics arising from stochastic noise injection and iterative refinement. Consequently, the geometric validity, centrality, and stability of attribute directions in diffusion latent space are underexplored and require dedicated evaluation mechanisms.

C. Semantic Embedding-Based Evaluation

Semantic alignment between generated images and corresponding textual descriptions is commonly evaluated using CLIP-based similarity metrics [3]. Embedding-based reward models, such as ImageReward, are designed to approximate human preference through learned scoring mechanisms [4]. These evaluation approaches provide scalable, automatic assessment and are now standard benchmarks for text-to-image generation.

However, distance-based similarity metrics do not explicitly separate intended attribute variation from unrelated visual changes. Existing literature on evaluation has shown that global error or similarity measures may not capture meaningful structural features of human perception [14]. Recent studies on human preference modeling in generative systems further emphasize that embedding-based metrics only partially correlate with human discretion [15]. For this reason, direction-sensitive analysis in embedding space is needed to verify that attribute progression is consistently preserved, not just semantic similarity increased.

D. Human Preference and Perceptual Reliability

Human evaluation remains fundamental to assessing the inherently subjective aspects of generative models. Human preference modeling systems typically employ experimental methods, such as two-alternative forced-choice (2AFC) and pairwise comparison frameworks, to statistically quantify human subjective discernment [7]. The resulting comparative data is then analyzed using the Bradley-Terry model [16] to construct latent preference scales that reflect human preferences. These modeling processes not only play a central role in evaluation but also serve as a key component for aligning reward systems within modern generative AI pipelines [4], [15].

Despite the growing adoption of human preference modeling techniques, most existing studies evaluate either automatic semantic alignment or human preference independently. They have not yet examined the relationship among latent geometry, semantic embedding behavior, and perceptual reliability collectively. A consolidated framework integrating geometric validation, semantic direction analysis, and human perceptual assessment is necessary to facilitate structured interpretation and understanding of controlled generation aligned with human factors.

Recent studies have increasingly emphasized the importance of controllability and systematic evaluation in generative AI systems [18], [19], highlighting the limitations of relying solely on semantic similarity metrics. These trends further motivate the need for a unified framework that systematically links latent space manipulation, semantic embedding behavior, and human perception.

III. Proposed Framework

A. Problem Formulation

Recent diffusion-based image generation models provide continuous manipulation of visual attributes in the latent space. However, controllable latent directions do not guarantee perceptual reliability.

In this study, we distinguish between three conceptually distinct layers:

- Latent geometry: the internal representation of the generative model.
- Semantic embedding space: a machine-interpretable meaning structure (e.g., CLIP).
- Human perception: subjective discernment of attributes.

Controllable generation implicitly assumes that a given perceptual attribute varies and can be approximated along a one-dimensional axis in the latent space. However, this geometric construction implicitly assumes that:

- Attribute variation is locally linear in latent space.
- The base image lies near the midpoint of the attribute axis.
- Progression along the axis corresponds to monotonic perceptual change.

These assumptions carry significant and non-trivial implications.

From a geometric standpoint, high-dimensional generative latent spaces do not ensure perfectly aligned semantic subspaces.

From a psychophysical standpoint, perceived attribute intensity is generally non-linear with respect to changes in

physical stimulus, as suggested by classical laws such as Weber–Fechner.

Perceptual reliability is, therefore, defined as cross-layer monotonicity and not just latent controllability. This approach ensures the ordering induced by the generation parameter α is preserved in both semantic and human perceptual evaluations.

More formally, let:

- α denote the generation parameter along the latent direction,
- $s(\alpha)$ denotes the semantic projection score,
- $\theta(\alpha)$ denotes the perceptual strength estimated from human evaluation.

An attribute manipulation is considered perceptually reliable if:

- The latent configuration satisfies geometric consistency,
- $s(\alpha)$ is monotonic with respect to α,
- $\theta(\alpha)$ is monotonic with respect to α,
- The ordering induced by $s(\alpha)$ and $\theta(\alpha)$ is positively correlated.

Thus, the central research question is: What geometric conditions enable latent attribute direction to preserve monotonicity through semantic embedding and human perception?

This formulation transforms attribute control from a generation problem into a cross-layer consistency problem.

