Trait Collapse in Text-to-Image Generation: Evaluating and Mitigating Personality Inconsistency in Visual Storytelling

Anonymous ACL submission

Abstract

In visual storytelling, a believable character must not only look consistent across scenes but also *behave* in ways that reflect their personality. Current text-to-image (T2I) models preserve visual appearance effectively, yet consistently fail to maintain personality-relevant behavior—a failure we term *trait collapse*. We show that conventional metrics (CLIP Score, DINO similarity) are blind to this failure, and that abstract personality descriptors yield no reliable improvement over neutral prompts regardless of abstraction level. To systematically diagnose trait collapse, we introduce **ViStoryBench-Personality**, a story-level benchmark combining an expert-validated Visual Cues Lexicon (Fleiss' $\kappa = 0.757$) with a six-dimensional metric suite covering personality alignment, cross-scene stability, and behavioral coherence. Evaluating seven T2I systems, we confirm a persistent semantic gap between abstract trait labels and visual evidence, and find that visual identity consistency does not imply personality consistency. To bridge this gap, we propose **Trait-Behavior-Visual (TBV)**, a training-free prompting framework that decomposes personality traits into renderable visual constraints. TBV improves Personality Alignment Score by up to $+22.1\%$, with larger gains in models with stronger language-conditioning capacity.

1 Introduction

In visual storytelling, character consistency extends beyond appearance. Personality psychology treats traits as stable behavioral dispositions rather than transient emotional states (McCrae and John, 1992), and observers readily infer personality from visible cues such as posture, gaze, and facial expression (Naumann et al., 2009). A convincing generated character should therefore not only remain visually recognizable across scenes, but also sustain behavior that is consistent with their underlying personality profile—a property we term *visual personality consistency*.

Current T2I models largely fail to maintain this property. Existing personalization methods (Ruiz et al., 2022; Gal et al., 2022; Ye et al., 2023; Li et al., 2023) and story-level approaches (Zhou et al., 2024) address the question of what a character looks like, with little attention to how a character acts. A model may faithfully preserve facial identity across frames while repeatedly producing generic smiles, rigid postures, or scene-driven emotional reactions, collapsing personality expression into a uniform behavioral default. We term this failure *trait collapse*, and attribute it to a *semantic grounding* gap: models receive abstract descriptors such as "friendly" or "reserved" as input, but cannot translate them into observable, context-stable visual behavior. This is distinct from general prompt non-compliance, as the failure is localized to personality-specific semantics rather than scene understanding. Conventional metrics offer no signal here—CLIP score does not improve, and sometimes decreases, under personality-grounded prompting, while DINO similarity remains largely unchanged ($|\Delta| < 0.03$; Table 1)—and existing story visualization benchmarks (Zhuang et al., 2025) similarly do not isolate personality-relevant behavior as an evaluation target.

To address this gap, we introduce **ViStoryBench-Personality**. The benchmark represents character personality through OCEAN profiles (McCrae and John, 1992), connects traits to observable visual manifestations via an expert-validated Visual Cues Lexicon ($\kappa = 0.757$), and quantifies performance along six dimensions covering personality alignment, cross-scene stability, visual readability, consistency, drift, and narrative coherence. Our evaluation shows that abstract personality descriptors fail to reliably improve over the neutral baseline across all tested architectures ($C_1 \approx C_2 \approx C_3$), confirming that

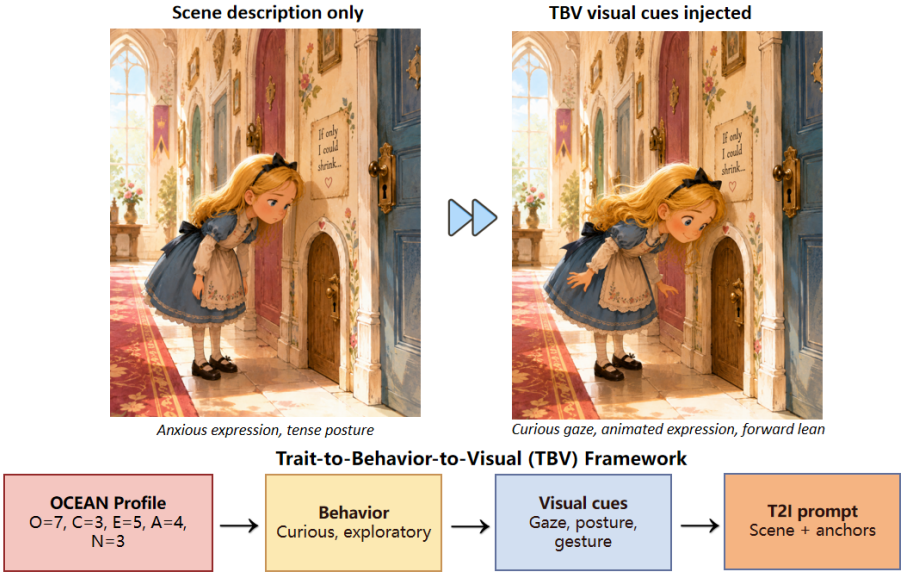

Figure 1: **Trait collapse vs. trait expression.** Given the same narrative scene, baseline prompts (C1, left) produce generic or emotionally dominated behavior, while TBV prompts (C4, right) translate the OCEAN profile into concrete visual constraints, maintaining personality-consistent expression. Bottom: the TBV framework decomposes abstract personality traits into behavioral tendencies and then into renderable visual cues via an expert-validated lexicon.

trait collapse is a systematic rather than incidental failure.

To mitigate trait collapse without retraining, we propose the **Trait-Behavior-Visual (TBV)** framework. TBV decomposes personality grounding into three sequential stages: anchoring the character's OCEAN profile, deriving context-sensitive behavioral tendencies, and mapping those tendencies to concrete visual constraints drawn from the lexicon. TBV improves Personality Alignment Score by up to $+22.1\%$, with greater gains observed in models with stronger text-encoding capacity.

2 Related Work

2.1 Character Consistency in Story Visualization

Character preservation is a central requirement for visual narrative generation. T2I personalization methods preserve subject identity through test-time adaptation (Ruiz et al., 2022; Gal et al., 2022) or zero-shot visual adapters (Ye et al., 2023; Li et al., 2023). Story-level methods extend this objective to multi-shot narratives: StoryDiffusion promotes cross-frame coherence through consistent self-attention (Zhou et al., 2024), while recent multi-character systems introduce region-level control for identity stability (Banerjee et al., 2025; Cheng et al., 2025). However, these methods primarily address *who* the character looks like, not *how* the character behaves. A character may remain visually recognizable while becoming behaviorally generic, with personality-relevant posture, gaze, and interaction patterns overridden by scene-level emotion. This gap is consistent with findings that pretrained models often rely on broad social stereotypes rather than fine-grained trait reasoning in personality-related tasks (Hausladen et al., 2024; Hamidieh et al., 2024; Yang et al., 2026).

2.2 Controllable Action and Emotion Generation

Controllable generation methods improve control over character pose, layout, motion, and affective expression, using spatial conditions such as edges, depth, segmentation maps, or human poses (Zhang et al., 2023), decoupled text-to-pose or motion-to-video pipelines (Guo et al., 2023; Wang et al., 2025), or emotion-conditioned signals such as valence-arousal values (He et al., 2025). However, these methods mainly specify what a character is doing or what emotion is shown in a particular scene; they do not directly evaluate whether the same character maintains stable personality expression across changing contexts. A model may render the requested action or emotion correctly while still producing generic, scene-driven, or personality-inconsistent behavior. We therefore view trait col-

lapse as a previously under-specified failure mode: prior work has improved identity preservation, action controllability, and affective expression, but rarely examines whether characters continue to behave like the same psychological individual across scenes.

2.3 Evaluation Benchmarks for T2I Generation

T2I evaluation has progressed from single-image prompt adherence (Saharia et al., 2022; Huang et al., 2023) to compositional reasoning (Chen et al., 2025; Deng et al., 2025) and story-level coherence (Zhuang et al., 2025). Existing story-level benchmarks assess narrative content, appearance consistency, and cross-frame identity, but do not isolate whether the same character remains psychologically coherent across scenes. Likewise, standard metrics such as CLIP-based text–image similarity and DINO-based appearance similarity can measure prompt alignment or visual identity, but provide no direct evidence that traits such as cautiousness, sociability, agreeableness, or emotional stability are expressed through visible behavior. This motivates a dedicated evaluation protocol for visual personality consistency.

2.4 Personality Theory and Visual Behavior

We adopt the Big Five (OCEAN) model (McCrae and John, 1992), which represents stable individual differences along Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This trait-level representation is useful because it separates enduring dispositions from transient emotions, a distinction that current T2I models often blur. The Brunswik Lens Model (Brunswik, 1957) further explains how observers infer psychological attributes from observable cues: for example, extraversion is associated with open posture, active approach, and direct gaze; agreeableness with softened expression and reduced interpersonal distance; and neuroticism with tense posture, avoidant gaze, and self-protective gestures. This cue-based view provides the theoretical basis for our benchmark, which evaluates personality through observable visual evidence, and for TBV, which inverts this process by translating personality profiles into concrete visual generation constraints.

3 Problem Formulation

Let c denote a character with a predefined OCEAN personality profile $\mathbf{v} \in [1, 7]^5$, and let $\{I_1, \ldots, I_T\}$ denote an image sequence generated for c across T narrative scenes.

Visual personality consistency. A generated sequence exhibits *visual personality consistency* if: (i) each image I_t contains observable personality cues—posture, gaze, facial expression, and interpersonal distance—that align directionally with $\mathbf{v}$; (ii) this alignment remains stable across scenes without systematic drift; and (iii) the alignment is legible from visual cues alone, without access to personality labels. This does not require the same emotion in every scene: a character may react differently across contexts while still retaining personality-consistent cues such as gaze, posture, or interaction style. Section 4.4 operationalizes these requirements as automatic metrics.

Trait collapse. *Trait collapse* occurs when visual identity consistency is maintained but visual personality consistency fails: personality-relevant cues are degraded, suppressed, or replaced by generic scene-driven reactions. It is distinct from identity drift, where appearance changes, and from general prompt non-compliance, where scene understanding fails broadly; the failure is localized to the abstract-to-visual mapping of personality traits. Section 6 provides empirical evidence for this mechanism.

