Prompt-Character Divergence: A Responsibility Compass for Human-AI Creative Collaboration

Anonymous Author(s)
Affiliation
Address
email

Abstract

Distinguishing genuine user intent from model-driven echoes, whether of copyrighted characters, familiar styles, or training-derived identities, has become critical for creators as generative AI brings visual content creation to millions. Yet most detection tools remain computationally heavy, opaque, or inaccessible to the people they most affect. We present Prompt–Character Divergence (PC-D), a lightweight metric that quantifies semantic drift—how far a generated image aligns with known visual identities beyond what the prompt predicts. PC-D supports creator agency and responsibility in shared authorship by mapping outputs along two axes, name proximity and model drift, to produce a responsibility compass with four creative-agency zones: model-driven risk, mixed attribution, safe co-creation, and user-driven intent. Evaluated on three open-source models and ten iconic characters, PC-D captures drift patterns consistent with human judgment and runs on consumer hardware. Rather than resolving attribution, PC-D functions as a creator-facing diagnostic for self-auditing, helping practitioners determine when outputs reflect their intent, when they reflect the model's learned biases, and how the two interact. The result is a practical, transparent aid that invites accessible, reflexive, and accountable human-AI collaboration.

1 Introduction

When millions can create sophisticated visual content with a few words, the nature of creative authorship changes. This shift is not merely technical—it redefines authorship as a shared, yet asymmetrical, collaboration between humans and generative systems, shaped by both user prompts and model priors. A prompt like "a video-game plumber in overalls" might yield an image uncannily similar to Mario, yet the source of that resemblance remains blurred. As these systems absorb vast artistic corpora, they can reproduce distinctive visual identities, carrying not only legal implications but also cultural memory and symbolic meaning, raising questions of ethical and affective stewardship in human–AI creation.

Existing resemblance- and provenance-detection tools, including invisible watermarking systems and open provenance standards like C2PA [1], are often too complex and expensive to be within practical reach of independent creators. This challenge is intensifying as the generative ecosystem fragments into specialized models. Open-source variants and fine-tuned systems, each trained on different datasets and aesthetics, display varying tendencies to resemble or replicate existing works. Creators now need more than generation capabilities: they require tools that clarify the interplay between model biases and human intent, enabling them to reason about resemblance risks and select AI collaborators aligned with their aims.

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

We present Prompt–Character Divergence (PC-D), a metric that decomposes generated outputs into user-directed and model-contributed components. Formally:

$$\text{PC-D}(x, p, c) = \cos\big(f_I(x), f_T(c)\big) - \cos\big(f_I(x), f_T(p)\big) \tag{1}$$

Here, x is the generated image, p the text prompt, and c the character name. f_I and f_T are the image and text encoders, and $\cos(\cdot, \cdot)$ denotes cosine similarity in a shared embedding space. We compute similarities in a CLIP-style dual-encoder space [2, 3] using OpenCLIP ViT-g/14.

Name-proximity is defined as $\cos\big(f_T(p), f_T(c)\big)$, a simple proxy for prompt explicitness. When interpreting PC-D values, positive scores suggest the output has drifted closer to the character name than to the original prompt—indicating the model's prior knowledge is exerting influence. Conversely, negative or near-zero values indicate the output remains aligned with the user's prompt. Since our experiments typically yield negative PC-D values, we refer to "larger PC-D" to mean less-negative scores, where the model's pull toward the character becomes stronger. The formulation is intended for interpretability rather than binary classification, providing creators with a transparent measure of how influence is distributed in the generative process.

Rather than stopping at a raw quantitative definition, we translate PC-D into a navigational aid for interpreting model behavior in context. We conceptualize this as a "responsibility compass"—a tool for orienting within human–AI co-creation. By mapping name-proximity against model drift, PC-D reveals four zones of agency: model-driven risk, mixed attribution, safe co-creation, and user-driven intent. The full diagnostic method runs entirely on local hardware with open-source tools, enabling accessibility for independent creators. Because we apply PC-D systematically across architectures and character archetypes, we expose model- and content-specific drift patterns that are invisible in single-model or single-domain audits.

Generative systems now mediate a large share of artistic production, creating the need for new ways to understand and guide their use. PC-D aims to shift the conversation from reactive moderation toward proactive understanding—providing an interpretable, creator-focused signal of semantic drift, and laying groundwork for tools that make influence patterns across characters and models transparent.

2 Methods

PC-D Metric. Prompt–Character Divergence (PC-D; Eq. 1) measures how strongly a generated image resembles a known character relative to the prompt that produced it. Definitions of f_I, f_T, and name-proximity are given in Section 1.

Computation. PC-D is calculated for each image and can be aggregated to reveal quadrant distributions and broader trends. All scoring runs locally on open-source tools. A vision–language model (GPT-4.1-mini) is used only as a baseline in the interpretability study.

Models. We evaluate PC-D on three open-source diffusion models—Playground-v2.5 [4], Stable Diffusion XL (SDXL) [5], and DeepFloyd-IF [6]—spanning different architectures and text encoders.