B. Latent Attribute Direction Hypothesis

Let $z_0 \in \mathbb{R}^d$ denote the latent representation of a base image generated by the diffusion model.

For a given perceptual attribute a, we obtain two endpoint latent codes which are intended to represent the minimal and maximal perceptual strength of attribute a.

$$z_{a,\min}, z_{a,\max}$$

We define the latent attribute direction as Eq. (1):

$$\mathbf{d}_a = z_{a,\max} - z_{a,\min}. \quad (1)$$

To ensure symmetry and interpretability, we adopt a midpoint-centered and normalized parameterization.

The midpoint is defined as Eq. (2):

$$z_{a,mid} = \frac{z_{a,\max} + z_{a,\min}}{2} \quad (2)$$

The normalized direction is Eq. (3):

$$\hat{\mathbf{d}}_a = \frac{\mathbf{d}_a}{\| \mathbf{d}_a \|}. \quad (3)$$

The latent attribute axis is then expressed as Eq. (4):

$$z(\alpha) = z_{a,mid} + \alpha \frac{\| \mathbf{d}_a \|}{2} \hat{\mathbf{d}}_a, \qquad \alpha \in \mathbb{R} \quad (4)$$

whereby,

$$\alpha = \pm 1$$

correspond to the endpoints of the minimal and maximal attributes.

This formulation separates two properties:

- Axis validity — whether $z_{a,\min}$ and $z_{a,\max}$ define a coherent one-dimensional direction.
- Base centrality — whether z_0 lies near $z_{a,\text{mid}}$.

We therefore formulate the following hypothesis:

Hypothesis H_1:

The triplet $(z_{a,min}, z_0, z_{a,max})$ is approximately collinear in the ambient latent space $\mathbb{R}^d$, and displacement along the normalized direction $\hat{\boldsymbol{d}}_a$ preserves perceptual ordering.

If H_1 holds, varying α should induce a systematic perceptual change in attribute a without introducing unintended semantic drift. Because diffusion latent spaces are high-dimensional and stochastic, H_1 is not guaranteed a priori and must be empirically validated.

In the next section, we introduce geometric criteria for testing H_1.

C. Geometric Quality Control in Latent Space

1) Linearity deviation (Δ): To evaluate hypothesis H_1, we first introduce a geometric measure that quantifies deviation from ideal midpoint symmetry.

The midpoint of the attribute axis is defined by Eq. (2). Assuming ideal symmetry and perfect collinearity, the base latent representation coincides with the midpoint.

Equivalently,

$$z_{a,\max} + z_{a,\min} - 2z_0 = 0.$$

We, therefore, define the linearity deviation as Eq. (5):

$$\Delta = \| z_{a,\max} + z_{a,\min} - 2z_0 \| \quad (5)$$

Geometric interpretation:

- $\Delta = 0$ indicates perfect midpoint symmetry.
- Larger Δ indicates deviation from collinearity.
- Non-zero Δimplies geometric distortion of the attribute axis relative to the base configuration.

Notably, Δ evaluates structural consistency exclusively in latent space, independent of semantic or perceptual interpretation.

2) Centrality (c): While Δ measures absolute deviation, it depends on the scale of the attribute span. To obtain a scale-invariant quantity, we define centrality as Eq. (6):

$$c = 1 - \frac{\| z_{a,\max} + z_{a,\min} - 2z_0 \|}{\| z_{a,\max} - z_{a,\min} \|} \quad (6)$$

Properties:

$c \approx 1 \rightarrow$ base is near the midpoint.

$c \approx 0 \rightarrow$ base is near one endpoint.

$c < 0 \rightarrow$ base lies outside the attribute span.

Thus, centrality quantifies the extent to which the base image is centered relative to the attribute axis.

From a geometric perspective, c quantifies the symmetry of displacement. In psychophysical terms, poor centrality may induce asymmetric perceptual sensitivity along the attribute axis.

3) QC decision principle: Geometric validation must take place before semantic or perceptual evaluation.

We define a quality control (QC) mechanism based on Δ and c. An attribute direction is classified as:

- PASS if Δ is sufficiently small and c is sufficiently high.
- WEAK-PASS if only one condition is moderately satisfied.
- FAIL if both conditions indicate significant geometric distortion.