4 ViStoryBench-Personality

We introduce **ViStoryBench-Personality**, a diagnostic benchmark for evaluating visual personality consistency in sequential T2I generation. Built on ViStoryBench (Zhuang et al., 2025), it spans 80 stories and 1,317 sequential shots, and provides four components: structured OCEAN profiles, an expert-validated Visual Cues Lexicon, controlled prompt conditions, and a six-dimensional metric suite.

4.1 Dataset Construction

Each story features a focal character represented by an OCEAN vector $\mathbf{v} \in [1, 7]^5$. From the full corpus, we sample 213 evaluation shots balanced across five affective scene categories derived from Russell's Circumplex Model (Russell, 1980): Neutral (20.7%), Low-arousal positive (19.2%), Low-arousal negative (19.7%), High-arousal positive (18.8%), and High-arousal negative (21.6%). Each shot is also labeled as *Direct* (71.8%) or *Indirect* (28.2%) depending on whether personality cues are

expressed through explicit facial/gestural signals or more implicit proxemic and spatial cues. Full dataset statistics are provided in Appendix F.

Ablation conditions. Each evaluation shot is generated under four prompt conditions: **C1** uses only the narrative scene and character appearance; **C2** adds abstract OCEAN trait labels; **C3** further adds a deduced scene-level emotion label; and **C4** injects explicit TBV-derived visual constraints. C1–C3 test whether non-visual textual personality information can reduce trait collapse, while C4 tests whether concrete visual grounding is necessary. A prompt example is provided in Appendix G.

4.2 Visual Cues Lexicon

The Visual Cues Lexicon $\mathcal{L}_{\text{visual}}$ maps OCEAN dimensions to 150 observable cues across 10 trait poles and five signal types: facial expression, gaze behavior, body posture, clothing/color palette, and scene environment. Cues are assigned to three activation tiers: **Core** (≥ 5 or ≤ 3), **Extended** (≥ 6 or ≤ 2), and **Fine-grained** ($= 7$ or $= 1$), allowing visual constraint density to scale with trait extremity while avoiding over-specification for moderate profiles. The full lexicon is provided in Appendix E.

4.3 Expert Validation

We validate the Visual Cues Lexicon with 18 psychology-trained raters. Experts completed three tasks: OCEAN-pole mapping, trait–behavior concordance rating, and end-to-end scenario coherence judgment. The lexicon achieves substantial agreement in 10-way dimension mapping (Fleiss' $\kappa = 0.757$), high trait–behavior concordance (S-CVI/Ave $= 0.978$), and strong scenario-level coherence (S-CVI/Ave $= 0.975$). Full task formats, per-pole statistics, and effect-size analyses are provided in Appendix C.

4.4 Evaluation Metrics

We organize evaluation into two complementary levels: *snapshot-level* metrics assess personality expression within a single image, while *temporal-level* metrics assess consistency across a narrative sequence. All metrics use GPT-4o as the VLM judge with structured JSON outputs; the exact judge prompts and decoding parameters are provided in Appendix B. We validate this judge against human annotations (Spearman $\rho = 0.611$, polarity agreement $= 86.6\%$; Appendix I).

4.4.1 Snapshot-Level Metrics

Personality Alignment Score (PAS). PAS measures how faithfully a generated image reflects the intended personality. Given the preset OCEAN profile $\mathbf{v}$ and the VLM-inferred profile $\hat{\mathbf{v}}$, we compute centered cosine similarity as PAS $= \frac{(\hat{\mathbf{v}}-\mathbf{m})\cdot(\mathbf{v}-\mathbf{m})}{\|\hat{\mathbf{v}}-\mathbf{m}\|_2 \cdot \|\mathbf{v}-\mathbf{m}\|_2}$, where $\mathbf{m} = [4,4,4,4,4]$ is the neutral midpoint. Centering ensures that alignment reflects directional personality evidence rather than neutral-scale bias.

Personality Consistency Score (PCS). PCS measures stochastic stability under repeated generation. For the same prompt, we generate $N = 3$ images and compute PCS $= 1 - \frac{1}{5}\sum_{d\in\mathcal{D}} \frac{\mathrm{Var}(\{\hat{v}_d^{(i)}\}_{i=1}^N)}{\sigma^2_{\max}}$, where $\sigma^2_{\max} = (U-L)^2/4 = 9$ by Popoviciu's inequality. High PCS indicates stable rendering, but not necessarily correct personality alignment, so we interpret PCS jointly with PAS.

Readability Score (RS). RS evaluates whether personality is inferable from visual cues alone. The VLM receives only the image and plot context, with all personality descriptors removed, and infers $\tilde{\mathbf{v}}$. RS uses the same centered cosine formulation as PAS between $\tilde{\mathbf{v}}$ and the target profile $\mathbf{v}$. High RS indicates that the generated image itself carries visually readable personality evidence.

4.4.2 Temporal-Level Metrics

Character Consistency Score (CCS). CCS measures cross-situational stability of personality expression. For shots $\{I_1, \ldots, I_n\}$ depicting the same character, CCS is the mean pairwise centered cosine similarity, CCS $= \frac{1}{\binom{n}{2}}\sum_{i<j} \frac{(\hat{\mathbf{v}}_i-\mathbf{m})\cdot(\hat{\mathbf{v}}_j-\mathbf{m})}{\|\hat{\mathbf{v}}_i-\mathbf{m}\|_2 \cdot \|\hat{\mathbf{v}}_j-\mathbf{m}\|_2}$. High CCS indicates internally stable expression, but does not guarantee correctness; we therefore interpret it together with PAS and PDS.

Personality Drift Score (PDS). PDS quantifies whether personality expression systematically deviates over a narrative sequence. For each shot t, we compute the normalized distance $d_t = \frac{1}{5}\sum_{d\in\mathcal{D}} \frac{|\hat{v}_d^{(t)} - v_d|}{U-L}$ and fit $d_t = \beta_0 + \beta_1 \tilde{t}$ over normalized time $\tilde{t} = (t-1)/(T-1)$. PDS is the slope β_1: a positive value indicates progressive personality erosion, while a value near zero suggests no systematic drift.

Story Coherence Score (SCS). SCS evaluates whether the generated image aligns with both the preset personality and the behavioral logic implied by the narrative: $\text{SCS} = \alpha \cdot \text{PAS} + (1 - \alpha) \cdot \text{SCS}_{\text{reasoning}}$, where $\text{SCS}_{\text{reasoning}} \in [0, 1]$ is a VLM-assigned score measuring how well the image matches the personality-conditioned behavioral trace from Phase B. We set $\alpha = 0.4$ to emphasize context-specific behavioral coherence while retaining global trait alignment; a sensitivity analysis (Appendix D) confirms that model rankings are stable across $\alpha \in [0.2, 0.8]$.

5 Trait-Behavior-Visual (TBV) Framework

The ablation design of ViStoryBench-Personality (Section 4.1) isolates whether abstract personality descriptors can reduce trait collapse by comparing conditions C1–C3. As our experiments confirm (Section 6), these conditions yield no reliable improvement, pointing to a systematic semantic gap between personality labels and visual rendering. To mitigate trait collapse without model retraining, we propose the **Trait-Behavior-Visual (TBV)** framework, a training-free prompting intervention that explicitly decomposes this gap into three tractable phases. As illustrated in Figure 2, TBV translates abstract OCEAN profiles into concrete visual generation constraints through: Phase T (Trait Anchoring), Phase B (Behavioral Deduction), and Phase V (Vision Mapping). Unless otherwise stated, all LLM-based operations in TBV, including trait anchoring, behavioral deduction, reverse inference, and visual-cue selection, are implemented with GPT-4o using structured JSON outputs. Implementation settings are provided in Appendix H.

5.1 Phase T: Trait Anchoring

Phase T establishes a stable OCEAN personality anchor for each character. To mitigate the inherent variance of LLM-based trait extraction, we employ a multi-sample self-consistency strategy (Wang et al., 2022): we sample k independent inferences and measure their agreement using a Normalized Disagreement Index, defined as $\text{NDI} = \frac{1}{5k} \sum_{i=1}^{k} \sum_{d \in \mathcal{D}} \frac{|v_{i,d} - \bar{v}_d|}{U - L}$, where $\mathcal{D} = \{O, C, E, A, N\}$, $U = 7$, $L = 1$, and $\bar{v}_d$ is the mean inferred score for dimension d across the k samples. A three-tier routing mechanism determines the outcome: if $\text{NDI} < 0.15$, the mean vector $\bar{\mathbf{v}}$ is accepted as the personality anchor $\hat{\mathbf{v}}$; if $0.15 \leq \text{NDI} \leq 0.33$, a multi-agent debate is triggered to reconcile divergent estimates; otherwise, the system falls back to a population-norm baseline. This mechanism reduces the propagation of unstable personality anchors to subsequent phases.

5.2 Phase B: Behavioral Deduction

Given the anchored personality profile $\hat{\mathbf{v}}$ and a narrative scene description, Phase B generates a behavioral reasoning trace r describing how the character would act in this context. The key challenge is ensuring that the character's response reflects their stable personality rather than defaulting to a generic reaction driven by the scene's emotional content.

To address this, we introduce a self-correcting loop inspired by iterative refinement methods (Madaan et al., 2023). An independent critic agent performs reverse inference on the reasoning trace to reconstruct the implied personality profile $\tilde{\mathbf{v}} = \text{Inf}(r)$. We then compute a behavioral deviation score $s = \frac{1}{5} \sum_{d \in \mathcal{D}} \frac{|\hat{v}_d - \tilde{v}_d|}{U - L}$. If $s > \tau$ (we set $\tau = 0.15$, equivalent to a mean deviation of 0.9 points per dimension on the 7-point scale), the trace is rejected as out-of-character and regenerated with corrective feedback. This mechanism ensures that the behavioral description remains personality-consistent even under emotionally charged narrative contexts.

5.3 Phase V: Vision Mapping

Phase V translates behavioral descriptions into explicit visual constraints. Following an inverse application of the Brunswik Lens Model (Brunswik, 1957)—mapping known personality traits to the observable cues through which they become visible—we draw from the expert-validated Visual Cues Lexicon $\mathcal{L}_{\text{visual}}$ (Section 4.2). For each character, the lexicon selects cues matching the character's OCEAN profile at the appropriate tier level (Core, Extended, or Fine-grained).

To reduce semantic collisions (e.g., simultaneous "clenched jaw" from high Neuroticism and "gentle smile" from high Agreeableness), we apply a priority-based suppression rule: facial expression and gaze cues take precedence over posture cues, which in turn override clothing and scene-level signals. This hierarchy reflects evidence that facial behavior is the most salient channel for personality perception (Ryumina et al., 2026). The resulting conflict-reduced constraint set $\mathcal{C}_{\text{visual}}$ is injected directly into the T2I prompt as the C4 condition in our ablation design.

