Spot the Pipeline: Evaluating Perceptual Convergence Among Text-to-Image Tools

Martin Disley
Institute for Design Informatics
University of Edinburgh
m.disley@ed.ac.uk

Abstract

The open text-to-image ecosystem is often treated as a pluralising counterweight to the homogenising tendencies of foundation models, since it offers users a wealth of checkpoints and variants that appear to promise stylistic diversity. This short paper asks whether that apparent diversity is perceptually legible in practice. This short paper presents a pilot study in which 15 power users completed a forced choice identification task of images generated by five popular text-to-image tools, using stylistically underspecified prompts. The central result is that recognisability was unevenly distributed: the most experienced users could separate architecturally distinct models, but even for them popular variants of the Stable Diffusion family were not separable. Taking perceptual identification as a proxy for aesthetic distinction, this pattern complicates the assumption that a proliferation of variants fine-tuned by end users straightforwardly translates into meaningful stylistic diversity across tools.

Introduction

The contemporary text-to-image landscape presents users with an abundance of checkpoints, model variants, and adaptors, especially around open base model families such as Stable Diffusion, through repositories such as Hugging Face, CivitAI, and Replicate. This abundance implies stylistic diversity: if one model produces stale or overfamiliar imagery, another should offer a meaningfully different visual character. The abundance of named options, however, does not necessarily translate into an abundance of distinct outputs. For an ecosystem like this to flourish, the available options should provide meaningful, or at least perceptibly different, aesthetic territory under ordinary use.

If the proliferation of checkpoints is understood to reflect a diverse generative image culture, then the key empirical issue is whether different tools are actually distinguishable in practice. This short paper proposes users' consistent ability to recognise or separate images from various models as a minimal proxy for meaningful divergence in visual character. For text-to-image systems, this problem is complicated by the fact that sophisticated prompting can push a model far from its defaults. This study uses deliberately underspecified prompts as a controlled probe of each tool's house style: the characteristic visual tendencies that become visible when users provide little explicit stylistic steering. The question, then, is whether the apparent abundance of available tools yields perceptually distinct options at the point of use.

This question is addressed here through a pilot study in which power users of ComfyUI, a node-based image making software, attempted to identify which of five text-to-image pipelines popular on the platform had generated a given image. The study uses this task to ask whether recognisability is evenly distributed across the sampled tools, or whether apparent abundance masks structured perceptual convergence. We interpret the resulting pattern as evidence that model abundance should not be taken as a proxy for stylistic diversity across tools without examining whether those tools are perceptually distinct in practice.

Related Work

Questions of diversity and convergence have become increasingly important in computational creativity and adjacent discussions of generative AI. At a broad level, recent work has challenged the assumption that more generated content necessarily means more meaningful variation, instead suggesting that generative systems can stabilise conventions at scale (Doshi and Hauser 2024; Padmakumar and He 2024; Anderson, Shah, and Kreminski 2024; Sarkar 2024). In text-to-image systems, this concern is visible both in critical accounts of homogenised visual culture and in analyses of prompt corpora that show repeated aesthetic patterns and conventions emerging in practice (McCormack et al. 2024). This paper takes this broader concern as background, but narrows it to a more specific ecosystem question: whether the abundance of checkpoints and model variants actually yields perceptibly different outputs.

In text-to-image systems specifically, prompting is iterative and skilled rather than transparent or one-shot (Oppenlaender 2022; Ibarrola and Grace 2023; Lichtenberg and Akdag 2025). This matters because heavily engineered prompts can push a system far from its defaults, whereas underspecified prompts expose the output tendencies users inherit when they do not actively steer. Despite this, analyses of prompt corpora suggests that repeated aesthetic conventions emerge in practice (McCormack et al. 2024), making default behaviour a meaningful object of study rather than a trivial baseline.