The specific numerical thresholds are determined empirically and reported in the experimental section.

Conceptually, QC establishes a necessary geometric condition for ensuring perceptual reliability:

- If geometric symmetry fails, cross-layer monotonicity cannot be interpreted meaningfully.
- If QC is satisfied, subsequent semantic and perceptual evaluations assess whether geometric consistency propagates across layers.

Thus, geometric QC constitutes the structural foundation of the proposed three-layer evaluation framework.

D. Semantic Direction Validation (clip)

Section III-C establishes geometric consistency in latent space. To ensure perceptual reliability, the attribute progression should be preserved in the semantic embedding space.

We therefore analyze whether the latent attribute axis defined in Section III-B remains consistently oriented in the CLIP embedding space.

Let Eq. (7):

$$\mathbf{v}_i = f_{\text{img}}(x_i) \in \mathbb{R}^D \tag{7}$$

denote the CLIP image embedding of the generated image x_i, where D denotes the dimensionality of the CLIP embedding space.

We define the semantic attribute direction using endpoint embeddings [see Eq. (8)]:

$$\mathbf{d}_{\text{clip}} = \frac{\mathbf{v}_{\max} - \mathbf{v}_{\min}}{\| \mathbf{v}_{\max} - \mathbf{v}_{\min} \|} \tag{8}$$

where, $\mathbf{v}_{\min}$ and $\mathbf{v}_{\max}$ correspond to images generated at $\alpha = -1$ and $\alpha = +1$, respectively.

To ensure symmetry, we define the semantic midpoint as Eq. (9):

$$\mathbf{v}_{\text{mid}} = \frac{\mathbf{v}_{\max} + \mathbf{v}_{\min}}{2} \tag{9}$$

The directional projection score is then defined as Eq. (10):

$$s_i = (\mathbf{v}_i - \mathbf{v}_{\text{mid}}) \cdot \mathbf{d}_{\text{clip}} \tag{10}$$

This projection extracts only the component aligned with the semantic attribute direction.

1) Why projection instead of distance?

Conventional CLIP evaluation typically relies on global similarity or cosine distance measures. However, distance-based metrics do not distinguish between intended attribute variation and unrelated semantic drift.

Directional projection isolates attribute-consistent change while suppressing orthogonal variation.

2) Monotonic consistency in semantic space: To evaluate whether semantic progression aligns with latent parameter α, we compute the Spearman Rank Correlation [see Eq. (11)]:

$$\rho_{\text{clip}} = \text{Spearman}(\{\alpha_i\}, \{s_i\}) \tag{11}$$

Interpretation:

- $\rho_{\text{clip}} \approx 1 \rightarrow$ monotonic semantic progression.
- $\rho_{\text{clip}} \approx 0 \rightarrow$ semantic incoherence.
- $\rho_{\text{clip}} < 0 \rightarrow$ semantic inversion.

Thus, semantic validation tests whether geometric consistency propagates into the embedding space.

3) Duality with latent QC: The proposed semantic validation mirrors latent qc. The human-layer monotonicity measure is formally defined in Section III-E: ρ_{human}

TABLE I. DUAL STRUCTURAL CRITERIA ACROSS LAYERS

Layer	Structural Criterion
Latent Space	Midpoint symmetry (Δ, c)
Semantic Space	Directional monotonicity (ρ_{clip})
Human Space	Perceptual monotonicity (ρ_{human})

Table I summarizes the structural criteria used at each evaluation layer, where each layer evaluates structural consistency with respect to a distinct representational geometry. This framework enforces geometric validity in latent space, semantic monotonicity in embedding space, and perceptual monotonicity in human evaluation.

E. Human Perceptual Validation

While Section III-C and Section III-D establish geometric and semantic consistency, perceptual reliability ultimately depends on human discernment.

To address this, we introduce a psychophysically grounded evaluation based on pairwise comparison.

1) Two-Alternative Forced Choice (2AFC): For each attribute, we compare generated images along the latent axis using the two-alternative forced choice (2AFC) Protocol.az(α_i)

Given a pair of images (x_i, x_j), participants are asked:

- Which image exhibits a stronger perceptual intensity of attribute a?