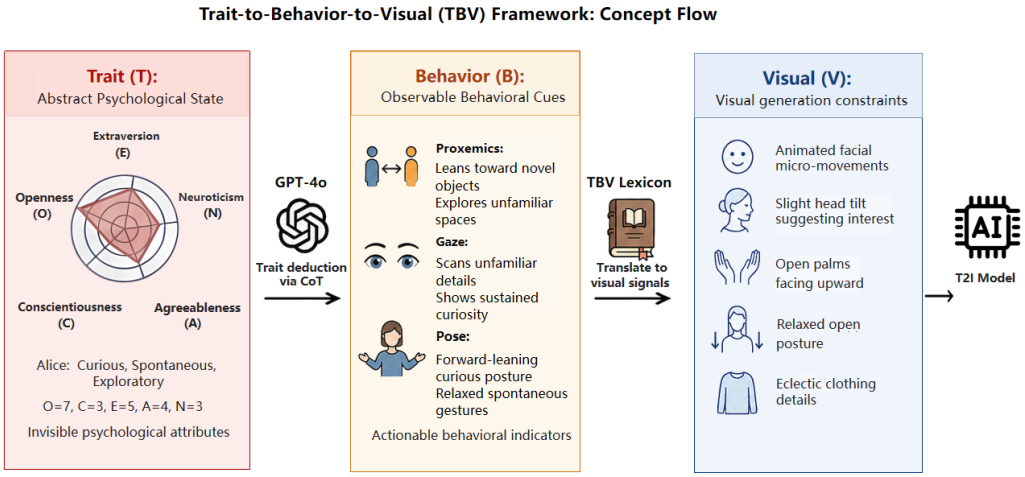

Figure 2: Overview of the Trait-Behavior-Visual (TBV) framework. Phase T anchors the character's OCEAN personality profile; Phase B derives context-sensitive behavioral tendencies via a self-correcting loop; Phase V maps behaviors into concrete visual constraints through the expert-validated Visual Cues Lexicon.

6 Experiments

6.1 Experimental Setup

Evaluated models. We evaluate seven T2I systems spanning three groups: (1) open-source generative models—**SDXL** (Podell et al., 2023), **FLUX.1-dev** (Labs, 2024), **SD3.5-Large** (Stability AI, 2024), and **Qwen-Image-2512**—covering representative UNet-, DiT-, MMDiT-, and LLM-conditioned architectures; (2) a visual consistency method, **StoryDiffusion** (Zhou et al., 2024), which builds on SDXL with Consistent Self-Attention; and (3) two closed-source APIs, **GPT Image 2** and **Gemini 3.1 Flash Image Preview (Nano Banana)**. Open-source models are evaluated in full diagnostic mode (C1–C4); closed-source APIs in benchmarking mode (C1 vs C4 only) due to internal prompt rewriting.

VLM judge. We employ GPT-4o with structured JSON output as the automated evaluator, validated against human annotations (Spearman $\rho = 0.611$, polarity agreement $= 86.6\%$; Appendix I). Detailed generation and evaluation settings are provided in Appendix H.

6.2 Limitations of Conventional Metrics

Before presenting our benchmark results, we first demonstrate that conventional T2I metrics are blind to personality dimensions. Table 1 reports CLIP Score (text–image alignment) and DINO similarity (cross-frame appearance consistency) under baseline (C1) and TBV (C4) conditions across all seven models.

	CLIP Score ↑			DINO Sim. ↑		
Model	C1	C4	Δ	C1	C4	Δ
SDXL	.274	.262	−.012	.443	.475	+.032
FLUX.1-dev	.256	.187	−.069	.501	.504	+.003
SD3.5-Large	.260	.252	−.008	.467	.497	+.030
StoryDiff.	.259	.256	−.003	.378	.392	+.014
Qwen-2512	.263	.259	−.004	.644	.631	−.013
GPT Image 2	.279	.254	−.025	.496	.504	+.008
Nano Banana	.276	.259	−.017	.487	.509	+.022
Mean	.267	.247	−.020	.488	.502	+.014

Table 1: CLIP Score and DINO similarity under C1 (baseline) and C4 (TBV). CLIP Score *decreases* under TBV prompting due to token truncation; DINO similarity remains unchanged ($|\Delta| < 0.03$). Neither metric captures the personality alignment improvements detected by our benchmark metrics (Table 2).

CLIP Score shows no meaningful improvement under personality-grounded prompting—and in several cases decreases, most notably on FLUX.1-dev ($\Delta = -.069$), where TBV produces the largest PAS gains. DINO similarity remains effectively unchanged across conditions (mean $|\Delta| = .014$). These results confirm that existing evaluation infrastructure cannot distinguish between personality-grounded and personality-absent generations, motivating the dedicated metrics introduced in Section 4.4.

6.3 Main Results

Table 2 reports evaluation results across six metrics. Qualitative case studies in Appendix N further illustrate this pattern using the same representative story shot across models and prompting conditions

Model / Setting	PAS ↑	PCS ↑	RS ↑	CCS ↑	PDS ↓	SCS ↑
Open-source models: baseline						
SDXL	0.446	0.686	0.429	0.357	+0.001	0.585
StoryDiffusion	0.480	0.763	0.467	0.499	+0.000	0.587
FLUX.1-dev	0.464	0.731	0.469	0.385	+0.003	0.614
SD3.5-Large	0.466	0.743	0.493	0.478	+0.003	0.605
Qwen-Image-2512	0.518	0.790	0.546	0.456	+0.001	0.665
Open-source models: +TBV						
SDXL + TBV	**0.467**	0.655	0.423	**0.406**	**-0.002**	0.560
StoryDiffusion + TBV	**0.502**	0.685	**0.508**	**0.527**	**-0.001**	0.564
FLUX.1-dev + TBV	**0.522**	**0.734**	**0.552**	**0.511**	**+0.002**	**0.642**
SD3.5-Large + TBV	**0.472**	**0.748**	0.488	0.449	+0.003	0.603
Qwen-Image-2512 + TBV	**0.633**	**0.816**	**0.641**	**0.558**	**-0.000**	**0.701**
Closed-source models: baseline						
GPT Image 2	0.538	0.793	0.556	0.504	-0.002	0.702
Nano Banana	0.542	0.754	0.551	0.459	+0.002	0.707
Closed-source models: +TBV						
GPT Image 2 + TBV	**0.648**	**0.815**	**0.659**	**0.606**	**-0.003**	**0.740**
Nano Banana + TBV	**0.638**	0.748	**0.661**	**0.505**	+0.002	**0.730**

Table 2: Personality consistency evaluation across seven T2I models. TBV yields up to $+22.1\%$ relative PAS improvement on Qwen-Image-2512, with larger gains observed in models equipped with stronger text encoders.

(Figures 4 and 5). Baseline models show consistently low personality alignment, with PAS ranging from 0.446 (SDXL) to 0.542 (Nano Banana). Applying TBV increases PAS numerically across all seven systems, with statistically significant gains concentrated in models with stronger text encoders, as further analyzed in Section 6.4. Qwen-Image-2512 (LLM-conditioned) and GPT Image 2 both exceed +20% relative PAS improvement, while SDXL (CLIP-only encoder) shows the smallest gain.

Notably, StoryDiffusion achieves the strongest baseline PCS (0.763) but only moderate PAS (0.480), confirming that visual identity consistency does not imply psychological consistency. After TBV, its PAS improves to 0.502, but the gain (+4.6%) is smaller than models with stronger language understanding, suggesting that personality grounding is bottlenecked by the text encoder rather than the visual consistency mechanism.

A related pattern emerges in PCS: TBV reduces PCS in three of five open-source systems (SDXL: $0.686 \rightarrow 0.655$; StoryDiffusion: $0.763 \rightarrow 0.685$; Nano Banana: $0.754 \rightarrow 0.748$). We interpret this as a diagnostic signal rather than a failure: baseline PCS is partly inflated by generic behavioral defaults, which are trivially consistent across seeds. TBV disrupts this default mode and introduces controlled personality-specific variation, which registers as lower stochastic consistency but higher alignment. PCS and PAS should therefore be read jointly; a high-PCS/low-PAS profile indicates stable but personality-absent generation, whereas a moderate-PCS/high-PAS profile indicates personality-expressive but naturally variable generation.

6.4 Ablation Study

We further examine whether trait collapse can be mitigated by adding non-visual textual information alone. Across the five open-source models, abstract OCEAN descriptors (C2) and scene-level emotion labels (C3) produce inconsistent effects: FLUX.1-dev slightly improves under C2 (+0.019), whereas SDXL decreases by 0.054. Paired Wilcoxon tests and TOST equivalence tests further support this pattern: 9 of 10 C1 vs. C2/C3 comparisons show no significant PAS improvement, and 6 of 10 fall within a practically negligible margin ($\delta = 0.05$; Appendix K). In contrast, C4 yields significant PAS gains for models with stronger text encoders, including FLUX.1-dev, Qwen-Image-2512, GPT Image 2, and Nano Banana. These results indicate that concrete visual grounding, rather than richer abstract labels, is the key factor for mitigating trait collapse, with efficacy modulated by text-encoding capacity. Full ablation results are provided in Appendix J.

6.5 Emotion Override Effect

We further examine whether affective scene context interferes with personality expression. Across valid generated images, PAS differs significantly across affective scene categories (ANOVA: $F(4, 5076) = 20.06$, $p < .001$; Kruskal–Wallis: $H = 82.00$, $p < .001$; Appendix L). Neutral scenes obtain the highest mean PAS, while all emotionally salient categories show lower PAS, suggesting that current T2I models tend to prioritize transient scene emotion over stable personality cues when the two signals compete. For example, a highly open character may be rendered primarily as anxious or withdrawn in a tense scene, even when the intended personality profile implies exploratory gaze or forward-leaning engagement. This finding supports our central claim that trait collapse is not merely an identity-preservation problem, but a semantic grounding problem: personality must remain visually legible even when emotional context changes. We also test whether personality alignment differs between direct and indirect visual-expression shots, but find no significant gap; this null result is reported in Appendix M.