Character Set. The evaluation set contains ten widely recognized characters from games, animation, and comics (e.g., Mario, Pikachu, Tinkerbell). These were chosen for diverse archetypes, cultural familiarity, and high resemblance sensitivity, serving as a rigorous stress test for semantic drift.

Prompt Construction. For each character, we generated 20 prompts via:

1. **Natural language prompting:** GPT-o3-mini produced prompts from vague (e.g., "blue hedgehog running") to highly specific, clustered using k-means ($k = 10$) [7].
2. **Keyword prompting:** Following the COPYCAT framework [8], we sampled 5–20 high co-occurrence keywords from LAION-2B.

This produced 200 prompts per model and 600 images total. Examples across name-proximity ranges appear in Appendix F (Table 5).

Human Participants. All human studies (image annotation and interpretability pilot) involved unpaid volunteers who gave informed consent to participate in minimal-risk, anonymized tasks. No personal or sensitive data were collected. Study-specific protocols appear in Appendix E and G.

Annotation Protocol. Three annotators labeled each image for resemblance to the target character. Final labels were by majority vote. Agreement was high (e.g., $\kappa = 0.811$ for Playground-v2.5) [9].

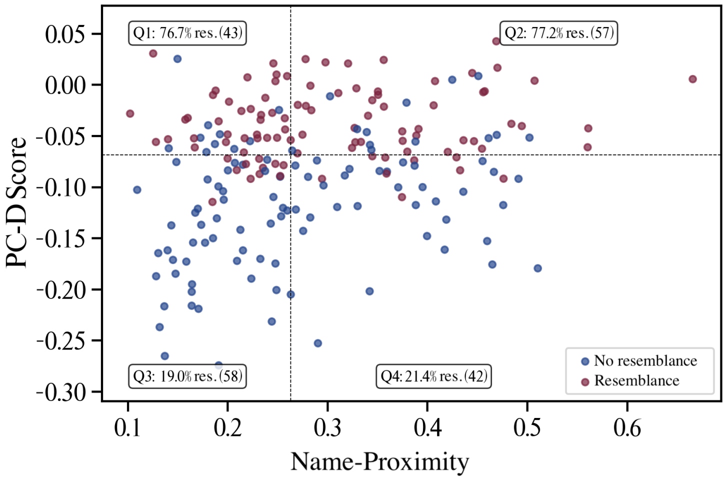


Figure 1: DeepFloyd-IF outputs in a 2D map: y-axis = PC-D (image–character minus image–prompt cosine; higher = more character-leaning), x-axis = name-proximity (prompt–character cosine). Dashed lines mark per-model medians to partition quadrants. Resemblance rates: Q1 (model-prior dominant/general) 76.7% (n=43); Q2 (prompt+model influence/explicit) 77.2% (n=57); Q3 (prompt-aligned/general) 19.0% (n=58); Q4 (prompt-dominant/explicit) 21.4% (n=42). Blue = no resemblance; maroon = resemblance.

Interpretability Study. Ten experienced generative tool users, including designers and artists, compared outputs using either (1) PC-D quadrant plots with examples (responsibility compass) or (2) VLM classification summaries. They rated interpretability, prompt–output clarity, and model selection support. VLM baselines used GPT-4.1-mini [10].

Methodological Note. PC-D is intended as a diagnostic signal of model drift towards replicating training data rather than a binary resemblance classifier. This focus emphasizes understanding the source of resemblance over categorical judgments. Appendix Figure 6 shows that PC-D aligns most strongly at extremes of annotator agreement.

Reproducibility and Anonymization. An anonymized repository[1] includes scripts, environment files, and partial data to reproduce core analyses including prompt and image generation and quadrant visualization. All materials allow anonymous access and contain no author identifiers.

3 Results

3.1 Interpreting PC-D: A Framework for Responsibility Attribution

Figure 1 illustrates the "responsibility compass" framework for DeepFloyd-IF outputs, providing an interpretable visualization of generative tendencies. For each model, thresholds were set at the median PC-D and median name-proximity across all prompt–character pairs, producing consistent quadrant boundaries within that model. Resemblance rates, as annotated by humans, differ substantially between quadrants. Quadrant plots for Playground-v2.5 and SDXL appear in Appendix Figures 4–5.

Q1 — Model-Driven Risk: High resemblance despite vague prompts, reflecting strong model bias.

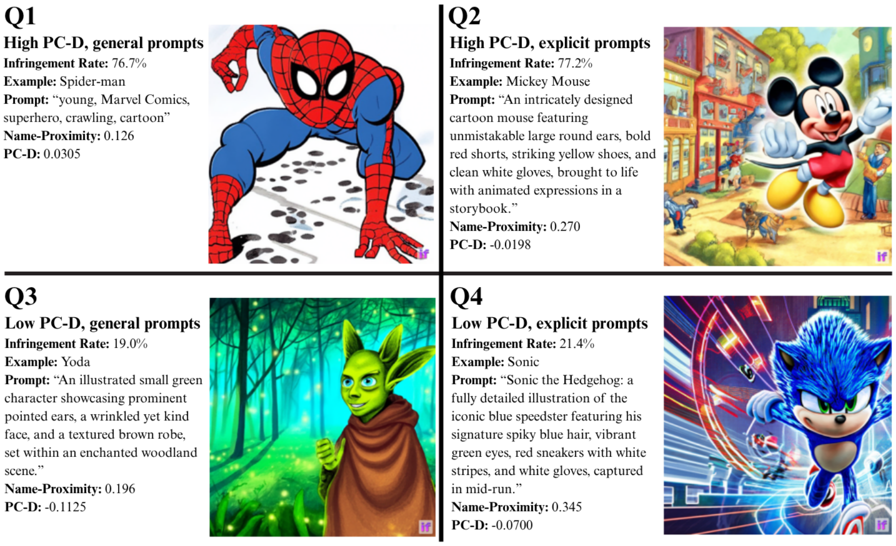
Q2 — Mixed Attribution: Explicit prompts plus model drift yield high resemblance.

