A Picture is Worth a Thousand Prompts? Efficacy Of Iterative Human-driven Prompt Refinement in Image Regeneration Tasks

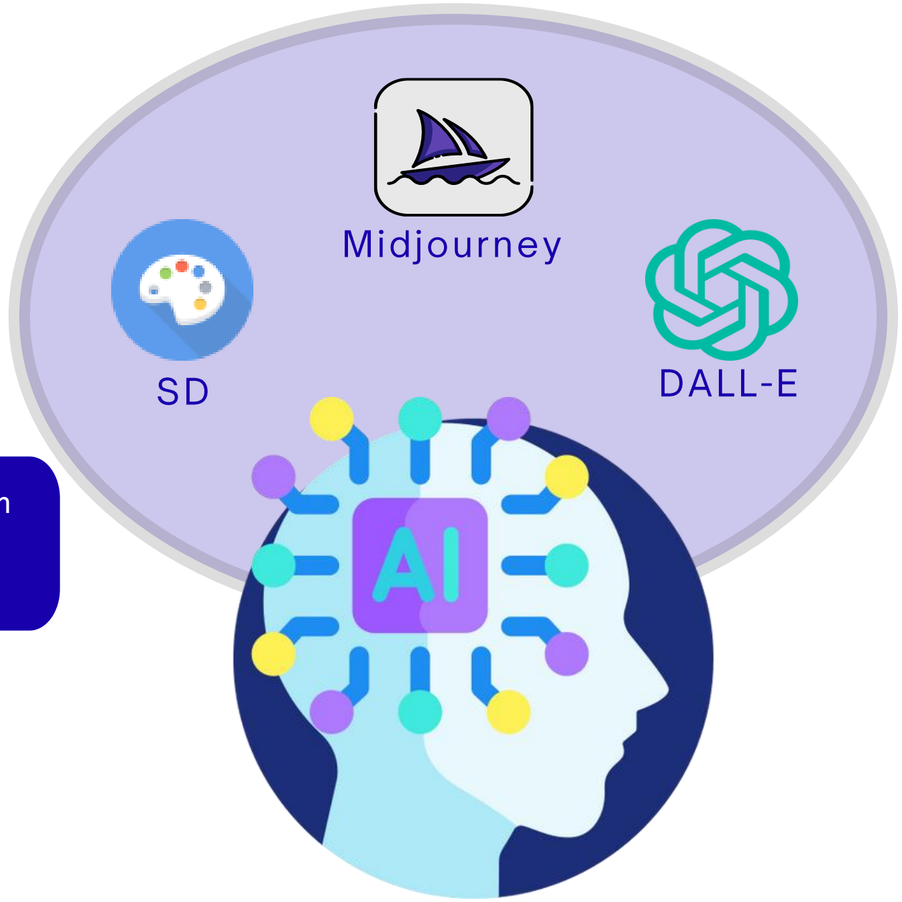
Khoi Trinh, Scott Seidenberger, Raveen Wijewickrama, **Murtuza Jadliwala**, Anindya Maiti

International Joint Conference on Artificial Intelligence (IJCAI)
Montréal, Quebec, Canada
2025