2AFC is preferred over absolute rating scales because:

- It reduces inter-observer bias.
- It avoids arbitrary internal calibration.
- It yields statistically consistent pairwise comparison data.

The outcome of the comparison shall be denoted as:

$$x_i \succ x_j$$

if image x_i is judged stronger than x_j.

2) Bradley–Terry preference model: We employ the Bradley–Terry Model to derive a continuous perceptual scale based on pairwise outcomes. Each image is associated with a latent perceptual strength parameter. The probability that is preferred over is defined as: $x_i\theta_i x_i x_j$ [see Eq. (12)]

$$P(x_i \succ x_j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)} \quad (12)$$

The parameters θ_i are estimated using maximum likelihood. This model yields a continuous perceptual strength scale:

$$\theta(\alpha_i)$$

associated with each generation parameter α_i.

3) Monotonic consistency in human space: To evaluate Perceptual Monotonicity, we compute the Spearman Rank Correlation between latent parameter and perceptual strength: α [see Eq. (13)]

$$\rho_{\text{human}} = \text{Spearman}(\{\alpha_i\},\{\theta_i\}) \quad (13)$$

Interpretation:

- $\rho_{\text{human}} \approx 1 \rightarrow$ monotonic perceptual progression.
- $\rho_{\text{human}} \approx 0 \rightarrow$ perceptual incoherence.
- $\rho_{\text{human}} < 0 \rightarrow$ perceptual inversion.

Human perceptual validation tests whether geometric and semantic consistency are reflected in subjective discernment. This completes the tri-layer validation structure of the proposed framework.

F. Cross-Layer Consistency Principle

The preceding sections introduce structural validation criteria across three representational layers:

- Latent geometry (Δ, c),
- Semantic embedding space ($\rho_{\text{clip}} > 0$),
- Human perception ($\rho_{\text{human}} > 0$).

These sections consolidate the principles integrating the stated criteria.

1) Necessary and progressive conditions: Geometric validity in latent space is a necessary prerequisite for ensuring perceptual reliability. When midpoint symmetry is violated (indicated by a Large Δ or Low C), the attribute axis becomes structurally distorted. In such cases, semantic or perceptual monotonicity loses meaningful interpretability.

Semantic consistency constitutes an important second-level condition. Even if latent geometry is coherent, the semantic embedding space must maintain directional monotonicity:

$$\rho_{\text{clip}} > 0.$$

Finally, perceptual reliability requires monotonic progression in human discernment:

$$\rho_{\text{human}} > 0.$$

2) Formal definition: We define the perceptual reliability of an attribute as the conjunction of cross-layer monotonicity conditions: a [see Eq. (14)]

$$\text{Reliable}(a) \iff \begin{cases} \Delta \text{ is small}, \\ c \text{ is near } 1, \\ \rho_{\text{clip}} > 0, \\ \rho_{\text{human}} > 0. \end{cases} \quad (14)$$

In practice, threshold values were empirically determined through preliminary validation and are specified in Section IV-B. We determine the thresholds based on preliminary validation of ranges that must be exceeded for geometric distortion to occur.

3) Interpretation: The proposed framework does not assume that controllable generation is universally reliable. Instead, it provides a structured mechanism for diagnosing where reliability fails:

- Latent distortion (geometric failure),
- Semantic misalignment (embedding failure),
- Perceptual ambiguity (human inconsistency).

Thus, perceptual reliability is treated as an emergent property of cross-layer structural consistency.

4) Theoretical implication: The three-layer framework suggests that controllable diffusion generation is not solely determined by latent manipulability. Instead, attribute reliability depends on whether structural coherence propagates across representational geometries:

$$\mathbb{R}^d \rightarrow \mathbb{R}^D \rightarrow \text{Human perceptual space.}$$

This perspective shifts evaluation from measuring similarity to analyzing structural consistency.

IV. Experimental Setup

A. Diffusion Model Configuration

To apply the proposed three-layer framework, we generated images using a Stable Diffusion 1.5–based inpainting model

(Realistic Vision V5.1). We employed an inpainting configuration to enable controlled attribute manipulation while preserving subject identity and overall scene structure.