Q3 — Safe Zone: General prompts with low drift show comparatively low resemblance rates.

Q4 — User-Driven Prompts: Explicit prompts that models sometimes resist overfitting.

These quadrants help distinguish the source of resemblance—user intent, model bias, or both. They also demonstrate that filtering on explicit prompts alone is insufficient: vague prompts can still produce outputs with strong, unintended resemblance to known visual identities if models

[1] `https://github.com/pcdanonymous/PC-Divergence`



Q1

High PC-D, general prompts

Infringement Rate: 76.7%
Example: Spider-man
Prompt: "young, Marvel Comics, superhero, crawling, cartoon"
Name-Proximity: 0.126
PC-D: 0.0305

Q2

High PC-D, explicit prompts

Infringement Rate: 77.2%
Example: Mickey Mouse
Prompt: "An intricately designed cartoon mouse featuring unmistakable large round ears, bold red shorts, striking yellow shoes, and clean white gloves, brought to life with animated expressions in a storybook."
Name-Proximity: 0.270
PC-D: -0.0198

Q3

Low PC-D, general prompts

Infringement Rate: 19.0%
Example: Yoda
Prompt: "An illustrated small green character showcasing prominent pointed ears, a wrinkled yet kind face, and a textured brown robe, set within an enchanted woodland scene."
Name-Proximity: 0.196
PC-D: -0.1125

Q4

Low PC-D, explicit prompts

Infringement Rate: 21.4%
Example: Sonic
Prompt: "Sonic the Hedgehog: a fully detailed illustration of the iconic blue speedster featuring his signature spiky blue hair, vibrant green eyes, red sneakers with white stripes, and white gloves, captured in mid-run."
Name-Proximity: 0.345
PC-D: -0.0700

Figure 2: Illustrative examples from each quadrant (DeepFloyd-IF). Each panel shows the prompt, generated image, name-proximity, and PC-D value. Quadrants are: Q1 model-prior dominant / general prompts, Q2 prompt + model influence / explicit prompts, Q3 prompt-aligned / general prompts, and Q4 prompt-dominant / explicit prompts.

overgenerate learned archetypes. Figure 2 shows the analysis grounded in specific (prompt, image output) pairs representative of each quadrant.

While PC-D aligns with human annotations, some high PC-D cases captured broad archetypes (e.g., superhero poses) without specific likeness as embedding similarity may conflate conceptual with distinctive resemblance. Additionally, narrative-style prompts often received lower specificity scores, even when describing notable characters, introducing a bias that can underestimate intent in natural-language prompts. This does not invalidate the framework but suggests that specificity should be interpreted cautiously in creative contexts.

3.2 Validation and Robustness Across Models

We computed Pearson correlations between PC-D scores and binary majority-vote resemblance labels across 200 generations per model. All models showed moderate, statistically significant correlations: Playground-v2.5 ($r = 0.65$, CI [0.58, 0.72]), SDXL ($r = 0.54$, CI [0.45, 0.62]), and DeepFloyd-IF ($r = 0.51$, CI [0.41, 0.60]), with 95% confidence intervals estimated via 1000 bootstraps. Despite architectural and training differences, PC-D captured consistent drift patterns and aligned moderately with human judgments. Appendix Figure 7 compares PC-D's correlations with human consensus against multiple VLM baselines across models.

We tested robustness by varying the PC-D threshold by ± 0.03. Q3 (Safe Zone) was most stable (3.8% std. dev. in resemblance rate), followed by Q4 (4.7%) and Q1 (5.2%). Q2 showed greater variability (6.7%), reflecting the inherent ambiguity of mixed human–AI influence.