Text-to-Image Models

User

Prompt: An Astronaut with an orange suit giving thumbs up



Text-to-Image Models

Generated Image

Targeted Image Generation

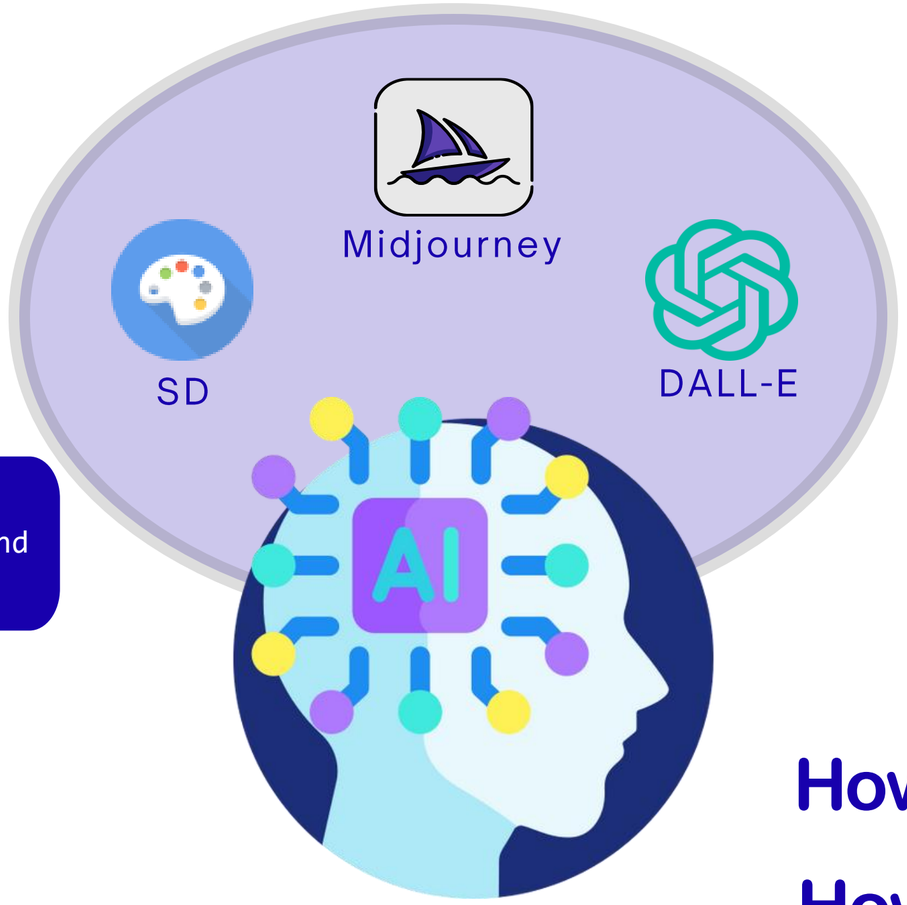
(Or Prompt Inference/Guessing)

User

? Prompt

Prompt guess: Astronaut wearing an orange suit and has two thumbs up

Target Image



Text-to-Image Models

Generated Image

How similar are the images?

How close are the prompts?

Iterative Image Generation

(Or Iterative Prompt Refinement)