7 Discussion and Conclusion

This paper identifies and formalizes *trait collapse*, a failure mode in which T2I models preserve visual identity while losing personality-relevant behavior. Through ViStoryBench-Personality, we show that this failure is systematic across seven T2I systems: abstract OCEAN and emotion labels do not reliably improve personality alignment, while TBV yields numerical PAS gains across all systems, with significant gains concentrated in stronger language-conditioned models. These results suggest that the bottleneck is not simply that models lack personality words in the prompt, but that they fail to ground abstract psychological semantics into renderable visual evidence.

A key implication is that visual identity consistency and personality consistency should be evaluated separately. StoryDiffusion achieves the strongest baseline PCS, indicating strong visual stability, yet its PAS remains moderate. This demonstrates that mechanisms designed for appearance preservation do not automatically preserve psychological coherence. For narrative generation, this distinction matters: a character can remain visually recognizable while becoming behaviorally generic, emotionally overwritten, or inconsistent with their intended traits. Moreover, TBV's improvement in PAS is accompanied by moderate PCS reductions in several systems, suggesting that personality expressiveness and stochastic consistency are not fully orthogonal objectives. Future architectures should aim to resolve this tension, generating images that are both personality-aligned and generation-stable rather than trading one for the other. Future story-level T2I benchmarks should therefore assess not only whether a character looks the same, but also whether they continue to behave like the same psychological individual.

Our results also reveal a model-capacity pattern. TBV produces the largest gains in systems with stronger language-conditioning pathways, such as Qwen-Image-2512 and GPT Image 2, while CLIP-based or visual-consistency-focused systems show smaller improvements. This suggests that TBV does not merely add more prompt tokens; it provides a structured semantic interface that stronger text encoders can exploit more effectively. At the same time, the limited gains on some open-source models indicate that prompt-level grounding has an upper bound: if the text-conditioning pathway cannot preserve fine-grained behavioral cues, explicit visual constraints may still be weakened during generation.

The Emotion Override Effect further clarifies why personality consistency is difficult. Emotionally salient scenes significantly reduce personality alignment, indicating that transient affective context can dominate stable trait expression. This interaction is important because narrative characters rarely appear in neutral scenes. A robust story-generation model should express personality through emotion rather than replacing personality with emotion: a cautious character and an exploratory character may both feel fear, but they should not display fear in the same visual way.

Overall, TBV suggests that abstract semantic attributes must often be decomposed into observable cues before they can be rendered reliably. We hope future T2I work moves beyond appearance-level consistency and evaluates characters as psychologically stable agents across narrative contexts.

Limitations

This work has several limitations. First, our evaluation relies on GPT-4o as a VLM judge for personality-related metrics. Although GPT-4o shows moderate agreement with human annotations and our Visual Cues Lexicon was validated by psychology-trained experts, larger-scale human evaluation is still needed across more visual styles, character types, and affective contexts.

Second, our framework is based on the Big Five (OCEAN) model, which is widely used but culturally situated. The visual expression of personality may vary across cultures, and some cues in the TBV Lexicon may not fully capture high-context or non-Western behavioral norms.

Third, our current evaluation focuses on one focal character per shot. Although the source stories may contain multiple characters, relational personality dynamics, such as dominance, avoidance, or interpersonal adaptation between characters, are not explicitly modeled.

Fourth, TBV depends on the model's ability to interpret and preserve fine-grained visual cues in the prompt. Its gains may therefore be limited for smaller T2I models or models with weaker text encoders, where explicit behavioral constraints can still be weakened during generation.

Finally, our temporal metrics require multiple generations per prompt and per story shot. This makes the evaluation computationally expensive and may limit direct extension to long-form video generation models.

Code and Data Availability

The benchmark metadata, Visual Cues Lexicon, prompt templates, evaluation scripts, and anonymized annotation data will be publicly released upon acceptance under a research-only license.

Ethical Considerations

This work studies how personality traits can be represented through visual behavior in generated images. Such capability has useful applications in storytelling, education, game design, and character-centered creative tools, but it also raises potential risks. More psychologically coherent generated characters may increase anthropomorphism or be misused in deceptive social interaction.

There is also a risk that mapping personality traits to visual cues could reinforce stereotypes. To reduce this risk, our Visual Cues Lexicon focuses on neutral and observable behavioral signals, such as posture, gaze, gesture, spatial positioning, and scene interaction. We avoid demographic-associated physical features and do not use protected attributes as personality indicators. All expert raters and human annotators participated voluntarily with informed consent. No monetary compensation was provided. No personally identifiable information was collected, and all annotations were used only in aggregate form for research evaluation.

We separately disclose the use of AI assistance in Appendix A.

References

Ayan Banerjee, Josep Llad'os, Umapada Pal, and Anjan Dutta. 2025. Talediffusion: Multi-character story generation with dialogue rendering. *ArXiv*, abs/2509.04123.

Egon Brunswik. 1957. Perception and the representative design of psychological experiments. *American Journal of Psychology*, 70:480.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2025. R2i-bench: Benchmarking reasoning-driven text-to-image generation. In *Conference on Empirical Methods in Natural Language Processing*.

Kai Cheng, Mingzhe Yu, Lei Wu, Manyi Li, and YaJie Xu. 2025. Storyweaver: Consistent multi-character open-ended story generation. In *International Conference on Intelligent Computing*.

Xueqing Deng, Linjie Yang, Qihang Yu, Chenglin Yang, and Liang-Chieh Chen. 2025. Leveraging panoptic scene graph for evaluating fine-grained text-to-image generation. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15107–15116.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ArXiv*, abs/2208.01618.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Y. Qiao, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ArXiv*, abs/2307.04725.

Kimia Hamidieh, Haoran Zhang, Walter Gerych, Tom Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. *ArXiv*, abs/2411.00997.

Carina Ines Hausladen, Manuel Knott, Colin F. Camerer, and Pietro Perona. 2024. Social perception of faces in a vision-language model. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.

Yi He, Shengqi Dang, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. 2025. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15218–15228.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *ArXiv*, abs/2307.06350.

Black Forest Labs. 2024. Flux.1 [dev]. https://huggingface.co/black-forest-labs/FLUX.1-dev.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2023. Photomaker: Customizing realistic human photos via stacked id embedding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8640–8650.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60 2:175–215.

Laura P Naumann, Simine Vazire, Peter J Rentfrow, and Samuel D Gosling. 2009. Personality judgments based on physical appearance. *Personality and social psychology bulletin*, 35(12):1661–1671.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Elena Ryumina, Alexandr Axyonov, Darya Koryakovskaya, Timur Abdulkadirov, Angelina Egorova, Sergey Fedchin, Alexander Zaburdaev, and Dmitry Ryumin. 2026. Ssl-mepr: A semi-supervised multi-task cross-domain learning framework for multimodal emotion and personality recognition. *Machine Learning and Knowledge Extraction*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487.

Stability AI. 2024. Stable diffusion 3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large. Model card.

Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, Lihong Liu, and Xingang Wang. 2025. Humandreamer: Generating controllable human-motion videos via decoupled generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12391–12401.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Yuqing Yang, Hao Zhou, Haoran Tu, Zhiwen Hui, Shiqi Yan, Haoyang Li, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. 2026. Persona-e^2: A human-grounded dataset for personality-shaped emotional responses to textual events.

Hu Ye, Jun Zhang, Siyi Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *ArXiv*, abs/2308.06721.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024. Storydiffusion: Consistent self-attention for long-range image and video generation. *ArXiv*, abs/2405.01434.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, Xuanyang Zhang, Xianfang Zeng, Zhewei Huang, Gang Yu, and Chi Zhang. 2025. Vistorybench: Comprehensive benchmark suite for story visualization. *ArXiv*, abs/2505.24862.

A AI Assistance Disclosure

AI tools were used as auxiliary support during the preparation of this paper. Their use was lim-

ited to literature exploration, grammar and language checking, LaTeX formatting assistance, figure/table polishing suggestions, and code-related troubleshooting.

All research questions, benchmark design, experimental settings, data construction, metric definitions, result interpretation, and scientific claims were determined and verified by the authors. The authors reviewed, edited, and approved the final manuscript, and take full responsibility for its content.

No AI tool was used to generate experimental results, fabricate data, alter evaluation outcomes, or replace human expert judgments. All reported quantitative results are derived from the described experimental pipeline and validation procedures.

B VLM Judge Prompts

This appendix lists the prompts and decoding parameters used for VLM-based evaluation. All prompts are sent as user messages together with the evaluated image; no separate system-role message is used. The API is called with `response_format={"type": "json_object"}` to enforce JSON-formatted outputs.

B.1 PAS Task: OCEAN Annotation

The PAS score is computed as the centered cosine similarity between the VLM-inferred OCEAN vector and the preset OCEAN vector.

```
You are a personality psychology expert analyzing
     a storybook illustration.
Scene: {scene_text}
Plot: {plot_text}
Focus on these character cues (even for non-human
     characters like animals):
- Interaction style: approaching vs retreating,
     embracing vs distant
- Spatial positioning: center/dominant vs
     peripheral/small
- Action energy: dynamic/expansive vs
     still/contained
- Emotional expression: open/warm vs closed/tense
     (in posture and gesture, not just face)
- Overall character demeanor conveyed by the
     illustration
Rate each Big Five dimension from 1 (very low) to
     7 (very high).
Do NOT default to 4 -- commit to a direction based
     on what you observe.
Respond ONLY with JSON:
{"observations": "...", "O": 4, "C": 4, "E": 4,
     "A": 4, "N": 4, "reasoning": "..."}
```

B.2 RS Task: Blind Personality Inference

The RS task evaluates whether personality can be inferred from the generated image without prior personality labels.

```
You are a personality psychology expert. Look at
     this storybook illustration.
WITHOUT any prior personality information, analyze
     ONLY the visual cues:
Plot context (NO personality labels): {plot_text}
Focus on:
- How the character interacts with others or the
     environment
- Body posture: open/closed, dynamic/static,
     approaching/retreating
- Spatial dominance: center vs edge, large vs small
- Emotional tone conveyed through gesture and
     composition
Infer Big Five traits from 1 (very low) to 7 (very
     high).
Do NOT default to 4 -- commit to a direction.
Respond ONLY with JSON:
{"visual_evidence": "...", "inferred_O": 4,
     "inferred_C": 4, "inferred_E": 4,
     "inferred_A": 4, "inferred_N": 4,
     "confidence": "medium", "reasoning": "..."}
```

B.3 SCS-Reasoning Task

The SCS-Reasoning prompt evaluates whether the generated image reflects the expected personality-driven behavior chain.

```
You are evaluating whether a generated image
     reflects a character's personality-driven
     behavior chain.
Plot: {plot_text}
Scene: {scene_text}
Reasoning trace (personality -> behavior ->
     expected visual):
{reasoning_trace}
Check:
1. Does the character's appearance match the
     predicted behavioral outcome?
2. Are specific visual manifestations from the
     trace present?
3. Rate alignment 0.0-1.0
Respond ONLY with JSON:
{"elements_found": ["..."], "elements_missing":
     ["..."], "alignment_score": 0.0, "rationale":
     "..."}
```

B.4 VLM Generation Parameters

All evaluations use `gpt-4o` with temperature T=0.2, `max_completion_tokens` = 2000, n=1, high image detail, JSON response format, streaming disabled, 120 s timeout, and up to 3 retries.