3.3 Contextual comparison to a VLM classifier (for orientation, not for replacement)

We compared PC-D with a GPT-4.1-mini baseline under three prompting strategies: binary classification, calibrated binary, and continuous confidence scoring. GPT-4.1-mini achieved higher correlations with human judgments ($r = 0.61$–0.89) than PC-D ($r = 0.60$–0.68). This gap is expected: unlike classifiers, PC-D is not designed to maximize agreement with human resemblance ratings in all cases. Human labels capture any visual likeness—whether arising from explicit user intent or model bias—whereas PC-D aims to capture the more nuanced relationship between both user and model in co-creation. As a result, images that closely match a target character due to deliberate human prompting may score low on PC-D, even though annotators classify them as resembling. This

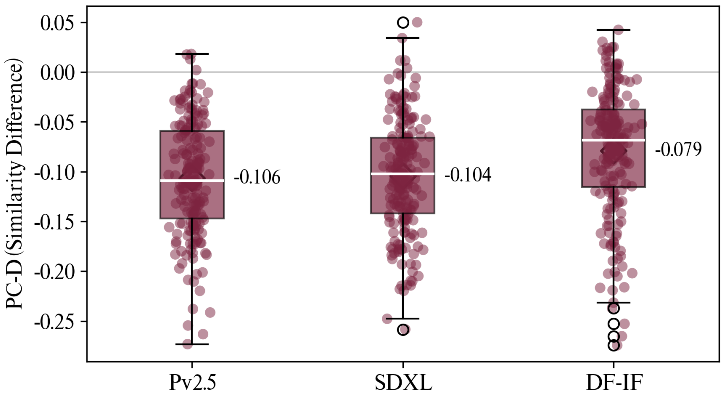


Figure 3: PC-D score distributions across Playground-v2.5, SDXL, and DeepFloyd-IF. Box plots show medians (white lines), interquartile ranges (boxes), and individual points. DeepFloyd-IF has the highest mean PC-D (-0.079) compared to Playground-v2.5 (-0.106) and SDXL (-0.104).

intentional divergence ensures that PC-D remains a diagnostic tool for isolating semantic drift rather than a general resemblance detector.

3.4 Creator Interpretability Study

In a pilot with ten experienced generative tool users—half with backgrounds in creative fields—PC-D scored higher than VLM classifiers on all interpretability measures: understanding model behavior (4.3 vs. 2.9), explaining prompt–output relationships (4.0 vs. 2.6), and informing model choice (4.1 vs. 2.5). Participants described PC-D as more explanatory and better suited for nuanced evaluation. One noted: “The compass made me think about how much was the model versus what the prompter wrote... The [VLM] just said what percentage matched, but I couldn’t tell why.” These responses underscore PC-D’s value in helping creators interpret and navigate generative systems (see Appendix G).

3.5 Practical Insights and Future Directions

Character analysis. Resemblance rates varied widely by character. Superheroes like Batman and Spider-Man showed the highest resemblance rates (76.7%) and PC-D scores (0.067–0.070), while Pikachu and Tinkerbell had much lower rates (20–22%; PC-D: -0.122 to -0.130). These differences were not explained by dataset frequency [11], suggesting that certain visual archetypes are inherently more prone to overgeneration. Detailed threshold performance for individual characters is provided in Appendix Tables 1–4.

Illustrative use case. A small studio fine-tuning a diffusion model could use PC-D to track drift over time. By auditing the same prompts across checkpoints, they could detect regressions—for example, vague prompts beginning to produce more human-labeled resemblances to protected characters—without relying on costly classifier APIs or exhaustive manual review.

4 Limitations and Future Work

PC-D’s main limitation is its reliance on semantic embeddings, which may blur visually distinct characters (e.g., Mario vs. Luigi). Name-proximity also favors keyword-heavy phrasing; richly descriptive prompts can score lower despite clearly describing a character. Our evaluation covers only ten characters and three open-source models, limiting generalizability to other domains, proprietary architectures, or less mainstream references. Finally, PC-D signals semantic drift but does not adjudicate copyright or intent.

We note that PC-D’s lightweight design is a deliberate choice: prioritizing transparency and local computability over algorithmic complexity enables creators, not just platforms, to audit model behavior on their own terms. Future work should integrate visually grounded embeddings, broaden character and model coverage, and assess PC-D in real creative workflows. While exploratory, this work contributes to a design space for creator-aligned interpretability—valuing transparency, contextual relevance, and human-centered reasoning alongside technical fidelity.

5 Discussion

5.1 Rethinking Creative Attribution in Human–AI Collaboration

PC-D offers a lightweight framework for analyzing the co-creative dynamics of generative systems. Instead of treating outputs as purely "human-created" or "AI-generated," it foregrounds shared—yet asymmetrical—agency. By operationalizing model bias as measurable drift, PC-D shifts creative auditing from binary filtering toward interpretive understanding. This distinction is particularly salient given growing evidence that diffusion models may unintentionally reproduce memorized content from their training data [12, 13].

For creators, it provides an interpretable signal of which visual concepts are most likely to be reproduced across architectures. For developers, it highlights regions where models over-rely on learned visual priors or exhibit poor generalization.

Unlike traditional approaches that treat resemblance monolithically, PC-D recognizes a fundamental duality in AI-mediated creation: outputs simultaneously reflect user intent and model priors, each requiring distinct consideration. The responsibility compass thus serves as an orientation tool rather than a normative arbiter, offering empirical grounding for decision-making while preserving the need for broader review.

5.2 From Centralized Oversight to Creator-Led Auditing

In a rapidly proliferating model ecosystem, platform-controlled assessments alone are insufficient. PC-D democratizes systematic oversight, bringing analytical capabilities previously reserved for large platforms within reach of independent creators. Users can inspect how prompts interact with training priors, detect unintended appropriation, and audit model behavior on their own terms. This capability also supports community-led oversight, where creative groups could share high-risk prompts, models, and characters, and collectively develop safer model-selection practices.