Unlike full image synthesis, the inpainting setting allows the torso, pose, and background to remain fixed while restricting modifications to the garment region. This configuration isolates attribute variation from confounding factors such as posture, lighting, or composition changes. Accordingly, perceptual variations caused by the latent parameter α are primarily due to garment-related transformations.

Let $z_0 \in \mathbb{R}^d$ denote the latent representation of a base image under a fixed subject condition. For each attribute a, we generated minimal and maximal endpoint latent codes $z_{a,\min}$ and $z_{a,\max}$ using prompt-controlled inpainting. These endpoints define the latent attribute direction introduced in Section III-B.

All latent manipulations occurred in the VAE latent space of the underlying Stable Diffusion architecture. The latent dimensionality is denoted by d, and the CLIP embedding dimensionality is denoted by D, consistent with the definitions in Section III-D.

After geometric validation (Section III-C), we generated α-series images using spherical linear interpolation (SLERP) between endpoints. SLERP preserves latent distribution geometry by interpolating along a hyper-spherical path instead of a straight Euclidean line. This reduces distributional drift during interpolation [17]. To improve perceptual discriminability, we distributed α values using logarithmic scaling and extended their range:

$$\alpha \in [-1.5, 1.5].$$

This extension preserves geometric consistency while increasing perceptual separability in attributes with weaker visual gradients.

Notably, generation and evaluation were structurally separated. The diffusion model serves solely as a transformation mechanism in latent space. We subsequently performed structural validation across three independent representational layers:

- Latent space (geometric QC: Δ, c)
- Semantic embedding space (projection-based monotonicity: ρ_{clip})
- Human perception (2AFC + Bradley–Terry: ρ_{human})

As illustrated in Fig. 1, perceptual reliability is not treated as a property of a single representational space.

Instead, the generation parameter α is propagated across three structurally distinct layers: latent space geometry, semantic embedding space, and human perceptual discernment. Monotonic consistency with respect to α is evaluated independently at each layer. We, therefore, define perceptual reliability as the preservation of cross-layer order rather than isolated similarity within any single space.

Fig. 2 shows the representative base images for each gender–garment condition. In all conditions, the torso, pose, and background remain constant, with only the garment region undergoing subsequent attribute manipulation.

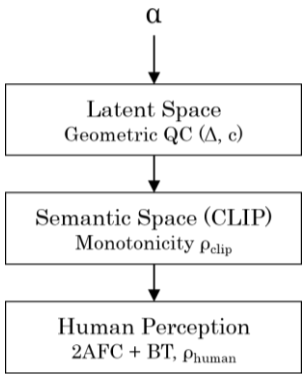

Fig. 1. Three-layer evaluation framework for perceptual reliability.

Fig. 2. Representative base images under fixed gender–garment conditions before attribute manipulation.

B. Latent Direction Construction and Geometric Quality Control

For each perceptual attribute a, latent endpoint representations were constructed under constant subject and garment conditions using the inpainting configuration described in Section IV-A. Using controlled prompt variations, we generated minimal and maximal attribute expressions that yielded the latent codes $z_{a,\min}$ and $z_{a,\max}$ in the VAE latent space of the underlying Stable Diffusion architecture.

We constructed the latent attribute direction $\mathbf{d}_a$ as defined in Eq. (1). This endpoint-based formulation derives the semantic axis in embedding space directly from the generative transformation itself, not from external text embeddings. As a result, the semantic direction remains structurally aligned with the latent manipulation.

Because diffusion latent spaces are high-dimensional and stochastic, simply constructing endpoint representations do not ensure a coherent one-dimensional attribute axis. To ensure structural validity prior to semantic or human evaluation, we employed a geometric validation. For each base image z_0, the

linearity deviation Δ, and centrality $c$were computed as defined in Eq. (5) and Eq. (6). For each attribute under each experimental condition, average values $\bar{\Delta}$ and $\bar{c}$ were obtained across base samples.

We defined the quality control criteria as:

- PASS: $|\bar{\Delta}| < 8.0$ and $\bar{c} > 0.30$
- WEAK-PASS: $|\bar{\Delta}| < 16.0$ and $\bar{c} > 0.15$
- FAIL: otherwise

We excluded attribute directions marked as FAIL and retained WEAK-PASS directions only when visual inspection confirmed they were perceptually interpretable with no visually observable structural distortion.