Within computational creativity, related work has there-

fore focused on how users navigate text-to-image systems, how style can be framed within creative applications, and how diversity can be evaluated within a single system rather than across named tools (Gajula et al. 2024; Ibarrola and Grace 2024). Adjacent work on CLIP embedding spaces and stylistic similarity helps us reason about images and prompts (Radford et al. 2021; Somepalli et al. 2024), while workflow studies show that generative image tools can differ in consistency, controllability, and refinement behaviour in practice (Ocampo Blanco and Bown 2024). Yet this leaves underexamined a broader ecosystem-level assumption: that limited stylistic range within any one model may be unproblematic if a plurality of aesthetic options exists across the wider field of tools. Concerned with evaluating ecosystem stylistic diversity, this pilot study investigates cross-model perceptual distinctiveness using underspecified prompting as a 'house style' probe.

Method

Here, this question is addressed with a forced-choice identification study deployed online as *Spot the Pipeline*. In each of the scored trials, participants were shown a single generated image and asked to identify which of five named text-to-image tools had produced it. The prompt text was visible on each trial, and participants could view the principal run parameters for the sample, but no example images or training phase were provided before the task began. The task comprised 25 scored trials, constructed from five underspecified prompts crossed with five tools, plus one duplicate-image attention check. After completing these scored trials participants complete a short experience questionnaire, which was required to reveal their score. Performance was evaluated against a five-way chance baseline of 20%.

The roster curation was informed by model and checkpoint popularity as given from benchmarks/leaderboards and observed conversations on forums and chat servers. It combined two architecturally distinct non-Stable-Diffusion tools with three models from the Stable-Diffusion family: FLUX.1-schnell, Z-Image Turbo, Stable Diffusion XL base, JuggernautXL v9, and RealisticVision v5.1. The three Stable-Diffusion variants represent related but distinct parts of the SD ecosystem: SDXL base is the official SDXL foundation model, JuggernautXL v9 is a community SDXL-family checkpoint, and RealisticVision v5.1 is an SD1.5-family realism checkpoint run with its recommended VAE, negative prompt, and hires/upscale pass. This mix allowed us to compare two non-SD tools with a set of widely used SD-derived pipelines distributed as distinct options in practice. Importantly, these were treated as tools-as-accessed rather than models as isolated artefacts: each tool was run using a fixed preset intended to reflect typical community use. In practice, this means that the study compares user-facing pipelines, not only abstract base-model behaviour.

Stimuli consisted of 25 images generated from five deliberately underspecified prompts crossed with the five tools. The prompts were: *a portrait of a person*, *a cosy living room interior*, *a city street at night, rain*, *a still life of a bowl of fruit on a table*, and *a small robot on a desk*. Fixed seeds were used within each prompt condition so that the sample

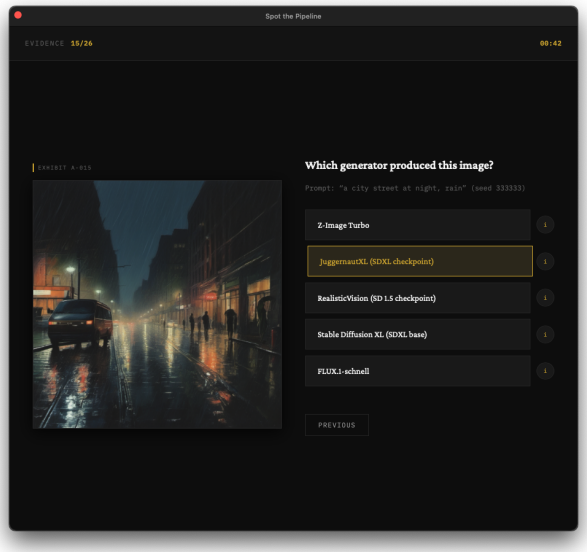

Figure 1: Spot the Pipeline interface: Participants viewed one generated image at a time and selected which of five named text-to-image tools they believed had produced it. The prompt text was shown, and principal run parameters could be viewed by clicking the "info" button

set was defined procedurally and comparison across tools was controlled at the level of prompt and stochastic initialisation.