5.3 Extending PC-D Beyond Named Characters

Future iterations could incorporate reference-image anchoring, allowing creators to define resemblance boundaries via visual exemplars rather than named tokens. A supplied image—such as a specific frame, template, or artwork—could serve as a style or identity anchor, with PC-D measuring drift toward that reference relative to the prompt. This would extend the compass framework to stylistic influence, personal style protection, and creative self-auditing, regardless of whether the style appears in pretrained embeddings.

5.4 Future Directions

PC-D is part of an emerging class of tools that give creators direct insight into the systems they use. As human–AI collaboration continues to evolve, a central challenge is understanding how model behavior varies across an increasingly diverse landscape—each system shaped by its own architecture, training data, and embedded assumptions. Creative outcomes are not solely the product of human direction or a single, monolithic "AI," but the result of interactions between a model's learned priors and a user's intent. Recognizing and interpreting these dynamics will be central to creative practice as model diversity increases.

While its embedding-based approach has limitations, our compass framework shifts attention from asking *what* models generate to probing *how* and *why*. It supports systematic auditing across generative ecosystems without prescribing boundaries, offering orientation rather than verdicts. We position this work as an initial operational step toward a broader framework for creator-facing responsibility attribution in generative AI.

References

[1] Coalition for Content Provenance and Authenticity. C2PA technical specification v2.2. `https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.pdf`, 2025. Accessed 2025-08-09.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[4] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

[5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

[6] DeepFloyd. Deepfloyd/if. `https://github.com/deep-floyd/IF`, 2023. GitHub repository, version v1.0.1.

[7] OpenAI. Openai o3 and o4-mini system card. `https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf`, 2025. Accessed 2025-08-09.

[8] Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them. *arXiv preprint arXiv:2406.14526*, 2024.

[9] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[10] OpenAI. Gpt-4.1-mini: Model overview and api docs. `https://platform.openai.com/docs/models/gpt-4.1-mini`, 2025. Accessed 2025-08-09.

[11] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[12] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Ballé, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

[13] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *arXiv:2305.20086*, 2023.

Appendix

A Quadrant Analyses Across Models

In addition to DeepFloyd-IF (Figure 1), we include responsibility compass quadrant plots for Playground-v2.5 and SDXL. These illustrate the generalizability of PC-D across model architectures.

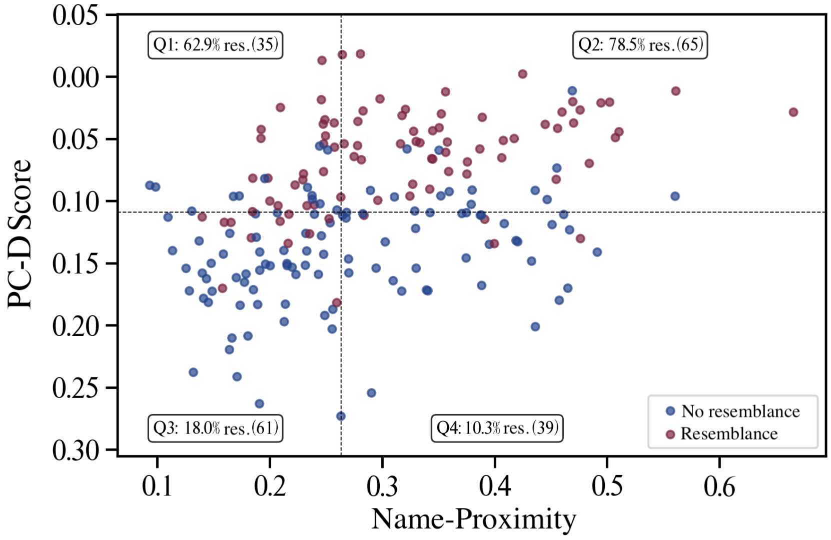


Figure 4: Quadrant plot of PC-D vs. name-proximity for Playground-v2.5. Q1 (model-prior dominant / general prompts): 62.9% human-labeled resemblance; Q2 (prompt + model influence / explicit prompts): 78.5%; Q3 (prompt-aligned / general prompts): 18.0%; Q4 (prompt-dominant / explicit prompts): 10.3%.