QC, thereby functions as a structural precondition for cross-layer analysis. If midpoint symmetry is substantially violated, monotonic consistency in semantic embedding space or human perceptual space cannot be meaningfully interpreted. We only propagated attribute directions that satisfied QC to the α-series generation and subsequent cross-layer evaluation.

Fig. 3 shows the representative endpoint.

Fig. 3. Representative minimal (min) and maximal (max) endpoint images for selected attributes.

The upper row corresponds to the fit attribute in the male T-shirt condition, illustrating a geometrically stable latent direction that creates a smooth variation in garment tightness. The lower row shows the pattern scale in the female dress-shirt condition, representing a geometrically weaker (WEAK-PASS) configuration with structurally complex visual variation. These examples offer visual context for the geometric differences evaluated before cross-layer analysis.

C. Semantic Projection Evaluation

To evaluate the preservation of latent attribute progression in the semantic embedding space, we analyzed generated images using a CLIP image encoder.

For each generated image x_i along the α-series, we computed the corresponding image embedding $\mathbf{v}_i \in \mathbb{R}^D$ defined in Eq. (7). Consistent with the endpoint-based formulation introduced in Section III-D, we then constructed the semantic attribute direction using the endpoint embeddings defined in Eq. (8). The semantic midpoint and directional projection scores were computed following Eq. (9) and Eq. (10), respectively.

This projection-based evaluation isolates variation aligned with the semantic attribute direction while suppressing orthogonal changes in embedding. Unlike global similarity or cosine distance measures, directional projection explicitly assesses if attribute progression remains a consistent semantic axis.

We evaluated monotonic consistency in semantic space, using the Spearman rank correlation defined in Eq. (11). A positive correlation indicates that semantic embedding responses preserve the ordering induced by latent manipulation, while near-zero or negative values indicate semantic incoherence or inversion.

Importantly, we treated semantic validation as an intermediate structural layer, not as a substitute for human evaluation. This verifies whether geometric consistency in the latent space propagates to the embedding space before assessing perceptual monotonicity.

D. Human Perceptual Evaluation

To evaluate how latent attribute manipulation reflects human perception, we conducted a two-alternative forced-choice (2AFC) experiment.

We recruited forty participants assigned to experimental conditions defined by gender and garment type. After excluding three responses due to incomplete submissions, 37 valid responses remained for analysis. For each perceptual attribute, participants viewed pairwise comparisons of images generated along the α-series and selected the image with the stronger perceptual intensity.

We chose pairwise comparison over absolute rating scales to reduce inter-observer calibration bias and improve statistical consistency. The 2AFC protocol produces ordinal preference data suitable for probabilistic modeling of latent perceptual strength.

We then estimated perceptual strength parameters θ_i using the Bradley–Terry model defined in Eq. (12). The resulting preference scale provides a continuous representation of perceived attribute intensity for each generation parameter α_i.

To assess perceptual monotonicity, we computed Spearman's rank correlation coefficient between the generation parameter α and the estimated perceptual strength, as defined in Eq. (13). A positive correlation coefficient indicates monotonic perceptual progression, whereas near-zero or negative values indicate perceptual incoherence or inversion.

We randomized image order and left–right presentation to mitigate positional bias. Human evaluation constitutes the final validation layer in the proposed framework. Section V analyzes these results to assess cross-layer alignment. The framework is

designed as a diagnostic tool to identify structurally unreliable attribute manipulations before deployment.

V. Results

Table II summarizes cross-layer monotonicity results aggregated by attribute. All attributes satisfied geometric admissibility, with fit, lightness, and glossiness classified as PASS, and pattern scale categorized as WEAK-PASS according to the criteria specified in Section IV-B.

Semantic monotonicity was consistently positive across all attributes, with high values of ρ_{clip} indicating stable directional progression in embedding space. These results confirm that latent manipulations were structurally preserved at the semantic level.

TABLE II. Cross-Layer Monotonicity Results Aggregated by Attribute.

