Jasmine Poon - TM Interactivity Grant Proposal

📢 **Principal Investigator(s):** Jasmine Poon

- **Primary contact email:** jasminejpoon@gmail.com
- **Location:** New York City, USA
- **Date:** June 19, 2026
- **Organization (if any):** Independent researcher
- **Tax ID (if applicable):** N/A

Discovery as continuous interaction: image generation controls synthesized from multimodal references

1.1 Abstract

Creation is a loop: you make something, see it, and discover what you wanted by reacting to what you got (Schön, 1983). Generative image models collapse that loop into discrete prompt-and-reroll turns across a lossy channel. The problem is not that intent is unclear, but that language is imprecise: a user can recognize the right result on sight yet not put the deciding quality into words, and reading what the model changed is just as hard. These are the gulfs of execution and evaluation (Hutchins, Hollan & Norman, 1985; Park et al., 2026), widest when the target is sub-verbal. This project's interaction mode is discovery: converting taste into an actionable preference on top of articulation, inside a continuous loop. From a user's reference images, the system extracts a shared quality and exposes it as a continuous control, an “adjective”, analogously, learned from images rather than stated in language, which the user adjusts until they recognize the result. Decomposing every image into look and content (via the generator's own JSON description) makes the change interpretable, addressing the evaluation gulf. We evaluate against the same user working with full words, skill, and tools, measuring whether discovery reaches that result and whether the reference's look transfers without its content leaking. I executed a one-week pilot (~$25 compute), whereby the mechanism, a coherent style control extracted from an image model, isolated the open problem: can we aim the control at the user's target rather than the model's default style, and hold it constant across subjects? This study proposes an interaction mode where we convert taste into actionable preference by discovery on top of articulation.

1.2 Problem & Motivation

The gap between skill and articulation is evident when everyone has become a “vibe-coder” as AI models increase in intelligence and capabilities beyond coding and software engineering tasks. This research aims to help regular people who use AI, understand what’s possible to alter, or create with AI by giving them palpable directions (synthesized from both VLM and textual information, i.e. prompts) during image creation, alongside a model’s latent space and steerable activations.

Steering the latent space with isolated possible directions yields a more successful creation loop (see Schön, 1983); because the control over the derivative, e.g. “more of this soft lighting”, allows for recognition through steering, or continuous interaction, which is important for arriving at an indeterminate preference (see Park et al.), for something so objective as taste and style.

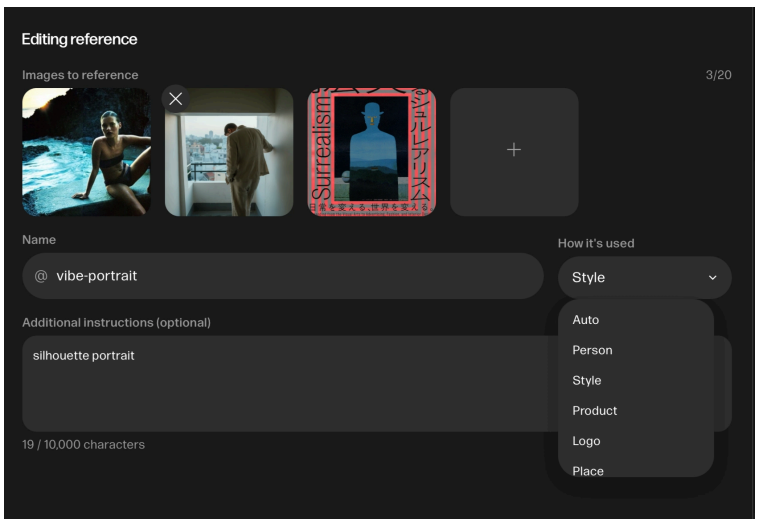
I argue for converting taste into preference by tackling the articulation gap, with the key points being a 2-part mechanism:

1. Discovery as the interaction mode, to transfer taste onto creation without semantic articulation (expert vocabulary), i.e. having the right words;
2. Synthesis of the bounded discovery space from a combination of multimodal references and AI understanding of the target output.

The current trend of vibe-coded ai-enabled creations consists of custom interfaces, toggles and sliders for 3D shader art, see “Generative Remix” (author’s own), “Mesh Gradient Tool” [https://v1.cloan.me/0.7513130781587641] by Chase Davis, a physics generator “Tutti Space Program” by Grant Kot [https://tsp.grantkot.com/].