Participants were recruited from ComfyUI communities on the messaging platform Discord. This power user sampling strategy was informed by an informal pre-study pilot with colleagues who had little or no regular experience using text-to-image systems. Those pilot versions tested whether the task was viable for inexperienced participants under both named-tool and less explicitly labelled versions of the quiz. In both cases, inexperienced participants performed at or below chance, suggesting that the task required more practical familiarity with contemporary text-to-image tools than a general audience could be assumed to have. We therefore treated the final deployment as a power-user baseline: a test of whether people with active practical exposure to current text-to-image workflows could detect tool-specific output signatures without a training phase, with the assumption that users with less experience would not have this capacity. The principal data points measured were overall accuracy relative to chance, per-tool identifiability, confusion patterns across tools, and associations between performance and self-reported experience.

Thirty submissions were recorded, of which 15 contained complete responses and were retained for analysis. Attrition occurred mostly at the start of the quiz - the incomplete submissions recorded a mean of 3.6 scored responses. Among the incomplete submissions, pooled partial accuracy was 18.9%, approximately at the 20% chance level. Like the novice users who attempted the informal pilot study, the low scores of the dropout suggest that they found the task ex-

ceedingly difficult. This filtering likely selected for participants who were more confident in their capacity to succeed in the task.

Results

Participants identified tools significantly above chance overall. Across the 15 completed submissions, mean accuracy was 30.7% against a chance baseline of 20%, corresponding to an average score of 7.67 out of 25. A one-sample test against chance showed that this effect was statistically significant ($t(14) = 2.87$, $p = .006$, with Cohen's $d = 0.74$). This indicates that some signal of tool identity was perceptible in the outputs by this group of power users without participants any explicit training phase.

This signal was not evenly distributed across the five tools. FLUX.1-schnell and Z-Image Turbo were the most recognisable, with accuracies of 48.0% and 45.3% respectively, both well above the 20% chance level. By contrast, the Stable-Diffusion-family cluster was only weakly separable: SDXL base was identified at 24.0%, JuggernautXL at 22.7%, and RealisticVision at 12.0%, the latter falling below chance. The principal empirical implication of this is therefore not simply that power-users can identify tools, but that recognisability is uneven and concentrated in a subset of them.

The confusion matrix shows in Figure 2 shows that this pattern is structured rather than random. FLUX and Z-Image each maintained relatively clear identities, whereas SDXL, Juggernaut, and RealisticVision were systematically confused with one another. The strongest off-diagonal confusion was RealisticVision being identified as Juggernaut (28%), which occurred more often than RealisticVision being correctly identified at all (12%). Juggernaut, in turn, was most often misidentified as Z-Image (27%). Taken together, these confusions indicate that the key result is an Stable Diffusion-family cluster whose members do not function as clearly distinct perceptual options under the conditions tested.

Experience also mattered, but in a specific way. Accuracy was positively associated with hours per week spent generating images (Spearman $p = 0.62$, $p = .015$), whereas experience with text-to-image tools was not significantly associated with performance ($p = 0.22$, $p = .43$). This suggests that the task rewards active, current engagement with the latest suite of models, more than simple accumulated exposure to image making with generative AI/ML. Some prompts made the tools easier to distinguish: portrait images were the strongest discriminator, whereas the still-life prompt was effectively at chance. This indicates either that some prompt types surface default output signatures more effectively than others or that users are more familiar with portrait images and are better able to identify the output signature in those images. This remains an open question and a potential avenue for further work.

Discussion

The significance of these results is not primarily that individual participants can identify some models above chance.