Attribute	QC Status	ρ_{clip}	ρ_{human}	Cross-layer Alignment
Fit	PASS	0.97	0.67	Strong
Lightness	PASS	0.99	0.31	Moderate
Glossiness	PASS	1.00	0.48	Moderate
Pattern Scale	WEAK-PASS	1.00	0.51	Limited

Perceptual monotonicity, evaluated with 37 valid participant responses, exhibited attribute-dependent variation. Fit demonstrated the strongest perceptual monotonicity. Glossiness exhibited moderate perceptual monotonicity, whereas lightness was comparatively weaker. This suggests that although luminance variation is consistently represented in embedding space, its perception is affected by contextual factors such as shading and local contrast. Pattern scale exhibited limited perceptual reliability despite positive semantic monotonicity.

For fit, latent manipulation primarily affected the garment's contour and silhouette, producing visibly coherent adaptive fitting—tightening and loosening—that appeared across geometric, semantic, and perceptual layers. In contrast, the pattern scale involved changes in motif size together with variations in texture density and contrast. As illustrated in Fig. 3, pattern enlargement introduced composite visual effects beyond simple geometric scaling. This factor presumably reduced perceptual monotonicity despite preserved semantic ordering.

All reported correlations were statistically significant at $p < 0.01$.

Joint evaluation of geometric validity, semantic embedding behavior, and human evaluation shows that perceptual reliability varies across attributes. Fit satisfies all cross-layer conditions defined in Eq. (14), demonstrating strong alignment across representational layers. In contrast, pattern scale demonstrates geometric admissibility but reduced perceptual monotonicity, indicating partial cross-layer misalignment.

These findings support the formulation that perceptual reliability cannot be determined solely from geometric or semantic validity, but arises only when monotonic ordering is preserved across latent geometry, embedding space, and human perception.

VI. Discussion

In this study, we introduced a three-layer evaluation framework to assess perceptual reliability in controllable diffusion-based fashion image generation. By explicitly separating latent geometry, semantic embedding space, and human perception, our proposed framework enables structured diagnosis of attribute controllability across representational levels.

A. Significance of the Three-Layer Decomposition

The central contribution of this study lies in the structural separation of representational layers. Previous evaluations of controllable generation typically rely on either automatic semantic metrics or human preference modeling independently. However, as demonstrated in the present results, geometric validity in latent space does not automatically imply perceptual reliability.

Our proposed framework demonstrates that perceptual reliability is achieved only when the monotonic order induced by the generation parameter α is preserved across all layers. This condition enables identification of alignment failure in geometric construction, semantic embedding behavior, or human perceptual interpretation, thereby providing diagnostic capabilities essential to understanding controllable diffusion models beyond visual realism.

B. Attribute-Dependent Reliability

The results reveal that perceptual reliability is attribute-dependent. Fit exhibited strong cross-layer alignment, satisfying geometric, semantic, and perceptual monotonicity conditions. This suggests that attributes closely tied to structural garment draping are more consistently encoded across representational spaces.

In contrast, pattern scale demonstrated reduced perceptual monotonicity despite geometric admissibility and positive semantic monotonicity. This indicates that certain attributes may be geometrically constructible yet perceptually ambiguous. Such ambiguity may stem from the composite nature of pattern scale, which includes density, contrast, and texture complexity rather than a single structural factor.

Glossiness exhibited moderate behavior, while lightness showed comparatively weaker perceptual monotonicity despite satisfying geometric and semantic criteria. These findings highlight that perceptual interpretation may depend on contextual or experiential factors even when semantic embeddings capture directional consistency.

C. Implications for Automatic Evaluation Metrics

The results further demonstrate that embedding-based monotonicity cannot replace human evaluation. While ρ_{clip} remained positive for all attributes, perceptual monotonicity varied significantly. This partial decoupling suggests that the semantic embedding space captures structural ordering but does not fully encode perceptual salience.

Consequently, evaluation pipelines that rely solely on CLIP similarity or reward models may overestimate controllability when perceptual ambiguity exists. Our proposed cross-layer

framework provides a robust and effective approach to identifying such discrepancies.

D. The Role of Geometric Quality Control

Geometric quality control (QC) plays a foundational role in the framework. By validating midpoint symmetry and collinearity before semantic and perceptual evaluation, QC establishes a critical structural condition for meaningful monotonic analysis.

However, the WEAK-PASS cases indicate that geometric symmetry alone does not determine perceptual reliability. Instead, QC acts as a structural filter that ensures interpretability of subsequent layers, while final reliability depends on cross-layer alignment.