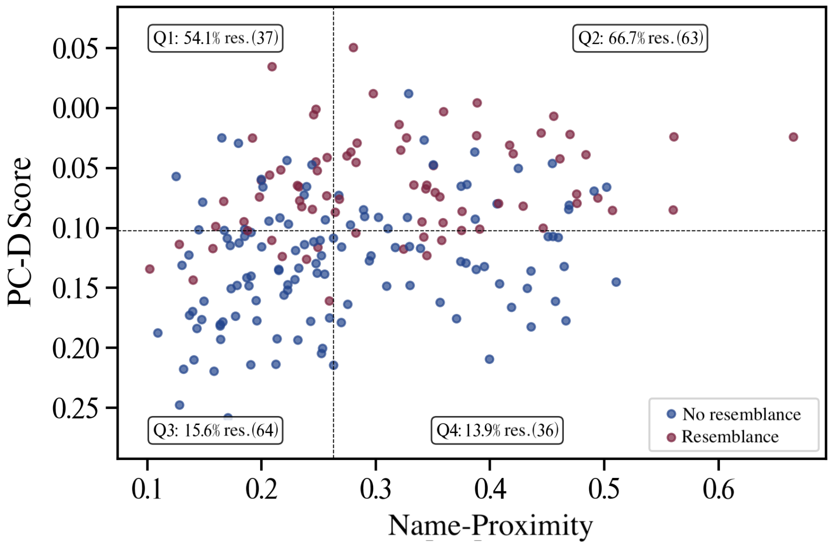


Figure 5: Quadrant plot of PC-D vs. name-proximity for SDXL. Q1: 54.1%, Q2: 66.7%, Q3: 15.6%, Q4: 13.9% human-labeled resemblance.

B PC-D vs. Human Consensus

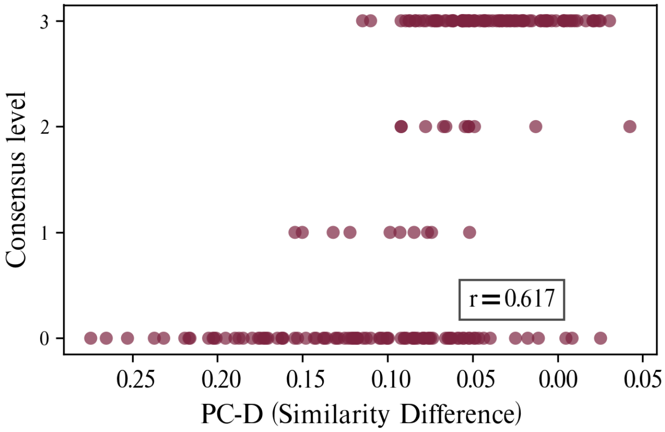


Figure 6: PC-D scores vs. human consensus (DeepFloyd-IF). A moderate positive Pearson correlation ($r = 0.617$) indicates alignment with human resemblance judgments while retaining independence from direct mimicry.

C Precision of Optimal PC-D Thresholds

Threshold performance tables for selected characters, showing the number of images above threshold, their human-labeled resemblance status, and resulting precision.

Table 1: Spider-Man PC-D threshold performance.

Model	Threshold	Above	Res.	No Res.	Precision
Pv2.5	−0.145	12	10	2	0.83
SDXL	−0.135	10	9	1	0.90
DeepFloyd-IF	−0.162	14	13	1	0.93

Table 2: Sonic the Hedgehog PC-D threshold performance.

Model	Threshold	Above	Res.	No Res.	Precision
Pv2.5	−0.122	9	7	2	0.78
SDXL	−0.090	8	6	2	0.75
DeepFloyd-IF	−0.090	10	8	2	0.80

Table 3: Pikachu PC-D threshold performance.

Model	Threshold	Above	Res.	No Res.	Precision
Pv2.5	−0.057	9	6	3	0.67
SDXL	−0.043	7	5	2	0.71
DeepFloyd-IF	−0.052	8	5	3	0.63

Table 4: Tinkerbell PC-D threshold performance.

Model	Threshold	Above	Res.	No Res.	Precision
Pv2.5	-0.058	10	7	3	0.70
SDXL	-0.080	8	6	2	0.75
DeepFloyd-IF	-0.075	9	7	2	0.78

D Correlation with Human Judgements

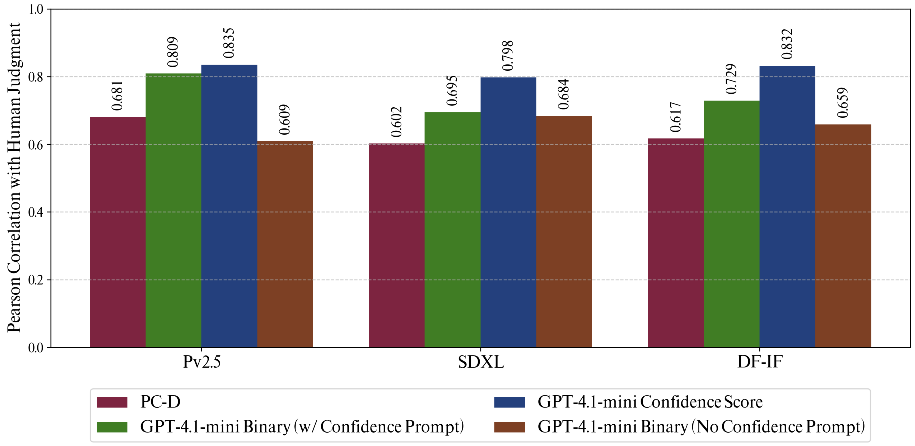


Figure 7: Pearson correlations between human consensus and four resemblance-detection methods across three models: PC-D (maroon), GPT-4.1-mini binary without confidence (brown), GPT-4.1-mini binary with confidence (green), and GPT-4.1-mini continuous confidence (blue). Confidence-aware GPT-4.1-mini achieved the highest alignment ($r \approx 0.84$–0.89), followed by binary with confidence ($r \approx 0.70$–0.81), binary without confidence ($r \approx 0.61$–0.68), and PC-D ($r \approx 0.60$–0.68).