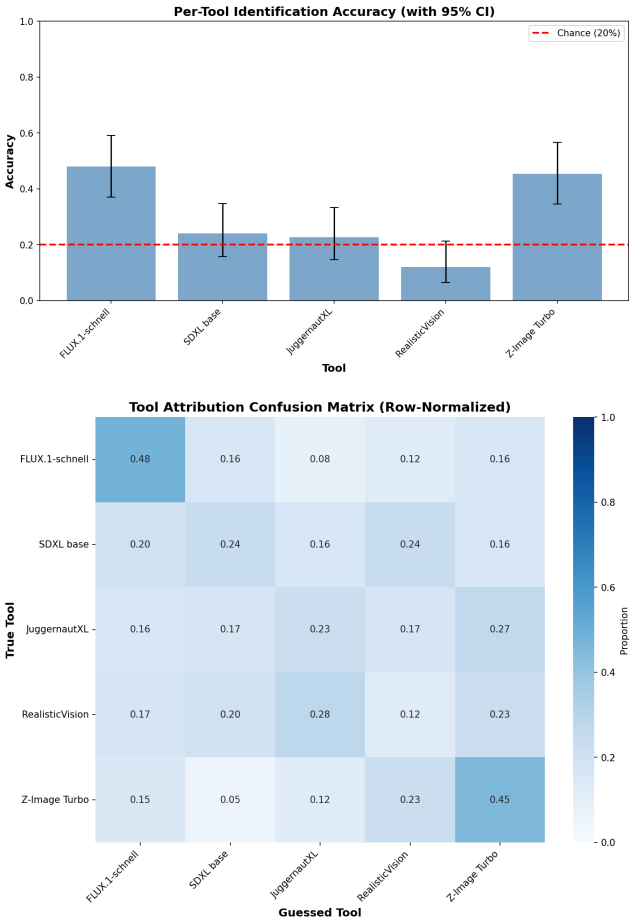

Figure 2: Perceptual recognisability of five text-to-image tools under default, underspecified prompting conditions. Top: per-tool identification accuracy, with chance performance at 20%. Bottom: confusion matrix showing that recognisability was unevenly distributed: FLUX.1-schnell and Z-Image Turbo were relatively distinct, whereas SDXL, JuggernautXL, and RealisticVision formed a structured confusion cluster.

Rather, the study uses recognisability as an ecosystem-level probe. Open and community-driven generative-image ecosystems are often positioned as a pluralising counterweight to the homogenising tendencies of centralised foundation models. That promise depends, however, on whether the proliferation of named checkpoints and pipelines actually presents perceptibly diverse options in practice. Our results complicate this assumption: while FLUX.1-schnell and Z-Image Turbo were relatively recognisable, the Stable-Diffusion-derived pipelines formed a structured confusion cluster. This suggests that checkpoint proliferation alone should not be equated with meaningful stylistic plurality.

More precisely, this study should not be read as a pure test of style in a narrow art-historical sense. It measures the perceptual distinctiveness of outputs generated under stylistically underspecified prompting conditions, which is closer

to perceptual recognisability than to style attribution alone. The term *output signature* is useful here because it captures the broader bundle of cues participants may have been using: rendering tendencies, composition, fidelity, artefacting, prompt adherence, and other pipeline specific behaviours as well as style. This studies results supports a narrower claim: that the house styles or output signature of some models are distinct and perceivable by the most experienced users in images generated without specific style guidance, while others models are not, and thus suggests that these house style converge.

Given this, we would caution interpreting the recognisability of FLUX and Z-Image as indicative of them being stylistically distinct, in the art-historical sense. Differences in model architecture, scale, and optimisation target mean the separability observed here may partly reflect quality-related or optimisation-related cues rather than stylistic difference alone. Differences in fidelity, or characteristic artefacts could all contribute to a recognisable output signature. This is not necessarily a weakness of the study. Because users encounter tools-as-accessed rather than purified latent style manifolds, a user-facing evaluation should remain sensitive to the full bundle of perceptible properties through which one tool comes to feel different from another.

Additionally, a lack of ecosystem stylistic diversity would not imply that community checkpoints or fine-tuned variants offer no value. A checkpoint may be useful because it is more reliable on certain prompt classes, easier to work with, cheaper to run, faster, or better integrated into an existing workflow. It may also help users reach a desired stylistic output with less friction, even if its default outputs are not cleanly separable in a blind identification task. The present result is therefore narrower: within the conditions tested here, these variants did not present as strongly distinct perceptual options. That is a meaningful qualification for ecosystems in which checkpoints are frequently marketed or selected as though each one opens a clearly different visual regime.

For computational creativity research, the implication is that model abundance is a weak proxy for meaningful diversity. The next step should move beyond further perceptual tests and take a dual approach. First, directly measuring image diversity across much larger samples of images and prompts would help quantify image diversity. Second, a qualitative study engaging practitioners in questions regarding tool choice would help us understand the broader ecosystem value of model plurality, beyond stylistic diversity.