B.5 Phase T: Trait Anchoring Prompt

Used in Phase T for each of the k=5 independent OCEAN inferences per character.

```
You are a personality psychology expert. Given the
     following character
description and narrative context, infer the
     character's Big Five (OCEAN)
personality profile.
Character: {character_description}
Narrative context: {story_synopsis}
Rate each dimension from 1 (very low) to 7 (very
     high).
Do NOT default to 4 -- commit to a direction based
     on the narrative evidence.
Respond ONLY with JSON:
{"O": 4, "C": 4, "E": 4, "A": 4, "N": 4,
     "reasoning": "..."}
```

When NDI $\in [0.15, 0.33]$, the following two-round debate critic prompt is appended:

```
The following OCEAN estimate was produced for this
     character: {estimate}.
```

```
Identify any dimensions where the estimate seems
     inconsistent with the
narrative evidence and propose corrections with
     justification.
Respond ONLY with JSON:
{"revised_O": 4, ..., "revised_N": 4, "critique":
     "..."}
```

B.6 Phase B: Behavioral Deduction Prompt

Used in Phase B to generate the behavioral reasoning trace r.

```
You are a character behavior analyst. Given a
     character's personality profile
and a narrative scene, describe specifically how
     the character would behave
in this scene. Your description must reflect the
     character's stable
personality, not just the emotional tone of the
     scene.
OCEAN profile: {ocean_vector}
Scene: {scene_description}
Respond ONLY with JSON:
{"behavioral_trace": "...", "key_behaviors":
     ["...", "..."],
 "reasoning": "..."}
```

The critic prompt for reverse inference (behavioral deviation check):

```
Given this behavioral description: {trace}, infer
     the implied OCEAN profile
of the character exhibiting this behavior.
Respond ONLY with JSON:
{"O": 4, "C": 4, "E": 4, "A": 4, "N": 4,
     "reasoning": "..."}
```

B.7 Phase V: Vision Mapping Prompt

Used in Phase V to select and compose the final visual constraint string.

```
Given the following behavioral description and the
     character's OCEAN profile,
select the most relevant visual cues from the TBV
     Lexicon and compose them
into a visual constraint string suitable for
     injection into a T2I prompt.
Behavioral trace: {behavioral_trace}
OCEAN profile: {ocean_vector}
Available cues (tier-filtered): {lexicon_subset}
Apply the priority rule: facial/gaze cues >
     posture cues >
clothing/scene cues. Return 4-6 non-conflicting
     cues.
Respond ONLY with JSON:
{"selected_cues": ["...", "..."],
     "constraint_string": "..."}
```

C Expert Validation of the Visual Cues Lexicon

We recruited 18 raters with graduate-level training in psychology, including Master's and doctoral students specializing in personality or clinical psychology with demonstrated familiarity with the Big Five model and affective theories. Raters were recruited through academic research networks and psychology graduate student mailing lists. Experts completed three evaluation tasks independently without discussion.

Before participation, all raters were informed that their responses would be used for research purposes to validate the Visual Cues Lexicon and the TBV pipeline. The instructions explained that their judgments would be analyzed only in aggregate form and reported anonymously. Participation was voluntary, and no personally identifiable information was collected. By completing the validation tasks, raters provided informed consent for their anonymized responses to be used in this study.

C.1 Task 1: Dimension Mapping

Each expert was presented with a visual cue description (e.g., "character leans forward with open arms") and asked to classify it into one of 10 OCEAN poles (O+, O−, C+, C−, E+, E−, A+, A−, N+, N−). The task format is illustrated below:

Task 1 Example Item

Visual cue: "The character stands with arms crossed, avoids eye contact, and maintains a large interpersonal distance."
Question: Which OCEAN pole does this cue most strongly indicate?
Options: O+ O− C+ C− E+ E− A+ A− N+ N−

Aggregated results ($n = 18$):

- Exact-match hit rate: 56.9% (chance = 10%, lift = 5.7×)
- Dimension-level accuracy (ignoring polarity): 58.3% (chance = 20%, lift = 2.9×)
- Fleiss' κ (10 categories): 0.757 (substantial agreement)

Table 3 reports per-pole accuracy.

Pole	Hit Rate (%)	# Cues	Lift
O+	68.8	16	6.9×
O−	20.3	16	2.0×
C+	78.3	15	7.8×
C−	74.2	16	7.4×
E+	67.2	16	6.7×
E−	68.8	12	6.9×
A+	67.0	14	6.7×
A−	34.4	12	3.4×
N+	51.0	13	5.1×
N−	33.7	13	3.4×

Table 3: Per-pole exact-match accuracy ($n = 18$). All poles exceed the 10% chance baseline. Low poles of O, A, and N show lower accuracy, reflecting genuine psychological overlap between conceptually adjacent dimensions (e.g., O− frequently mapped to C+).

Table 4 reports accuracy grouped by visual signal type.

Signal Type	Hit Rate (%)	# Cues
Facial expression	48.4	23
Body posture	59.0	25
Clothing	63.2	34
Color palette	61.4	11
Scene environment	54.5	50

Table 4: Dimension mapping accuracy by signal type. Clothing cues show the highest discriminability; facial expression cues are most ambiguous.

C.2 Task 2: Trait-Behavior Concordance

Each expert rated the semantic fit between a personality trait and its corresponding behavioral description on a 5-point Likert scale. The task format:

Task 2 Example Item

Trait: High Agreeableness (A+)
Behavioral description: "The character approaches others with open posture, maintains warm eye contact, and positions themselves at close interpersonal distance."
Question: How well does this behavioral description reflect the given personality trait?

Scale: 1 (Strongly Disagree) – 2 (Disagree) – 3 (Neutral) – 4 (Agree) – 5 (Strongly Agree)

Aggregated results ($n = 18$, 50 items):

- Mean rating: 4.85 ($SD = 0.40$)
- Percentage rated $\geq$ 4 (Agree or Strongly Agree): 98.3%
- S-CVI/Ave: 0.978 (excellent, threshold $\geq$ 0.80)
- Items with I-CVI $\geq$ 0.78: 30/30 (100%)

C.3 Task 3: Scenario Coherence

Experts rated 30 representative scenarios—six per Russell's macro-emotion quadrant (High-Arousal-Positive, High-Arousal-Negative, Low-Arousal-Positive, Low-Arousal-Negative) plus six Neutral—for end-to-end pipeline coherence. Each scenario presented a character's OCEAN profile, a narrative scene, and the complete TBV-generated visual cue fragment. Experts rated whether the visual cues were psychologically plausible for the given character and context. The task format:

Task 3 Example Item

Character profile: O=5, C=3, E=6, A=5, N=2 (high extraversion, low neuroticism)
Scene (High-Arousal-Positive): "The character wins an unexpected award at a community gathering."
TBV-generated visual cues: "The character steps forward with wide-open arms, beaming smile, upright and expansive posture, direct eye contact with the audience, animated hand gestures, relaxed facial muscles with no tension."
Question: How well do these visual cues reflect a psychologically plausible response for this character in this scene?

Scale: 1 (Strongly Disagree) – 2 (Disagree) – 3 (Neutral) – 4 (Agree) – 5 (Strongly Agree)

Aggregated results ($n = 18$, 30 scenarios):

- Mean coherence rating: 4.80 ($SD = 0.46$)
- Percentage rated ≥ 4: 97.5%
- S-CVI/Ave: 0.975 (excellent)

All thirty scenarios received unanimous or near-unanimous endorsement, confirming that the TBV pipeline maintains psychological plausibility across diverse emotional contexts.

C.4 Effect Sizes

To contextualize the magnitude of expert agreement:

- **Task 1 vs. chance:** Cohen's $h = 1.103$ (large effect), confirming that expert classification performance far exceeds random guessing.
- **Task 2 vs. neutral midpoint (3.0):** One-sample t-test: $t(179) = 68.02$, $p < 10^{-130}$, Cohen's $d = 5.07$ (large effect), indicating concordance ratings are strongly skewed toward agreement.

D SCS Weight Sensitivity

We set $\alpha = 0.4$ in the SCS formula to emphasize context-specific behavioral coherence over global trait recovery, reflecting the principle that story-level personality evaluation should prioritize *how* a character behaves in a given scene over whether the overall trait profile is recovered.

To verify that this choice does not affect our conclusions, we swept $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ and recomputed model rankings. As shown in Table 5, model rankings remain identical for $\alpha \in [0.3, 0.7]$ (Kendall's $\tau = 1.0$, $p < .001$). At the extremes ($\alpha = 0.2$ and 0.8), $\tau = 0.905$ ($p < .01$), with only the two lowest-ranked models (SDXL and StoryDiffusion) exchanging positions. These results confirm that our conclusions are not

sensitive to the weighting choice.

α	Kendall's τ	p
0.2	0.905	.003
0.3	1.000	<.001
0.4	1.000	—
0.5	1.000	<.001
0.6	1.000	<.001
0.7	1.000	<.001
0.8	0.905	.003

Table 5: Kendall's τ between model rankings at each α and the paper's $\alpha = 0.4$. Rankings are fully stable across the central range.

E Full Visual Cues Lexicon

This appendix presents the complete set of 150 visual cues used in our annotation framework, organized by OCEAN personality dimensions and polarity. Tier annotations are C = Core ($\geq$5 / $\leq$3), E = Extended ($\geq$6 / $\leq$2), and F = Fine (=7 / =1). Cue types are Fa = facial expression, Bo = body posture, Cl = clothing, Co = color palette, and Sc = scene environment.