E Annotation Protocol and Inter-Rater Reliability

Three annotators acted as proxies for an “ordinary observer” resemblance judgment. They completed a calibration phase with 5 primer images per character (2 clear resemblance, 2 clear no resemblance, 1 ambiguous) and were given neutral examples to illustrate borderline cases.

During evaluation, annotators saw only the generated image and assigned a binary label (0 = no resemblance, 1 = resemblance). They were blinded to model identity, prompt text, and PC-D scores, and worked independently. Final labels were determined by majority vote.

Inter-rater agreement was high: Playground-v2.5 ($\kappa = 0.811$), SDXL ($\kappa = 0.635$), DeepFloyd-IF ($\kappa = 0.860$). All $\kappa > 0.6$ indicate substantial agreement, validating label reliability.

F Name-Proximity Examples by Similarity Score Range

Table 5: Example prompts categorized by name-proximity levels.

Name-Proximity	Example Prompt	Character	Score
Low Proxy			
	A detailed illustration of a mustachioed plumber wearing a red cap with an emblem, blue overalls, and a matching red shirt, set in a vibrant video game scene with pipes, coins, and whimsical brick blocks.	Mario	0.057
	A detailed cartoon portrait of a diminutive green sage with large pointed ears, deep facial creases, and a flowing brown robe, set amid ancient ruins and mystical foliage.	Yoda	0.100
	An energetic illustration of a cartoon sea sponge with a square yellow body, white shirt, red tie, and brown pants, accented by big blue eyes and a joyful smile, against an underwater backdrop.	Spongebob Squarepants	0.164
	An expressive digital painting of a yellow cartoon figure with dark-tipped pointed ears, red cheeks, and a lightning bolt tail, set against a dynamic urban backdrop.	Pikachu	0.178
	A dynamic portrait of a regal character with braided platinum hair and a light-blue gown featuring frost patterns, set in a magical winter forest with drifting snow.	Elsa	0.188
High Proxy			
	Disney, Fairy, Peter Pan, wings, light, cute, pixie dust, Neverland, magical, blonde, fairy tale, animated, fantasy, tiny, flight, bright, green dress, sassy, enchanting, glowing, cartoon	Tinkerbell	0.605
	spider-sense, web-shooter, stealth, friendly neighborhood, wall-crawler, swinging, mask, symbol, superhero, costume	Spider man	0.609
	Star Wars, The Mandalorian, The Child, Disney, Cute, meme, Grogu, adorable, green, soft, merchandise, small, robe, alien, popular, sci-fi, species, puppet, viral, miniature, cartoon	Yoda	0.625
	magical, fairy tale, Neverland, Disney, blonde, tiny, wings, enchanting, fantasy, cartoon, sassy, Fairy, animated, cute, glowing	Tinkerbell	0.627
	dark, batmobile, black, Batarang, night	Batman	0.644
	Genesis, zone, Sega, cool, fast, Sonic boom, tails, rings, running, collectible	Sonic the Hedgehog	0.673

G Interpretability Study Protocol

We conducted a small interpretability study to evaluate whether PC-D supports creative users in reasoning about model behavior, understanding prompt–output dynamics, and informing model selection. The study included 10 non-author participants who regularly use generative image tools such as Midjourney, DALL·E, or Stable Diffusion, 5 of which are artists and designers in the creative space.

Study Conditions. Participants were randomly assigned to one of two conditions:

1. **PC-D condition:** Four generated images from DeepFloyd-IF, each labeled with its prompt, PC-D score, name-proximity, and quadrant classification (e.g., "model-prior dominant / general prompts"), plus quadrant plots for all three models.
2. **VLM condition:** The same images with only binary human-labeled resemblance annotations (0/1) from GPT-4.1-mini, plus a summary table of resemblance rates per model.

Interpretability Measures. Participants then answered three 5-point Likert-scale questions:

- **Model Behavior Understanding:** "How much did this feedback help you understand how the model behaves or generalizes?"
- **Prompt–Output Relationship Clarity:** "How well did this help you understand the relationship between your prompt and the model's output?"
- **Influence on Model Choice:** "Would this kind of feedback influence which model you choose for future creative work?"

They also provided qualitative feedback.

Results. PC-D outperformed the VLM classifier across all interpretability metrics: 4.3 vs. 2.9 (model understanding), 4.0 vs. 2.6 (prompt–output clarity), and 4.1 vs. 2.5 (model selection). Participants described PC-D as "explanatory," "interesting," and "informative." One noted: "The quadrant map made me think about how much was the model versus what the prompter wrote... It made me think more about training differences."

Limitations. This was a small pilot for early insights. Participants were unpaid volunteers, and although half work or study in creative fields, the sample size and recruitment method limit generalizability. Nonetheless, these preliminary findings suggest PC-D offers a more transparent and interpretable framework for creators evaluating model behavior.