The current state of image generation tools offer two approaches, both requiring specification.

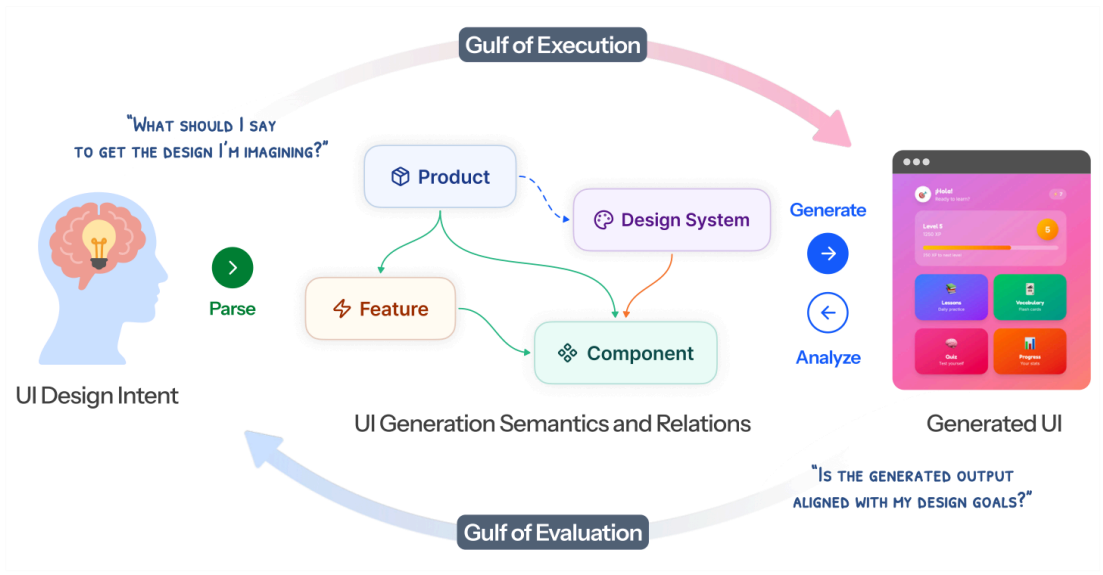
1. Structural: Canva's Magic Layers (March 2026) decomposes a flat image into separate, nameable, editable parts.
2. Reference-by-example: Reve's References carries an example's attributes into new work, but the user must assign each reference a fixed role.
    - The "how it's used" menu offers Auto, Person, Style, Product, Logo, Place, but users report that the result tends to copy the reference's content.



- A Reve user (post retweeted by Reve, 5 June 2026) reports it "reproduce[s] my references too literally and made collages of them rather than applying qualities, like the lighting, composition, textures, or style," and asks to "say why a reference matters and what should be extracted from it."

Structural tools (Canva's Magic Layers, March 2026) decompose an image into nameable parts. Explorable continuous control exists today only as hand-built, single-purpose interfaces (e.g., Generative Remix [author's own]; Mesh Gradient Tool; Tutti Space Program) — controllable because a human authored each one's knobs. General image models ship none — only a prompt and a seed.

The gap. Recent work bridges these gulfs with *nameable semantics* as an intermediate representation between intent and output (Park et al., 2026). We target the residue that semantics cannot name. The research question: *can a shared quality be extracted from reference images as a continuous control, and can a user reach a target by adjusting it?* This matters more as models improve — the output space grows faster than anyone can specify a point in it, so navigation by recognition is what keeps human control in step with the model.



Source: Park et al., "Bridging gulfs of execution and evaluation in generative UI design".

1.3 Research Direction Addressed

- Evaluations for real-time multimodal interaction
- Explaining complex information with generative UI *(controls generated per reference set, not hand-built)*
- Steering agents during long-horizon tasks
- **Other:** Discovery — an interaction mode that turns references into continuous controls steered by recognition, augmenting on top of natural language prompting

1.4 Proposed Approach / Method

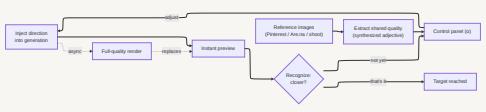
There are two components and one principle:

(1) Controls from references are analogously, "synthesized adjectives."