E.1 Openness to Experience

E.1.1 High Openness

Tier	Type	Visual cue
C	Fa	Animated or expressive facial micro-movements
C	Fa	Wide open or exploratory gaze
C	Bo	Forward-leaning curious posture
C	Bo	Relaxed open hand gestures
C	Bo	Slight head tilt suggesting interest
E	Cl	Layered or mixed-texture clothing
E	Cl	Asymmetric or flowing silhouettes
E	Cl	Artistic or handcrafted accessories
E	Co	Warm or vivid accent colors (mustard, rust, teal)
E	Cl	Unconventional or eclectic outfit combinations
F	Sc	Bookshelf, map, or artwork in background
F	Sc	Natural outdoor light or dappled sunlight
F	Sc	Creative or slightly cluttered studio environment
F	Sc	Plants, cultural objects, or travel items as props
F	Sc	Warm diffuse ambient lighting

E.1.2 Low Openness

Tier	Type	Visual cue
C	Fa	Composed neutral or flat facial expression
C	Fa	Forward-facing steady gaze, no exploratory movement
C	Bo	Upright symmetrical stance
C	Bo	Arms close to body or hands clasped
C	Bo	Deliberate controlled movements, no fidgeting
E	Cl	Well-fitted traditional or classic silhouette
E	Cl	No accessories or minimal functional accessories
E	Cl	Conventional or uniform-style clothing
E	Co	Muted neutral palette (navy, grey, beige, black)
E	Co	Single-color or subtle pattern outfit
F	Sc	Plain or uniform background wall
F	Sc	Orderly indoor space with consistent lighting
F	Sc	Minimal props, no clutter
F	Sc	Flat even artificial lighting
F	Sc	No decorative or cultural objects visible

E.2 Conscientiousness

E.2.1 High Conscientiousness

Tier	Type	Visual cue
C	Fa	Subtle confident closed-lip expression, no tension
C	Fa	Focused and direct gaze
C	Bo	Tall upright posture with engaged core
C	Bo	Deliberate purposeful hand placement, no fidgeting
C	Bo	Still body with minimal movement
E	Cl	Neatly pressed formal or semi-formal attire
E	Cl	Classic or professional silhouette
E	Cl	Minimal but purposeful accessories
E	Co	Coordinated color palette, no clashing tones
E	Cl	Well-maintained fabric with no wrinkles
F	Sc	Clean organized desk or workspace in background
F	Sc	Planner, checklist, or clock visible as prop
F	Sc	Neat filing or structured shelving in background
F	Sc	Consistent balanced overhead lighting
F	Sc	Nothing out of place in background

E.2.2 Low Conscientiousness

Tier	Type	Visual cue
C	Fa	Relaxed jaw, slightly open mouth, unfocused expression
C	Fa	Wandering or unfocused gaze
C	Bo	Slouched or leaning posture
C	Bo	Hands fidgeting or in pockets
C	Bo	Weight shifted to one side, asymmetric stance
E	Cl	Wrinkled or untucked casual clothing
E	Cl	Shirt untucked or collar askew
E	Cl	Comfort-first athleisure or oversized fit
E	Co	Mismatched or random color combinations
E	Cl	Worn or slightly disheveled items
F	Sc	Cluttered or disorganized background
F	Sc	Scattered objects or papers visible
F	Sc	Unmade surface or piled items in background
F	Sc	Dim or inconsistent ambient lighting
F	Sc	Items visibly out of place in background

E.3 Extraversion

E.3.1 High Extraversion

Tier	Type	Visual cue
C	Fa	Broad confident smile showing teeth (Duchenne smile)
C	Fa	Direct sustained eye contact
C	Bo	Arms extended outward, open gesturing
C	Bo	Erect expansive posture occupying space
C	Bo	Body oriented toward viewer or others
E	Cl	Attention-drawing patterns or prints
E	Cl	Open necklines or expressive cuts
E	Cl	Bold statement accessories
E	Co	Bright warm colors (red, orange, yellow, warm pink)
E	Cl	Fitted clothing that emphasizes physical presence
F	Sc	Social gathering or crowd in background
F	Sc	Bright high-key lighting
F	Sc	Open spacious environment
F	Sc	Stage, podium, or center-of-room position
F	Sc	Lively colorful setting

E.3.2 Low Extraversion (Introversion)

Tier	Type	Visual cue
C	Fa	Minimal facial expression, closed mouth, neutral affect
C	Fa	Downward or averted gaze
C	Bo	Arms crossed or drawn close to body
C	Bo	Body angled away from viewer
C	Bo	Weight inward, occupying minimal space
E	Cl	Close-fitting minimal silhouette, no statement elements
E	Cl	Covered or layered clothing
E	Cl	No statement accessories
E	Co	Dark muted tones (charcoal, navy, deep grey)
E	Cl	Understated fabric textures
F	Sc	Quiet solitary corner or alcove
F	Sc	Dim soft interior lighting
F	Sc	Single chair or isolated seating
F	Sc	Window with soft natural light, solitary setting
F	Sc	Few or no other people visible in background

E.4 Agreeableness

E.4.1 High Agreeableness

Tier	Type	Visual cue
C	Fa	Gentle closed-lip or soft open smile
C	Fa	Slight head tilt indicating empathy
C	Bo	Open palms facing upward
C	Bo	Body leaning slightly forward toward others
C	Fa	Soft relaxed brow, no tension
E	Cl	Comfortable relaxed-fit clothing
E	Cl	Rounded gentle silhouettes, no sharp styling
E	Cl	Natural or organic fabric textures
E	Co	Soft warm tones (peach, cream, sage, dusty rose)
E	Cl	No sharp or aggressive styling elements
F	Sc	Warm cozy interior with soft furnishings
F	Sc	Plants, flowers, or natural elements as props
F	Sc	Warm golden or candlelit ambient lighting
F	Sc	Collaborative or shared space in background
F	Sc	Soft textures like cushions or throws visible

E.4.2 Low Agreeableness

Tier	Type	Visual cue
C	Fa	Sharp sustained eye contact, unblinking
C	Fa	Tense jaw or clenched teeth visible
C	Fa	Flat or downturned expression
C	Bo	Arms crossed high on chest or hands on hips
C	Bo	Body square-on, space-assertive stance
E	Cl	Sharp angular silhouettes or structured cuts
E	Cl	Stiff formal or militaristic styling
E	Cl	Power-signaling or imposing accessories
E	Co	Cold tones (black, steel grey, icy blue)
E	Co	High-contrast color blocking
F	Sc	Stark industrial or minimal background
F	Sc	Hard directional lighting creating strong shadows
F	Sc	Empty impersonal space
F	Sc	No decorative elements, bare walls
F	Sc	Cool color temperature lighting

E.5 Neuroticism

E.5.1 High Neuroticism

Tier	Type	Visual cue
C	Fa	Furrowed brow or creased forehead
C	Fa	Tight pressed lips or bitten lip
C	Fa	Downcast or darting anxious gaze
C	Bo	Raised and tensed shoulders
C	Bo	Hands wringing, fidgeting, or gripping
E	Cl	Disheveled or slightly unkempt presentation
E	Cl	Clothing that looks pulled at or self-touched
E	Cl	Self-concealing layered or oversized items
E	Co	Dark or dull desaturated colors
E	Co	Mismatched or incoherent outfit choices
F	Sc	Dark shadowy or low-key environment
F	Sc	Confined or enclosed space
F	Sc	Dramatic or harsh lighting with strong shadows
F	Sc	Visually chaotic or disordered background
F	Sc	Isolated setting with no comfort elements

E.5.2 Low Neuroticism (Emotional Stability)

Tier	Type	Visual cue
C	Fa	Calm smooth brow with no creasing
C	Fa	Steady unhurried gaze
C	Bo	Relaxed dropped shoulders
C	Bo	Easy natural hand placement at sides
C	Bo	Slow deliberate composed movements
E	Cl	Neat but relaxed presentation
E	Cl	No signs of self-touching or garment distortion
E	Cl	Well-coordinated comfortable outfit
E	Co	Clean neutral or balanced tones (soft white, light grey, calm blue)
E	Cl	Natural fabric, nothing stiff or aggressive
F	Sc	Bright airy open environment
F	Sc	Natural daylight or soft warm light
F	Sc	Calm water, greenery, or nature elements
F	Sc	Clean uncluttered background
F	Sc	Spacious setting with comfortable proportions

F Dataset Statistics

ViStoryBench-Personality is built on the narrative foundation of ViStoryBench (Zhuang et al., 2025), spanning 80 stories and 1,317 sequential shots. Table 6 summarizes the key statistics.

Statistic	Value
Total stories	80
Total sequential shots	1,317
Unique characters	80
Mean shots per story	16.5
Evaluation shots	213
Conditions per shot	4 (C1–C4)
T2I models evaluated	7
Total generated images	15,336

Table 6: Overview of the ViStoryBench-Personality dataset.

Evaluation Subset. The 213 evaluation shots were selected via balanced sampling across five affective scene categories derived from Russell's Circumplex Model (Russell, 1980). This design ensures that the evaluation covers both neutral and emotionally salient contexts, allowing us to analyze whether scene-level affect overrides personality-specific visual cues. Table 7 reports the affective category and visual difficulty distribution.

Category	Count	%
Neutral	44	20.7
Low-arousal positive	41	19.2
Low-arousal negative	42	19.7
High-arousal positive	40	18.8
High-arousal negative	46	21.6
Visual difficulty		
Direct (facial/gestural)	153	71.8
Indirect (spatial/implicit)	60	28.2

Table 7: Distribution of affective scene categories and visual difficulty across 213 evaluation shots.

OCEAN Distribution. Figure 3 shows the Big Five personality distribution across unique characters. The distribution reflects the narrative archetypes in the source corpus: high Conscientiousness is overrepresented, corresponding to goal-driven protagonists in adventure stories, while extreme low-pole scores are less frequent. Neutral scores also appear across all dimensions, providing a natural control group for evaluating whether T2I models default to generic personality expression when no strong directional trait signal is present.

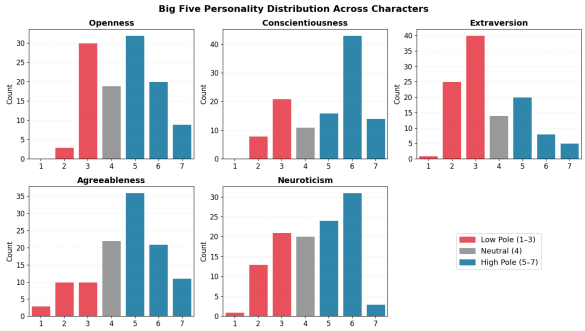

Figure 3: Big Five personality distribution across unique characters. Red = low pole (1–3), grey = neutral (4), blue = high pole (5–7).