H Example Images

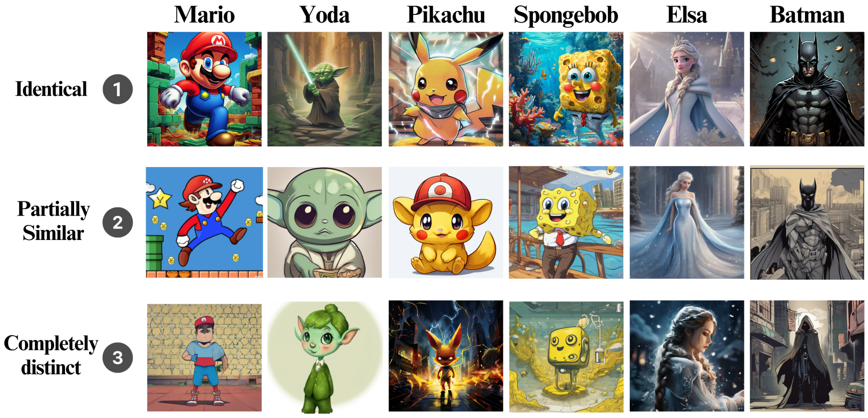


Figure 8: Outputs from three text-to-image models (Playground-v2.5, SDXL, and DeepFloyd-IF) for six characters (Mario, Yoda, Pikachu, Spongebob, Elsa, Batman) organized by human-labeled resemblance: row 1 – unanimous resemblance; row 2 – partial resemblance (at least one annotator flagged resemblance); row 3 – no resemblance (no annotators flagged resemblance).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly present the main contribution—Prompt–Character Divergence (PC-D)—and its purpose as an interpretable metric for evaluating semantic drift in generative image models. These claims are substantiated through both empirical and methodological sections (Sections 2 and 3).

2. **Limitations**

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

   Answer: [Yes]

   Justification: Sections 3.3 and 5 discuss PC-D's limitations, including its reliance on semantic embeddings that may conflate conceptual similarity with visual distinctiveness, name-proximity bias toward keyword-heavy prompts, limited evaluation scope (ten characters, three open-source models), and the fact that it does not provide definitive legal judgments.

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: The paper does not include formal theoretical results or proofs. It presents a metric and empirical framework rather than deriving new theorems.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper?

   Answer: [Yes]

   Justification: Section 2 and the Appendix describe the experimental setup, including models, datasets, generation process, and PC-D computation. All code is included in the supplementary material.

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

   Question: Does the paper provide open access to the data and code?

   Answer: [Yes]

   Justification: An anonymized GitHub repository contains scripts and partial data sufficient to reproduce the primary figures and experiments.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details?

   Answer: [Yes]

   Justification: Section 2 outlines how prompts were created and images were generated, including model architectures, prompt clustering, and evaluation procedures.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or statistical significance?

   Answer: [Yes]

   Justification: Pearson correlations are reported with bootstrapped 95% confidence intervals (Section 3.2). Error bars are included in figures where appropriate, supporting the statistical validity of results.

8. **Experiments compute resources**

   Question: Does the paper provide sufficient information on the compute resources?

Answer: [Yes]

Justification: Section 2 (Computation) specifies that all experiments were run locally on consumer hardware with open-source tools. Processing 600 images required only minutes and incurred no commercial compute cost.

9. **Code of ethics**

   Question: Does the research conform with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The research adheres to ethical guidelines, uses publicly available models, and aims to empower creators with transparency tools.

10. **Broader impacts**

    Question: Does the paper discuss both positive and negative societal impacts?

    Answer: [Yes]

    Justification: Section 4 discusses both positive impacts (democratizing model auditing, supporting creator agency) and potential negative impacts (embedding bias, overreliance on semantic similarity metrics), addressing societal implications from multiple perspectives.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release?

    Answer: [NA]

    Justification: The paper does not release new models or datasets with high misuse risk. All tools released are local, transparent, and open-source.

12. **Licenses for existing assets**

    Question: Are existing assets properly credited and licensed?

    Answer: [Yes]

    Justification: All models and tools used are open-source and properly cited in the references (e.g., DeepFloyd-IF, Playground-v2.5, SDXL, OpenCLIP).

13. **New assets**

    Question: Are new assets well documented?

    Answer: [Yes]

    Justification: While no new datasets or models are released, the codebase introduced is documented and included in supplementary materials.

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

    Question: For human subject research, does the paper include instructions and compensation info?

    Answer: [Yes]

    Justification: The paper describes both the annotation task and the interpretability pilot study, including participant recruitment, informed consent, and majority voting protocol. All participants were unpaid volunteers in minimal-risk tasks of under one hour, with no collection of personal or sensitive data. Full protocols for both studies appear in the Appendix.

15. **Institutional review board (IRB) approvals**

    Question: Does the paper describe potential risks and IRB approval?

    Answer: [NA]

    Justification: The work involved no personal data, sensitive topics, or identifiable human information; therefore, IRB review was not required.

16. **Declaration of LLM usage**

    Question: Does the paper describe LLM usage?

    Answer: [Yes]

    Justification: Section 2 notes that GPT-o3-mini was used to generate prompt variations and GPT-4.1-mini as a baseline in the interpretability study. These uses are disclosed and do not constitute core components of the proposed method.