The pilot question was direct: can a style be extracted from a cluster of images and applied as a factor to new images, the way an adjective works in a text instruction? The mechanism is a difference-of-means activation direction: average the model's internal activations over reference images that share a quality, subtract the average over a content-matched neutral set, and add the result, scaled by a knob α, during generation. It is training-free and established in diffusion models, including for abstract qualities like style (CASteer, 2025; also Concept Sliders, 2023; Rimsky et al., 2024). CASteer applies the direction at cross-attention and only where the quality is present, which limits side effects on the rest of the image — the property needed to avoid dragging content. We over-produce candidate directions and keep the ones that are clean, decoupled, and transferable by inspection — curate, not optimize. Our contribution is the source: the positives are the user's references for a quality they cannot name, not a named concept. The pilot's open problem: the direction tracked the model's default color grade and did not transfer across subjects.

(2) Decomposition, to aim the synthesis. Ideogram-4's Describe endpoint returns JSON that splits an image into **look** (`style_description`: medium, lighting, color palette) and **content** (`compositional_deconstruction`: objects, with bounding boxes). This is the disentanglement lever: build the control from the reference's look while holding the subject's content fixed. The same decomposition also supplies the evaluation's scoring substrate, run there on a separate model to avoid circularity (see Evaluation, below).

(3) Live loop. The user adjusts a control; a low-detail preview updates immediately; the full render completes in the background. Expressing intent and judging the result happen continuously, with no turn boundary.



Principle. Define the target by the user's response: their references and what they recognize, not by the model's internal representation. The pilot's failure was using the model as both the thing steered and the definition of the target, so it returned its default style.

Model: open-weights generator (Ideogram-4) for reproducibility.

Preliminary evidence (a one-week pilot, ~$25 compute; predictions and answer keys locked before scoring).

1. Description did not carry recognition. An expert sorted images by a quality, then wrote down what it was; two independent judges using only her description could not reproduce her sort (near chance; her first-named feature ran backward), though they agreed with each other.
2. The quality is a property of scenes, not model artifacts. Mixed blind with real photographs, her picks appeared about equally in real and generated images (≈16% vs ≈16%).

3. The mechanism produces a coherent control, but an unaimed one. A single control moved the image coherently while holding the subject, yet tracked the model's color grade rather than the target and did not transfer across subjects.

Figure: Pilot α-sweep (one-week pilot, ~$25 compute; keys were locked before scoring). A single synthesized control applied to three scenes (rows), each rendered across control levels α = −0.7 → +0.3 (left → right). The image moves coherently and monotonically with the subject held fixed; the visible change still tracks the model's own color grade rather than the target quality. We address this entanglement in this project.

1.5 Proposal Criteria

Relevance

- The loop is continuous and real-time: adjust → instant change → recognize → adjust (§1.4), the human–AI loop the program targets, in a creative domain its current work (audio, speech) has not reached. The control panel is interface generated per reference set, matching the generative-UI direction. It also contributes an independent evaluation.

Feasibility

- The mechanism is established (CASteer, ICLR 2026); the decomposition is a documented Ideogram endpoint; the pilot ran for $25. Effort goes to the open problem (aiming and transfer). Solo PI, scoped to the controls the pilot supports.

Construct validity

- One question: does discovery reach or beat the expert ceiling on a target the user can recognize but not specify? The evaluation (§1.4) measures exactly that: the look transferred, content preserved, no leakage; per category on a frozen generator, human-primary, with an explicit circularity guard and predictions sealed before scoring.
- **Simplicity & generality:** Training-free; built from standard parts. Generality is tested directly: the core result is replicated on a second, independent open-weights image generator (e.g., FLUX or SDXL), so the finding is a property of the method, not one model. Decomposition and scoring are model-agnostic; everything is released open.

1.6 Timeline, Milestones & Deliverables *(≤ 6 months)*

Month	Milestone	Deliverable
1	Recruit 3 experts; capture intent, references, and a "gold" result per task; pre-register the study	Open gold dataset + sealed protocol
2	Controls from references on Ideogram-4; decomposition pipeline (look/content from Describe JSON); aim control at look, hold content	Control library + decomposition code + before/after examples
3	Live loop: adjust → preview → background render	Working prototype
4	Within-subject study across interfaces (incl. reference-by-example baseline); blind judging + leakage check	Study results: per-trait transfer / preservation / leakage
5	Validate automatic scores against human labels; **stretch:** test cross-person agreement → shared benchmark or per-user finding	Validated metric (or negative result) + analysis
6	Generality: replicate the core result on a second open-weights image generator (e.g., FLUX or SDXL — Thinking Machines has no image model); write-up; open release	Report + open release (prototype, dataset, protocol, code)

1.7 Expected Outcomes & Impact

A working prototype, an open gold dataset (expert intent, references, gold result), and per-trait results (transfer, preservation, leakage) across five interfaces on one fixed generator. An automatic score validated against human judgment, or a negative result showing it cannot be. Stretch: a reusable benchmark if the quality is shared across people; otherwise a per-user finding. Everything open. Risk: separating the wanted quality from incidental color, and holding it across subjects, is the open bet; a negative result is reportable.