G Prompt Construction Example

Table 8 shows a representative prompt construction example across the four ablation conditions.

Cond.	Prompt content
C1	Alice wishes she could shrink to fit through the small door. Daytime, inside the hall, disappointed and hopeful atmosphere. A curious girl with golden hair and blue eyes.
C2	C1 + *[Personality: High Openness, Low Conscientiousness, High Extraversion, Low Neuroticism]*
C3	C2 + *feeling Disappointment*
C4	C1 + *animated facial micro-movements | wide open exploratory gaze | forward-leaning curious posture | relaxed open hand gestures | slight head tilt suggesting interest*

Table 8: Example prompt construction across four abstraction levels for Alice (O=7, C=3, E=5, A=4, N=3). C1–C3 test non-visual textual information, while C4 tests concrete visual grounding.

H Implementation Details

H.1 T2I Model Configurations

Table 9 summarizes the generation parameters for all seven models. Open-source models were run locally; closed-source models were accessed via their respective APIs.

H.2 Hardware and Runtime

Open-source models were run on NVIDIA RTX 4090 (24 GB) or A100 (80 GB). Qwen-2512 and StoryDiffusion were run on A100 due to higher memory requirements; remaining open-source models used RTX 4090 with `sequential_cpu_offload`. Table 10 reports approximate per-image generation time.

Model	Steps	CFG	Resolution	Arch.
SDXL	28	7.5	1024×1024	UNet+CLIP
FLUX.1-dev	28	3.5	1024×1024	DiT+T5
SD3.5-Large	28	4.5	1024×1024	MMDiT+3Enc
Qwen-2512	50	4.0	1328×1328	MMDiT+LLM
StoryDiff.	50	5.0	1024×1024	SDXL+CSA
GPT Image 2	—	—	1254×1254	Closed
Nano Banana	—	—	1408×768	Closed

Table 9: Generation parameters for the seven T2I models.

Model	Device	Time/image
SDXL	RTX 4090	∼8 s
FLUX.1-dev	RTX 4090	∼12 s
SD3.5-Large	RTX 4090	∼15 s
Qwen-2512	A100	∼25 s
StoryDiffusion	A100	∼20 s
GPT Image 2	API	∼15 s
Nano Banana	API	∼15 s

Table 10: Approximate per-image generation time.

H.3 Generation Protocol

Each of the 213 evaluation shots was generated under 4 ablation conditions (C1–C4). For the five open-source models (SDXL, FLUX.1-dev, SD3.5-Large, Qwen-2512, StoryDiffusion), each condition was generated 3 times with different random seeds, yielding $213 \times 4 \times 3 \times 5 = 12{,}780$ images. For the two closed-source models (GPT Image 2, Nano Banana), a total of $1{,}278 \times 2 = 2{,}556$ images were generated. The complete evaluation set comprises 15,336 generated images.

H.4 VLM Evaluation Configuration

Personality assessment was conducted using GPT-4o via API with the following parameters: temperature T=0.2, `max_completion_tokens` = 2000, n=1, high image detail, and JSON response format (`response_format={"type": "json_object"}`). Each generated image was evaluated on three tasks (PAS, RS, SCS-Reasoning), and consecutive shots within each story were evaluated for PCS.

H.5 TBV LLM Configuration

All LLM-based operations in TBV (Phase T, Phase B, and Phase V) are implemented using GPT-4o (`gpt-4o-2024-11-20`) with temperature T=0.2 and `max_completion_tokens` = 1000. Phase T samples k=5 independent OCEAN inferences per character. If the resulting NDI falls in $[0.15, 0.33]$, a two-round multi-agent debate is triggered: a sec-

ond GPT-4o instance critiques the initial estimate and proposes a revised anchor, after which the mean of non-outlier estimates is accepted. Phase B applies the self-correcting loop with a maximum of three regeneration attempts before falling back to the best-scoring trace. The full prompts for all three phases are listed in Appendix B (Sections B.5–B.7).

H.6 Cost Estimation

Table 11 summarizes the estimated API and compute cost.

Component	Est. Cost (USD)
GPT Image 2 generation	∼$240
Nano Banana generation	∼$240
VLM evaluation (GPT-4o)	∼$180
RTX 4090 GPU hours	∼200 h
A100 GPU hours	∼20 h
Total API cost	**∼$660**

Table 11: Estimated cost breakdown for the full evaluation pipeline. GPU hours reported separately as institutional compute.

I VLM-Human Agreement Details

To validate GPT-4o as an automated personality judge, we conducted a human agreement study. We sampled a stratified calibration subset of 100 generated images covering multiple T2I systems, all five affective scene categories, and both direct and indirect visual-expression shots. Six psychology-trained annotators, all Master's or doctoral students in psychology-related fields, independently rated each image on the five OCEAN dimensions using the same 7-point scale as the VLM judge. Final human scores were computed as the mean rating per dimension per image. Annotators participated voluntarily with informed consent; no personally identifiable information was collected, and all ratings were analyzed only in aggregate form. We then compared GPT-4o predictions with the averaged human ratings using rank correlation, absolute error, within-one-point accuracy, and polarity agreement.

I.1 Overall Agreement

Table 12 reports aggregate agreement statistics.

The Spearman $\rho = 0.611$ exceeds the typical accuracy of human personality inference from static photographs (Naumann et al., 2009), supporting the use of GPT-4o as a scalable proxy for human personality judgment.

Metric	Value
Spearman ρ	0.611
Pearson r	0.613
MAE (7-point scale)	0.871
Within ± 1 point	81.3%
Polarity agreement	86.6%

Table 12: VLM-human agreement on personality evaluation. Polarity agreement measures whether VLM and humans assign the same trait direction (above vs. below neutral); chance $= 50\%$.

I.2 Per-Dimension Polarity Agreement

Table 13 reports polarity agreement by OCEAN dimension. GPT-4o and human annotators agreed on trait direction in 86.6% of cases overall, with per-dimension agreement ranging from 77.3% (Extraversion) to 97.6% (Openness).

Dimension	Agreement	Chance
Openness (O)	97.6%	50%
Conscientiousness (C)	86.8%	50%
Extraversion (E)	77.3%	50%
Agreeableness (A)	81.2%	50%
Neuroticism (N)	89.1%	50%
Overall	**86.6%**	**50%**

Table 13: Polarity agreement between VLM and human annotators by OCEAN dimension (chance $= 50\%$). Neutral ratings ($= 4$) are excluded.

The relatively lower agreement on Extraversion may reflect the difficulty of distinguishing stable social-energy cues from transient emotional expressions in narrative illustrations, consistent with the Emotion Override Effect observed in our main analysis.

I.3 Score-Level vs. Directional Agreement

The moderate score-level agreement ($\rho = 0.611$) combined with high polarity agreement (86.6%) indicates that GPT-4o reliably captures the *direction* of personality expression while showing greater variability in exact magnitude estimates. This pattern is well-suited to our evaluation framework, which uses centered cosine similarity and is therefore more sensitive to directional agreement than to exact score magnitude.

J Full Ablation Results

Table 14 reports the complete C1–C4 ablation results for the five open-source models.

K Statistical Tests for Ablation Conditions

To formally validate the claim that abstract personality descriptors do not reliably improve PAS over baseline prompts (C1 $\approx$ C2 $\approx$ C3), we conducted paired Wilcoxon signed-rank tests and two one-sided tests (TOST) for equivalence ($\delta = 0.05$) between C1 and each of C2, C3, and C4 across all models ($n = 213$ per pair for open-source models). Table 15 reports the full results.

Of the ten C1 vs. C2/C3 comparisons across open-source models, nine show no significant improvement (Wilcoxon $p > .05$). The single exception is Qwen-Image-2512 under C2 ($p = .027$, $r = 0.21$), where the gain ($\Delta = +0.027$) is small in magnitude and inconsistent across models—SDXL shows a decrease of comparable magnitude under C2 ($\Delta = -0.054$, $p = .054$). TOST equivalence is confirmed for 6 of 10 comparisons; the remaining four involve either a borderline positive shift (Qwen-Image-2512 C2/C3, TOST $p = .051/.054$) or a negative shift exceeding the equivalence margin (SDXL C2/C3). This pattern indicates that abstract descriptors produce *inconsistent* rather than *uniformly absent* effects—neither constitutes reliable trait grounding.

In contrast, C1 vs. C4 yields significant improvements for FLUX.1-dev ($p = .005$, $r = 0.24$), Qwen-Image-2512 ($p < .001$, $r = 0.43$), and both closed-source models ($p < .001$). SDXL and SD3.5-Large show non-significant C4 gains ($p = .160$ and $.634$), consistent with the text-encoder capacity pattern discussed in Section 6.3.

L Emotion Override Statistical Analysis

We analyze PAS across five affective scene categories: Neutral, Low-Arousal Positive, Low-Arousal Negative, High-Arousal Positive, and High-Arousal Negative. After removing records with missing PAS or affective labels, the analysis contains 5,081 valid observations. Table 16 reports descriptive statistics for each category.

A one-way ANOVA shows a significant effect of affective category on PAS, $F(4, 5076) = 20.06$, $p < .001$. Because Levene's test indicates unequal variances across groups ($p < .001$), we also conduct a non-parametric Kruskal–Wallis test, which confirms the same pattern ($H = 82.00$, $p < .001$). Tukey HSD post-hoc comparisons show that Neutral differs significantly from all four emotionally salient categories, while several comparisons among emotional categories are not significant. This supports the Emotion Override Effect: emotionally salient contexts tend to reduce visually readable personality alignment relative to neutral contexts.

M Indirect Expression Gap (Null Result)

We hypothesized that shots requiring indirect personality expression (e.g., proxemics, spatial positioning, implicit behavioral cues) would yield lower personality alignment than shots with direct expression channels (e.g., facial expression, overt gestures). To test this, we compared PAS between direct and indirect shots using independent-samples Welch's t-tests for each model, pooling across all four conditions. Table 17 reports the direct–indirect PAS comparison for each model.

No model showed a significant difference between direct and indirect scenes (all $p > .20$). The mean PAS gap across seven models was $|\Delta| = 0.016$, with the direction inconsistent (three positive, four negative).