E. Limitations and Future Work

Several research limitations should be noted. First, our study employed only one diffusion backbone and inpainting configuration. Although the proposed framework is model-agnostic, validating it empirically across additional architectures could improve generalizability.

Second, we relied solely on visual discernment for perceptual evaluation. Attributes such as glossiness and fit may also depend on tactile or experiential cues not captured in image-based evaluation. Extending the framework to multimodal perceptual assessment is therefore an important consideration for future research.

Third, we did not explicitly analyze cultural and experiential variability among participants. We recognize that future studies should examine whether perceptual monotonicity differs by demographic or expertise groups.

Finally, several methodological limitations should be acknowledged. The selection of endpoint latent codes $z_{a,min}$ and $z_{a,max}$ was based on controlled generation and visual inspection, and was not formulated as a systematic or optimization-based procedure, which may affect reproducibility. In addition, although the human evaluation employed a two-alternative forced-choice design with 37 participants, a formal statistical power analysis was not conducted, and potential order effects or response dependencies were not explicitly modeled. Furthermore, the study focused on four representative attributes, which may introduce selection bias. Addressing these limitations through principled endpoint selection, more rigorous experimental design, and broader attribute coverage remains an important direction for future work.

F. Broader Implications

The proposed framework could also apply to other fields beyond fashion-image generation, whereby controllable generative models are assessed against human interpretation. Separating geometric validity, semantic embedding behavior, and perceptual monotonicity provides a generalizable framework for assessing human-aligned controllability.

Our study moves from assessing methods that rely solely on isolated similarity to a broader cross-layer monotonicity approach, providing a structured framework for understanding how generative transformations propagate across representational geometries.

VII. Conclusion

In this study, we proposed a three-layer evaluation framework for assessing perceptual reliability in controllable diffusion-based fashion image generation. Explicitly separating latent geometry, semantic embedding space, and human perception allowed for structured validation of attribute manipulation across representational levels.

We introduced geometric quality control to ensure midpoint symmetry and collinearity in the latent space prior to semantic or perceptual evaluation, and evaluated semantic consistency using projection-based monotonicity in the CLIP embedding space. We subsequently assessed perceptual monotonicity using a 2AFC experiment with 37 valid participants, applying the Bradley–Terry model.

Empirical results show that perceptual reliability depends on the attribute being tested: Fit exhibited strong cross-layer alignment, whereas pattern scale showed limited perceptual monotonicity despite geometric admissibility. These findings confirm that controllable generation cannot be evaluated solely through semantic similarity or latent manipulability. Instead, perceptual reliability arises only when a monotonic ordering is preserved across latent geometry, embedding space, and human judgment.

Our proposed framework provides a structured methodology for diagnosing controllable generation beyond visual realism and offers a foundation for human-aligned evaluation of generative models.

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 et al., "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 10684–10695, 2022.

[3] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), pp. 8748–8763, 2021.

[4] J. Xu et al., "ImageReward: Learning and evaluating human preferences for text-to-image generation," arXiv preprint arXiv:2304.05977, 2023.

[5] 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.

[6] 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.

[7] 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.

[8] H. Wang et al., "MV-VTON: Multi-view virtual try-on with diffusion models," arXiv preprint arXiv:2404.17364, 2024.

[9] Z. Zhang, L. Wang, and W. Ding, "LiteMP-VTON: A knowledge-distilled diffusion model for realistic and efficient virtual try-on," Information, vol. 16, no. 5, p. 408, 2025.

[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4401–4410, 2019.

[11] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in Proc. Int. Conf. Mach. Learn. (ICML), pp. 4114–4124, 2019.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, 2013.

[13] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it?," IEEE Signal Process. Mag., vol. 26, no. 1, pp. 98–117, 2009.

[14] 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.

[15] M. Jogan and A. A. Stocker, "A new two-alternative forced choice method," J. Vis., vol. 14, no. 3, 2014.

[16] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs," Biometrika, vol. 39, no. 3–4, pp. 324–345, 1952.

[17] T. White, "Sampling generative networks," arXiv preprint arXiv:1609.04468, 2016.

[18] 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.

[19] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv:2207.12598, 2022.