Conclusion

This pilot study shows that, under underspecified prompting, active and experienced power users can perceptually recognise some text-to-image tools, and that many apparently distinct variants are not recognisable. In our sample, architecturally distinct models were distinguishable, and Stable Diffusion family variants formed a confusable cluster. The significance of this finding concerns ecosystem diversity rather than individual capacity. Open and community driven model ecosystems are often treated as a pluralising counterweight to the homogenising tendencies of centralised foundation models. That promise depends on whether proliferating checkpoints and pipelines produce perceptibly distinct creative options in practice. Our results suggest that the abundance of named checkpoints does not automatically provide meaningful stylistic diversity. For computational creativity, this means that the proliferation of open text-to-image models should not be assumed to mitigate homogeneity in generative AI media without evidence that they produce genuinely distinct outputs.

References

Anderson, B. R.; Shah, J. H.; and Kreminski, M. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Creativity and Cognition*, 413–425. Chicago IL USA: ACM.

Doshi, A. R., and Hauser, O. P. 2024. Generative artificial intelligence enhances creativity but reduces the diversity of novel content.

Gajula, P.; Abhishek Dangeti; Vivek Srivastava; and Vikram Jamwal. 2024. Style-Frame: A Foundational Framework for Artistic Style Driven Applications. In Grace, K.; Llano, M. T.; Martins, P.; Hedblom, M. M.; Cardoso, A.; Association for Computational Creativity; and Jönköping University., eds., *Proceedings of the Fifteenth International Conference on Computational Creativity: ICCC'24, Jönköping, Sweden, 17-21 June*. Coimbra: Association for Computational Creativity (ACC).

Ibarrola, F., and Grace, K. 2023. Prompt diversification for iterating with text-to-image models. In Pease, A.; Cunha, J. M.; Ackerman, M.; Brown, D. G.; and Association for Computational Creativity., eds., *Proceedings of the Fourteenth International Conference on Computational Creativity: ICCC'23, Ontario, Canada, 19-23 June*. Coimbra: Association for Computational Creativity (ACC).

Ibarrola, F., and Grace, K. 2024. Measuring Diversity in Co-creative Image Generation. In *Proceedings of the 15th International Conference on Computational Creativity*.

Lichtenberg, S., and Akdag, A. 2025. A Creativity Assessment Scale for Text-to-Image Prompting: Challenges & Observations. In Oliveira, H. G.; Spendlove, B.; Gervás, P.; and Ventura, D., eds., *Proceedings of the Sixteenth International Conference on Computational Creativity: ICCC25*.

McCormack, J.; Llano, M. T.; Krol, S. J.; and Rajcic, N. 2024. No Longer Trending on Artstation: Prompt Analysis of Generative AI Art. In Johnson, C.; Rebelo, S. M.; and Santos, I., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 279–295. Cham: Springer Nature Switzerland.

Ocampo Blanco, R., and Bown, O. 2024. Integrating Generative AI into Creative Workflows: Dealing with Consistency, Scene Control, and Refinement in a Professional Image Generation Case Study. In Grace, K.; Llano, M. T.; Martins, P.; Hedblom, M. M.; Cardoso, A.; Association for Computational Creativity; and Jönköping University., eds., *Proceedings of the Fifteenth International Conference on Computational Creativity: ICCC'24, Jönköping, Sweden, 17-21 June*. Coimbra: Association for Computational Creativity (ACC).

Oppenlaender, J. 2022. The Creativity of Text-to-Image Generation. In Hedblom, M. M., ed., *Proceedings of the Thirteenth International Conference on Computational Creativity: ICCC'22*. Coimbra: Association for Computational Creativity (ACC).

Padmakumar, V., and He, H. 2024. Does Writing with Language Models Reduce Content Diversity?

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Sarkar, A. 2024. Intention Is All You Need.

Somepalli, G.; Gupta, A.; Gupta, K.; Palta, S.; Goldblum, M.; Geiping, J.; Shrivastava, A.; and Goldstein, T. 2024. Measuring Style Similarity in Diffusion Models.