Through-line. The proposal opens on a gap: recognition outruns specification, and closes on why it widens. As interaction becomes continuous and real-time, so does intent: it forms and sharpens while a person looks, and turn-based specification cannot keep up. A general

way to express continuous intent by recognition, evaluated with the generator held fixed, is the contribution — reusable for any slow generative medium a person steers by eye, not images alone.

1.8 References (selected)

Industry / problem

- Reve user report — @0xnichy, X, 5 June 2026 (post retweeted by @reve). `https://x.com/0xnichy/status/2062790808080617947`
- Reve, "Introducing References," 2026. https://blog.reve.com/posts/introducing-references/
- Canva, "Magic Layers" (public beta), March 2026.

Prior interfaces (bespoke, hand-built control)

- Generative Remix — author's own. https://www.thirdplane.io/works/generative-remix/
- Mesh Gradient Tool — Chase Davis. https://v1.cloan.me/0.7513130781587641
- Tutti Space Program (physics generator) — Grant Kot. https://tsp.grantkot.com/

Method

- Ideogram 4.0 — Describe endpoint (image → JSON: `style_description`, `compositional_deconstruction`, bounding boxes). https://developer.ideogram.ai/api-reference/api-reference/describe-v4
- "CASteer: Cross-Attention Steering for Controllable Concept Erasure," 2025 (ICLR 2026). arXiv:2503.09630 — training-free difference-of-means steering in diffusion, selective at cross-attention, handles style.
- Gandikota et al., "Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models," ECCV 2024. arXiv:2311.12092 — continuous control directions buildable from a few sample images, including visual concepts hard to describe in text.
- Rimsky et al., "Steering Llama 2 via Contrastive Activation Addition," ACL 2024.
- Arditi et al., "Refusal in Language Models Is Mediated by a Single Direction," NeurIPS 2024.
- Ostermann et al., "From Weights to Activations: Is Steering the Next Frontier of Adaptation?" 2026. arXiv:2604.14090.

Evaluation

- Zhang et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" (LPIPS), CVPR 2018.
- Chiang et al., "Chatbot Arena," ICML 2024.

Interaction & articulation (HCI)

- Schön, "The Reflective Practitioner: How Professionals Think in Action," Basic Books, 1983 — creation as a reflective loop of action and discovery.
- Hutchins, Hollan & Norman, "Direct Manipulation Interfaces," Human-Computer Interaction 1(4), 1985 — the gulfs of execution and evaluation.
- Park et al., "Bridging Gulfs in UI Generation through Semantic Guidance," CHI 2026. arXiv:2601.19171 — nameable semantics as an intermediate representation between intent and output.
- Choi et al., "CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI," CHI 2024. arXiv:2312.11949.
- Dang, Mecke & Buschek, "GANSlider: How Users Control Generative Models for Images using Multiple Sliders," CHI 2022. arXiv:2202.00965.
- Ma et al., "POET: Automated Expansion of Text-to-Image Generation," 2025. arXiv:2504.13392.

2. Budget

Category	Description	Amount (USD)
Researcher time (solo PI)	6 months full-time; builds method, prototype, study	$72,000
Compute (cloud GPU)	Activation capture/injection, decomposition, study runs (open weights)	$12,000
Human evaluation	Expert creators (gold results) + screened blind judges + participants	$9,000
Part-time contributor	`[FILL — optional]`; study logistics, rating pipeline	$5,000
Storage, tooling, misc.	Dataset hosting, preview infrastructure	$2,000
Total cash		**$100,000**

---

3. Team

- **Jasmine Poon** — Independent researcher. (CV attached)
- Actively looking for co-contributor for peer review in own network and at South Park Commons Community NYC

Expert creators who produce gold results are paid participants (budgeted under Human evaluation), not team.