Interpretation. This null result admits several explanations. First, current T2I models may not produce sufficiently fine-grained differences in indirect personality channels (e.g., interpersonal distance, spatial dominance) for VLM-based evaluation to detect. Second, the VLM judge itself may lack sensitivity to subtle proxemic and compositional cues, attending primarily to facial and gestural signals regardless of scene type. Third, the direct/indirect distinction may be less categorical than hypothesized: even "indirect" scenes often contain some facial or postural information that the VLM can exploit. Future work should design stronger indirect manipulations (e.g., scenes depicting only environmental or object-level personality cues without any character face visible) and employ more fine-grained evaluation metrics targeting specific visual channels.

N Per-Model Qualitative and Quantitative Analysis

This appendix provides a model-wise qualitative and quantitative analysis complementing the aggregate results in Table 2. Figures 4 and 5 use the same representative story shot across systems to visualize

Table 14: Full ablation study results. C1 = base narrative prompt (plot + appearance); C2 = C1 + raw OCEAN trait tags; C3 = C2 + deduced scene-level emotion labels; C4 (Full TBV) = C1 + explicit visual cues from the TBV lexicon. Appending abstract psychological vocabulary (C2, C3) yields marginal or negative effects, confirming the semantic gap. Only concrete visual grounding (C4) produces consistent gains. $n = 213$ per condition for open-source models.

	SDXL		**StoryDiffusion**		**FLUX.1-dev**		**SD3.5-Large**		**Qwen-2512**	
Cond	PAS	RS	PAS	RS	PAS	RS	PAS	RS	PAS	RS
C1	0.446	0.429	0.480	0.467	0.464	0.469	0.466	0.493	0.518	0.546
C2	0.392	0.382	0.462	0.449	0.483	0.475	0.452	0.469	0.545	0.560
C3	0.408	0.400	0.471	0.444	0.484	0.460	0.460	0.471	0.540	0.575
C4	0.467	0.423	0.502	0.508	0.522	0.552	0.472	0.488	0.633	0.641
Δ(C4–C1)	+4.8%	–1.4%	+4.6%	+8.8%	+12.6%	+17.7%	+1.5%	–1.0%	+22.1%	+17.4%

Model	Comparison	ΔPAS	Wilcoxon p	r	TOST p	Equiv.
FLUX.1-dev	C1 vs C2	+0.015	.161	0.13	.009	✓
	C1 vs C3	+0.020	.372	0.08	.045	✓
	C1 vs C4	+0.059	.005	0.24	—	—
SDXL	C1 vs C2	−0.054	.054	−0.17	.561	×
	C1 vs C3	−0.038	.134	−0.13	.279	×
	C1 vs C4	+0.021	.160	0.12	—	—
SD3.5-Large	C1 vs C2	−0.014	.703	−0.03	.018	✓
	C1 vs C3	−0.005	.996	0.00	.002	✓
	C1 vs C4	+0.007	.634	0.04	—	—
StoryDiffusion	C1 vs C2	−0.018	.436	−0.07	.043	✓
	C1 vs C3	−0.009	.478	−0.06	.012	✓
	C1 vs C4	+0.022	.148	0.12	—	—
Qwen-Image-2512	C1 vs C2	+0.027	.027*	0.21	.051	×
	C1 vs C3	+0.021	.307	0.10	.054	×
	C1 vs C4	+0.115	<.001	0.43	—	—
GPT Image 2	C1 vs C4	+0.108	<.001	0.50	—	—
Nano Banana	C1 vs C4	+0.094	<.001	0.39	—	—

Table 15: Paired Wilcoxon signed-rank tests and TOST equivalence tests ($\delta = 0.05$) for PAS across ablation conditions. r = rank-biserial correlation. Equiv. = TOST equivalence at $\delta = 0.05$: ✓ confirmed, × not confirmed, — not applicable. TOST is reported only for C1 vs. C2/C3 comparisons, which are the focus of the equivalence claim; C1 vs. C4 is tested as the intervention condition. *Significant at $p < .05$. Closed-source models are evaluated under C1 and C4 only.

Category	**N**	**Mean**	**SD**	**Median**
Neutral	1053	0.575	0.311	0.655
High-Arousal Positive	952	0.507	0.337	0.571
Low-Arousal Negative	998	0.498	0.338	0.541
High-Arousal Negative	1097	0.477	0.374	0.567
Low-Arousal Positive	981	0.446	0.351	0.474

Table 16: PAS distribution across affective scene categories. Neutral scenes obtain the highest mean PAS, while all emotionally salient categories show lower mean PAS.

Model	**Direct**	**Indirect**	t	p
SDXL	.430	.423	+0.24	.812
FLUX.1-dev	.497	.467	+1.15	.253
SD3.5-Large	.461	.466	−0.19	.847
StoryDiffusion	.477	.484	−0.29	.770
Qwen-2512	.551	.580	−1.25	.212
GPT Image 2	.590	.601	−0.32	.749
Nano Banana	.596	.576	+0.55	.584

Table 17: Direct vs. indirect PAS comparison (Welch's t-test). No model shows a significant difference ($p < .05$).

how different models respond to increasingly explicit prompting conditions. For open-source models, Figure 4 compares C1–C4; for closed-source APIs, Figure 5 compares C1 and C4 only because these systems perform internal prompt rewriting and are evaluated in benchmarking mode.

The qualitative patterns support the quantitative results. In C1–C3, most models preserve the broad scene layout and character identity, but their behav-

ior is often generic or dominated by the immediate scene emotion. For example, several outputs emphasize the corridor setting and anxious atmosphere without clear personality-specific evidence beyond appearance. In contrast, C4 more often introduces TBV-specified behavioral cues, such as exploratory gaze, forward-leaning posture, open gesture, or more active engagement with the environment. This shift matches the PAS and RS improvements in Table 2: models with stronger language-conditioning pathways show larger gains, while visual-consistency-oriented systems improve less.

N.1 SDXL

SDXL exhibits the weakest response to TBV among the evaluated systems. Its PAS increases only from 0.446 to 0.467, while RS slightly decreases from 0.429 to 0.423. This pattern suggests that CLIP-based text conditioning can partially absorb explicit visual cues, but remains limited in grounding abstract personality semantics into stable visual behavior. Qualitatively, SDXL often preserves the requested scene and character appearance, but the personality-relevant cues remain generic or emotionally dominated.

N.2 StoryDiffusion

StoryDiffusion provides a useful contrast between visual identity consistency and personality consistency. It achieves the strongest baseline PCS among open-source systems, but its baseline PAS remains moderate. This indicates that consistent self-attention improves cross-frame visual stability without necessarily preserving personality-relevant behavior. After TBV prompting, PAS improves from 0.480 to 0.502, but the gain remains smaller than models with stronger language understanding. This supports our claim that personality grounding is not solved by visual consistency mechanisms alone.

N.3 FLUX.1-dev

FLUX.1-dev shows a stronger response to TBV than SDXL and StoryDiffusion. PAS increases from 0.464 to 0.522, and RS increases from 0.469 to 0.552. This suggests that FLUX can better exploit concrete visual constraints once personality is translated into renderable cues. Qualitatively, TBV prompts tend to produce clearer posture, gaze, and interaction cues, although the model still occasionally prioritizes scene-level affect over stable personality expression.

N.4 SD3.5-Large

SD3.5-Large shows only marginal improvement under TBV, with PAS increasing from 0.466 to 0.472 and RS slightly decreasing from 0.493 to 0.488. This indicates that a stronger generative architecture does not automatically yield stronger personality grounding. One possible explanation is that the model may normalize or weaken fine-grained prompt details during generation, reducing the effect of additional visual constraints. Its results suggest that architectural sophistication alone is insufficient unless the text-conditioning pathway preserves personality-relevant behavioral cues.

N.5 Qwen-Image-2512

Qwen-Image-2512 is the strongest open-source responder to TBV. PAS increases from 0.518 to 0.633, RS from 0.546 to 0.641, and SCS from 0.665 to 0.701. This model benefits most from the TBV decomposition, suggesting that LLM-conditioned T2I systems are better able to use explicit visual constraints derived from abstract traits. Qualitatively, Qwen-Image-2512 more reliably renders exploratory gaze, forward-leaning posture, and spontaneous gestures when these cues are explicitly specified.

N.6 GPT Image 2

GPT Image 2 shows the strongest improvement among closed-source systems. PAS increases from 0.538 to 0.648, while RS increases from 0.556 to 0.659. This result supports the broader trend that models with stronger language understanding are more responsive to TBV prompting. However, because GPT Image 2 performs internal prompt rewriting, we evaluate it only under C1 and C4 conditions and do not use it for full C1–C4 ablation analysis.

N.7 Nano Banana

Nano Banana starts from a relatively strong baseline and still benefits substantially from TBV. PAS increases from 0.542 to 0.638, and RS increases from 0.551 to 0.661. PCS slightly decreases, suggesting that TBV may introduce more personality-specific variation across scenes while improving snapshot-level personality alignment. This trade-off reinforces the need to evaluate both alignment and stability rather than relying on a single consistency metric.

N.8 Summary

Across models, the largest TBV gains occur in LLM-conditioned or stronger language-conditioned systems, while CLIP-only or visual-consistency-focused systems show smaller gains. These model-wise patterns support two conclusions. First, visual identity preservation does not imply personality preservation. Second, the key bottleneck in trait collapse lies primarily in semantic grounding: models require concrete visual constraints before abstract personality traits become reliably renderable.

Model	C1	C2	C3	C4 / TBV
SDXL				
StoryDiffusion				
FLUX.1-dev				
SD3.5-Large				
Qwen-Image-2512				

Figure 4: Qualitative case study across five open-source systems on the same representative story shot. C1 uses only the base narrative prompt; C2 adds abstract OCEAN trait labels; C3 adds a deduced scene-level emotion label; C4 injects explicit TBV-derived visual constraints. Across models, C1–C3 often preserve the broad scene layout but tend to produce generic, weakly visible, or emotion-dominated behavior. C4 more consistently introduces personality-relevant visual evidence, such as exploratory gaze, forward-leaning posture, and open gestures, illustrating the qualitative mechanism behind the PAS and RS gains in Table 2.

Model	C1	C4 / TBV
GPT Image 2		
Nano Banana		

Figure 5: Qualitative case study for the two closed-source APIs. Because GPT Image 2 and Nano Banana perform internal prompt rewriting, we evaluate them only under C1 and C4 conditions. Compared with C1, C4 produces more explicit personality-relevant behavior, including more active body orientation, exploratory engagement with the environment, and less purely scene-driven affect. This qualitative shift is consistent with the quantitative PAS and RS gains reported in Table 2.