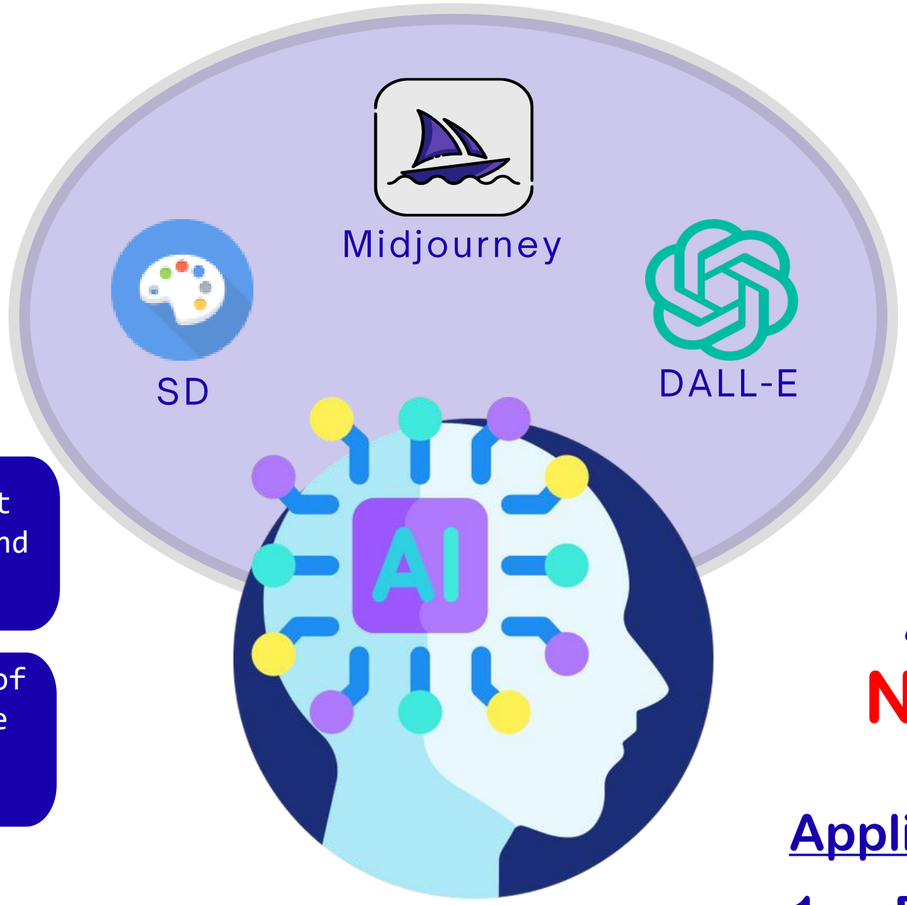
User

? Prompt

Prompt guess 1: Astronaut wearing an orange suit and has two thumbs up

Prompt guess 2: Cartoon of an astronaut in an orange suit on a lunar surface with thumbs up

Target Image



Text-to-Image Models

Generated Image

Are the images similar?
Not really! Let's try again!

Applications

1. Bypassing prompt marketplaces
2. Teaching effective prompt creation
3. Restoring damaged/lost artwork

Iterative Image Generation

(Or Iterative Prompt Refinement)

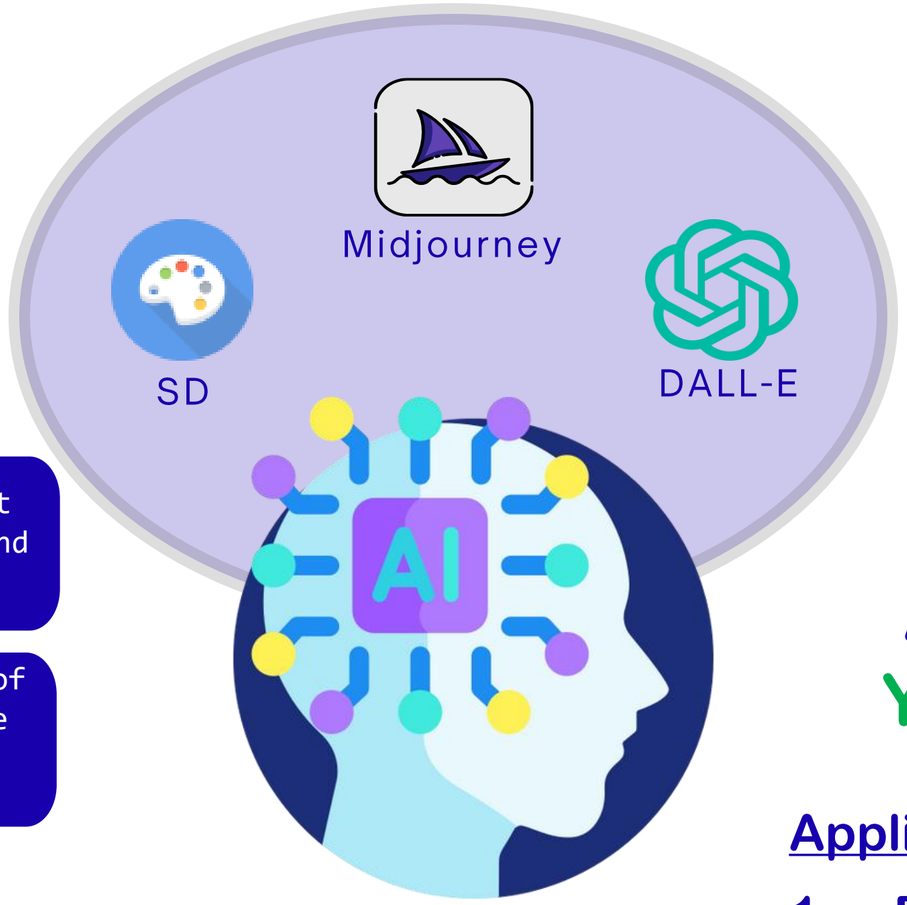
User

? Prompt

Prompt guess 1: Astronaut wearing an orange suit and has two thumbs up

Prompt guess 2: Cartoon of an astronaut in an orange suit on a lunar surface with thumbs up

Target Image



Text-to-Image Models

Generated Image

Are the images similar?

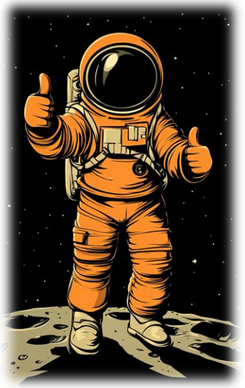
Yes, this is much better!

Applications

1. Bypassing prompt marketplaces
2. Teaching effective prompt creation
3. Restoring damaged/lost artwork

Image Similarity Metrics (ISM)

Target Image



Generated Image

How to measure similarity?

Subjective Metrics

- Human Judgement

Quantitative Metrics

- Perceptual Similarity (PS)
- Contrastive Language Image Pre-training (CLIP) scores
- ImageHash

! We do not use traditional metrics such as L2 distance and SSIM as they fail to capture perceptual nuances due to their pixel-wise independence assumption

Research Gaps & Objectives

Research Gap 1: Unclear if humans are good at iteratively refining prompts

- Objective: Assess effectiveness of iterative human-driven prompt refinement in image generation tasks

Research Gap 2: Unclear if quantitative ISMs (PS, CLIP scores, ImageHash) align well with human judgment during such iterative image generation tasks

- Objective: Evaluate alignment between human judgement of image similarity and ISM scores

Research Questions & Hypotheses

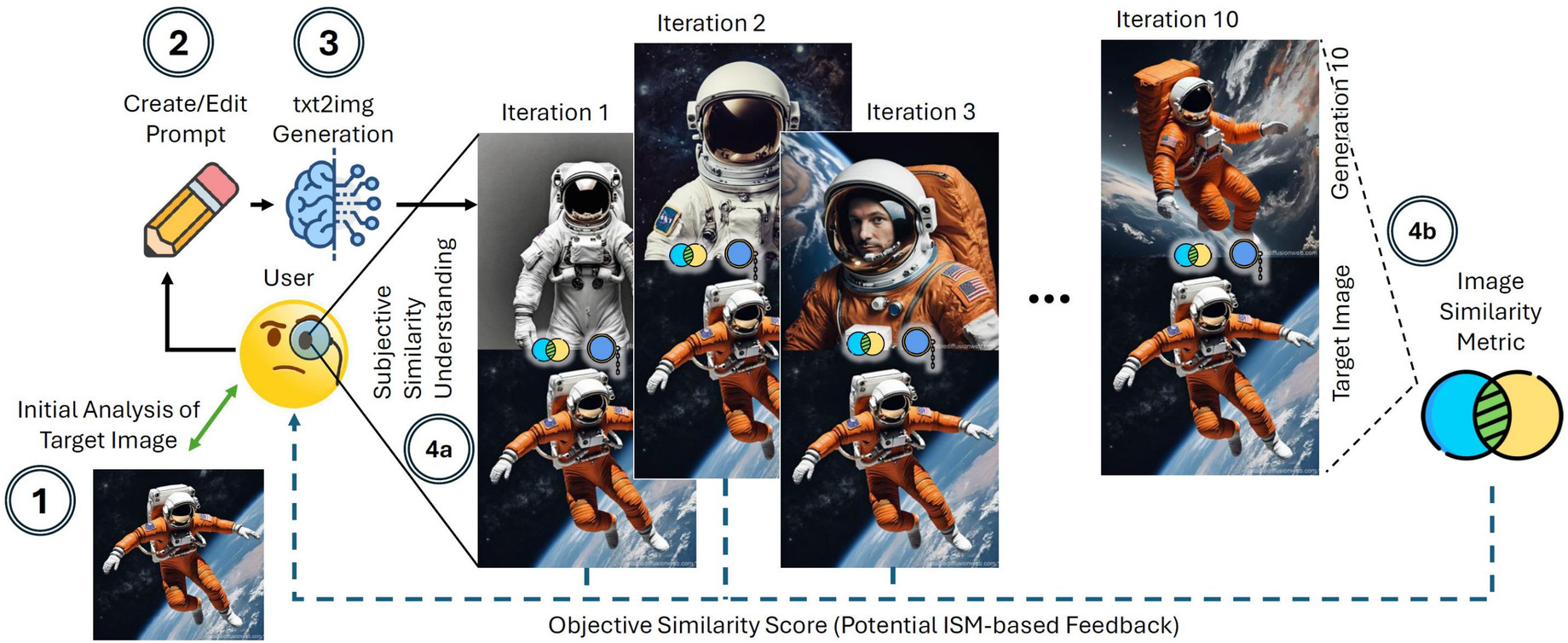
RQ1: Do humans agree that ISMs reliably reflect image similarity?

- **Hypothesis 1.1**: Good human-ISM agreement regarding image similarity
- **Hypothesis 1.2**: No significant differences among ISMs

RQ2: Does iterative refinement improve (generated) image similarity?

- **Hypothesis 2.1**: Iterative refinement increases image similarity
- **Hypothesis 2.2**: Improvement diminishes over iterations
- **Hypothesis 2.3**: Users perceive later iterations as superior

Experiment



Users

- A total of 20 human subject participants divided into 4 groups
- Diverse demographics and varying levels of expertise in usage of text-to-image tools

Image Dataset

- Each participant assigned 10 different target images generated using SD-3
- Each target image used in a prompt refinement task over 10 iterations

Tasks

- Iterative prompt refinement, with ISM feedback provided for half of them
- Ranking images generated across all iterations for each refinement task

Evaluation Methodology

RQ1: Do humans agree that ISMs reliably reflect image similarity?

- Intraclass Correlation Coefficient (ICC)

RQ2: Does iterative refinement improve (regenerated) image similarity?

- Mixed-Effects Model
  - Fixed Effects: Demographic information (age, gender, etc.), subject of the target image, visibility of the ISM on the UI, and type of ISM shown to the user
  - Random Intercepts: For each participant and iteration
- Chi-squared test
  - Aggregated the iteration at which each user selected their top-ranked image into one distribution and performed a chi-square goodness-of-fit test against the null hypothesis of a uniform distribution

Findings

RQ1: Do humans agree that ISMs reliably reflect image similarity?

- ISMs align moderately well with human judgment (Hypothesis 1.1)
- Alignment results are consistent for CLIP models and Perceptual Similarity (Hypothesis 1.2)
- ImageHash performs poorly - excluded from further evaluation!

ISM	ICC	Alignment Rating
PS	0.686	Moderate
CLIP (B32)	0.620	Moderate
CLIP (L14)	0.527	Moderate
ImageHash	0.250	Poor

Findings

RQ2: Does iterative refinement improve (regenerated) image similarity?

- Significant improvements in iterations 1-6, plateau afterward → confirms iterative improvement

Effect	F	p-value
Iteration	11.486	< 0.001
Gender	0.001	0.999
Education	0.001	0.999
Native language	0.253	0.615
txt2img familiarity	0.253	0.615
Subject	5.758	< 0.001
Visibility of ISM	2.061	0.151
Type of ISM	0.446	0.504

Findings

RQ2: Does iterative refinement improve (regenerated) image similarity?

- Subject matter significantly influences outcomes.

Effect	F	p-value
Iteration	11.486	< 0.001
Gender	0.001	0.999
Education	0.001	0.999
Native language	0.253	0.615
txt2img familiarity	0.253	0.615
Subject	5.758	< 0.001
Visibility of ISM	2.061	0.151
Type of ISM	0.446	0.504

Findings

RQ2: Does iterative refinement improve (regenerated) image similarity?

- Visibility of ISM feedback and ISM types had negligible effect.

Effect	F	p-value
Iteration	11.486	< 0.001
Gender	0.001	0.999
Education	0.001	0.999
Native language	0.253	0.615
txt2img familiarity	0.253	0.615
Subject	5.758	< 0.001
Visibility of ISM	2.061	0.151
Type of ISM	0.446	0.504

Findings

RQ2: Users' perception of the iterative refinement process

- Significant preference for later iterations (9 & 10) → Users perceive substantial iterative improvements.

Iteration	User Preference
1	12
2	9
3	9
4	7
5	10
6	11
7	13
8	14
9	21
10	44

Summary of Findings

🔄 Iterative refinement → Improved similarity

✅ Reliable ISMs → PS & CLIP scores

📉 Early iterations → Highest improvement

More interesting results/findings in the paper!

Additional Insights

Limitations

- 👥 🔻 Human subject participants sample (20 university students)
- 🔄 📌 Fixed upper bound on iterations
- 📊 🤔 Design of the feedback (ISM) interface not considered

Future Research

- 💬 🎧 Other modalities of feedback during iterative image generation
- ⚠️ 🛡️ Ethical Issues surrounding prompt inference

Thank You!

Paper: