

Closed-Loop Multimodal Generative AI under Real-World Constraints

Evaluation · Alignment · Generation

Yasi Zhang

University of California, Los Angeles

Advisors: Prof. Ying Nian Wu & Prof. Oscar Leong

June 2026

Outline

○ **About Me** (CV)

○ **Why a Closed-Loop System?** — Research Vision

PART I

● **Evaluating:** How to enable scalable & precise evaluation without expensive human labels?

EdiVal-Agent (ICLR 2026)

PART II

● **Aligning:** How to better align LLM-as-a-Judge with inconsistent human feedback?

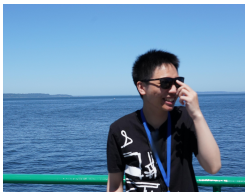
REAL (ICML 2026)

PART III

● **Generating:** How to learn generative models when clean data is exhausted?

RSD (ICLR 2026)

About Me



Yasi Zhang (She/They)

yasminzhang.github.io

my resume

Education

Ph.D., Statistics & Data Science, UCLA (2022–present)

B.S., Data Science & Big Data Tech., Fudan University

Industry Experience

Student Researcher, ByteDance Seed (Spring 2026)

Student Researcher, Google Research (Spring&Summer 2025)

Applied Scientist Intern, Amazon AWS AI Labs (Summer 2024)

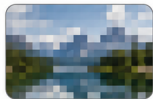
Research Interests

Generative AI, Multimodality, Reinforcement Learning.

Visual Generation Progression

2014–2020 Resolution & Realism Era

2014



GANs

Learning to generate realistic images

2018



BigGAN / StyleGAN

Higher resolution and photorealism

2020



Diffusion

Stable training and high-fidelity synthesis



Key trend:

Higher resolution ·
Better realism ·
Fidelity

Key shift: From visual fidelity to instruction-following creativity

2021–Now Creative & Controllable Generation Era

2021



DALL-E

Text-to-image generation

2022



Stable Diffusion

Open ecosystem and controllable generation

recent



Seedream

Better text rendering
Compositional control



GPT-Image

Style transfer (Ghibli)
Infographics
Creativity



Nano Banana

OOTD
Visual personalization

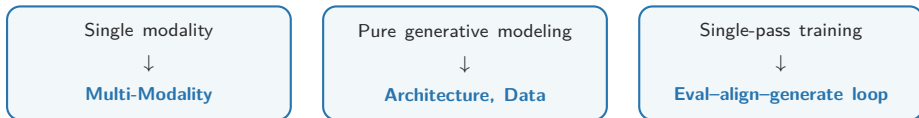


Key trend:

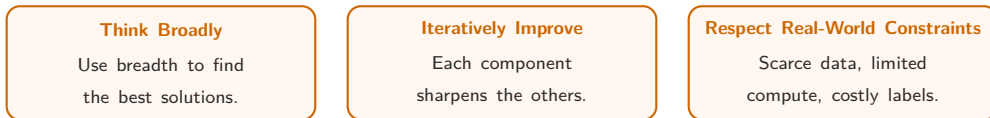
Creative ·
Stronger control ·
Richer use cases

Why a Closed-Loop System?

Three observations from the progression of visual generation:



Three principles to build such powerful multimodal generative AI:



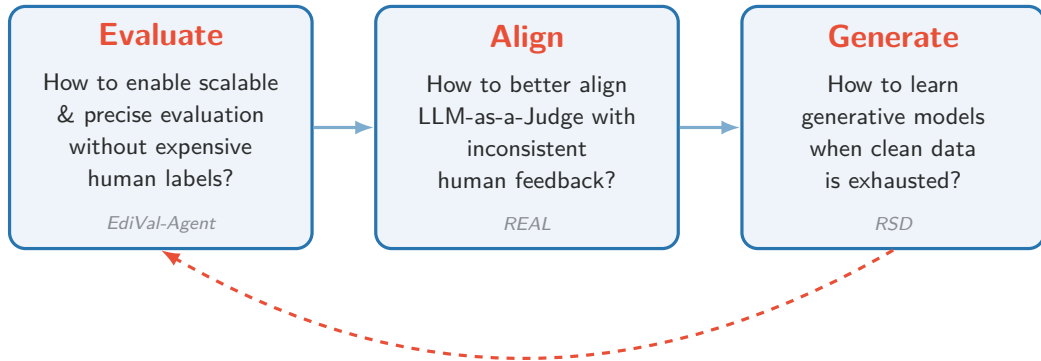
My research direction:

Closed-Loop Multimodal Generative AI under Real-World Constraints

Evaluation · Alignment · Generation

Closed-Loop Multimodal Generative AI under Real-World Constraints

My research aims to improve **each stage** of this loop by tackling:



Each component sharpens the others and contributes a virtuous closed loop.

PART I

Evaluating

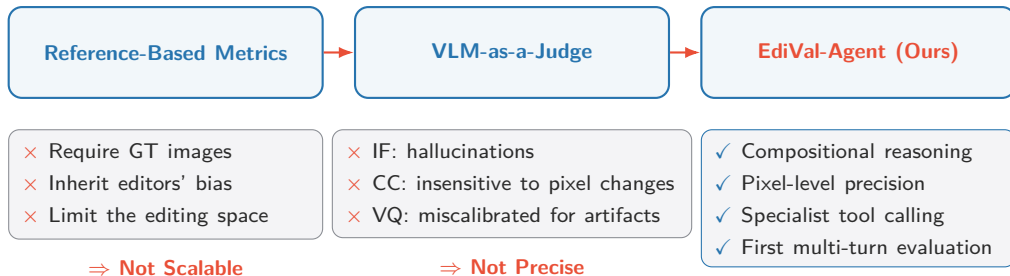
How to enable scalable & precise evaluation
without expensive human labels?

EdiVal-Agent: An Object-Centric Framework for
Automated, Fine-Grained Evaluation of Multi-Turn Editing. ICLR 2026.

MT-EditFlow: Reinforcement Learning for
Multi-Turn Image Editing with Flow Matching. In submission.

EdiVal-Agent — Motivation

Problem: Instruction-based image editing has advanced rapidly, yet **scalable**, **precise**, and **multi-turn** evaluation remains a bottleneck.



EdiVal-Agent — Pipeline

Stage 1: Decomposition

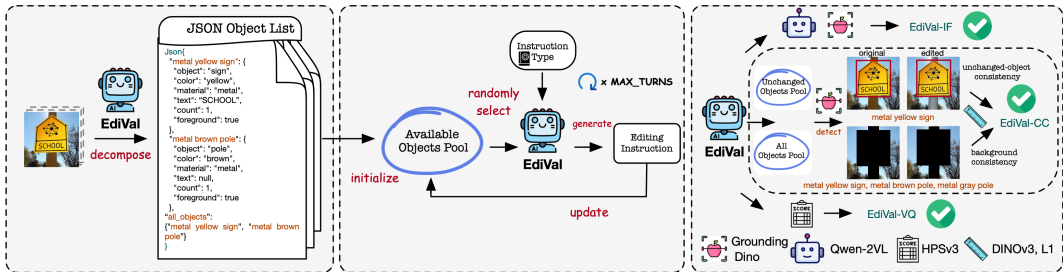
Parse images into structured object-level descriptions

Stage 2: Instruction Gen

Generate diverse multi-turn editing instructions

Stage 3: Evaluation

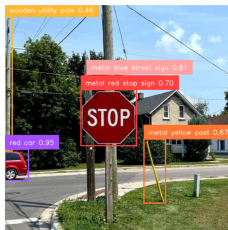
Assess edits on IF, CC, VQ with specialist tools



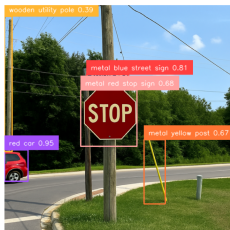
Note: EdiVal metrics are enabled **only** through our structured instruction generation pipeline.

EdiVal-Agent — Qualitative Visualization

Instruction: "Remove brick beige house."



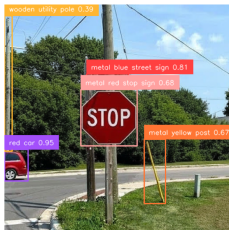
Base Image



GPT-Image-1

EdiVal-IF: Good

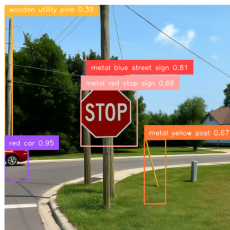
EdiVal-CC: position shifted, appearance changed



Nano Banana

EdiVal-IF: Good

EdiVal-CC: Good



Qwen-Image-Edit

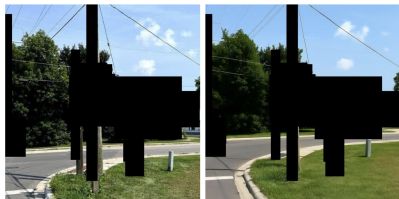
EdiVal-IF: Good

EdiVal-CC: Position shifted

- **EdiVal-IF:** "brick beige house" should not be detected in the edited image.
- **EdiVal-CC (object-level):** unchanged objects like the stop sign should remain unchanged in both position (detection box) and appearance (DINO features).

EdiVal-Agent — Qualitative Visualization (cont.)

EdiVal-CC: Background Consistency



Base image

Qwen-Image-Edit

EdiVal-VQ: Beautification vs. Preservation



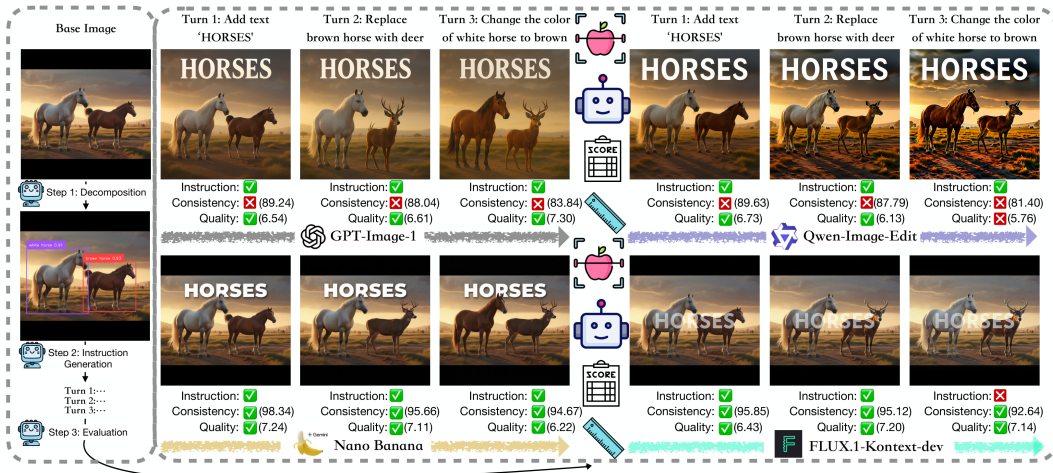
(a) Base image

(b) GPT-Image-1

(c) FLUX.1-max

- **EdiVal-CC (background-level):** the **background** (excluding the area of all objects) should remain the same as the base image.
- **EdiVal-VQ:** GPT-Image tends to **beautify** the image (good or bad depends on user preference).

EdiVal-Agent — Qualitative Comparison



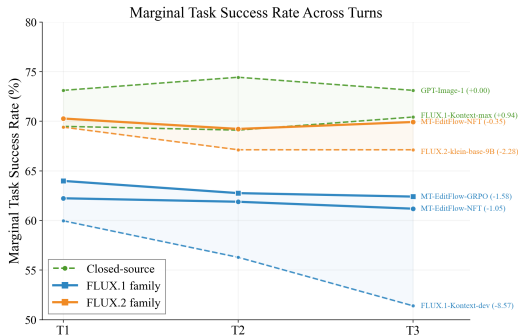
GPT-Image-1: good IF, but fails at consistency.
Nano Banana: achieves a good balance.

Qwen: exposure bias degrades later turns.
FLUX.1-Kontext: fails at multi-turn IF.

EdiVal-Agent — Summary

1. **Shifts** from holistic evaluation to **compositional, object-level reasoning** with **agentic specialist tool calling**, and tailored **pixel-level precision** metrics.
2. **Introduces** the first **multi-turn** image editing evaluation framework, exposing **compounding failure modes** that single-turn benchmarks systematically miss.
3. **Achieves** significantly stronger **human agreement accuracy** than all prior automated metrics including reference-based metrics and VLM-as-a-Judge.
4. **Reveals** through fine-grained analysis that **no single model dominates all axes** (IF, CC, VQ), informing targeted improvements.

Follow-up: MT-EditFlow — Improving Multi-Turn Editing with RL



Key Takeaway

- Open-source editors (FLUX.1/2) exhibit **marginal editing capability** degradation.
- MT-EditFlow **stabilizes multi-turn performance** by optimizing the reward mechanism.

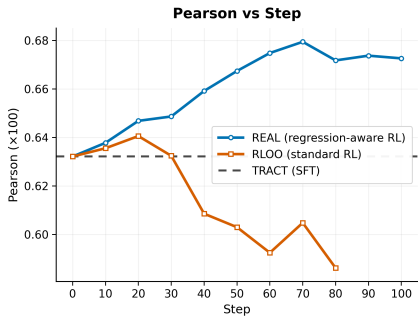
PART II

Aligning

How to better align LLM-as-a-Judge
with inconsistent human feedback?

REAL: Regression-Aware Reinforcement Learning for LLM-as-a-Judge. ICML 2026.

EBAMA: Object-Conditioned Energy-Based Attention Map
Alignment in Text-to-Image Diffusion Models. ECCV 2024.



Standard RL fails to improve LLM-as-a-Judge.

Problem: standard RL uses binary 0/1 rewards: predicting 4 when truth is 5 is penalized the same as predicting 1.

Standard RL for LLMs Given prompt x , CoT c , answer y , policy π_θ , standard RL is to maximize:

$$\mathcal{L}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, (c, y) \sim \pi_\theta(\cdot | x)} [r(x, y)]$$

Standard REINFORCE Gradient Policy gradient:

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, (c, y) \sim \pi_\theta(\cdot | x)} [r(x, y) \nabla_\theta \log \pi_\theta(c, y | x)]$$

Key assumption: reward does not depend on θ , i.e. $\nabla_\theta r = 0$.

RAIL Inference (Lukasik et al., 2024) Expected value over digit tokens $\mathcal{K} = \{0, \dots, 9\}$:

$$\hat{y}_\theta(x, c) = \sum_{k \in \mathcal{K}} k \cdot \pi_\theta(k | x, c)$$

Improves Pearson & Spearman **for free** at inference time.

REAL Objective

Replace binary reward with a **regression-aware** reward:

$$r_{\text{REAL}}(\theta, x, c) = -(\hat{y}_{\theta} - y^*)^2 + \lambda \log \pi_{\theta}(y^* | x, c)$$

- **Squared-error:** predicting 4 when truth is 5 is *much better* than predicting 1.
- **Log-likelihood:** maintains next-token prediction ability.

The Challenge: Reward depends on θ , so standard REINFORCE is **invalid**.

⇒ **Our solution: Generalized Policy Gradient.**

Lemma 4.1 Generalized Policy Gradient with Policy-Dependent Rewards

Standard RL assumes $\nabla_{\theta} r = 0$. REAL's reward $r(\theta, x, c) = -(\hat{y}_{\theta} - y^*)^2 + \lambda \log \pi_{\theta}(y^* | x, c)$ depends on θ . The generalized gradient:

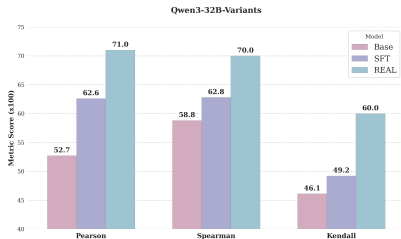
$$\nabla_{\theta} \mathcal{L} = \mathbb{E}_{x, c \sim \pi_{\theta}} \left[\underbrace{r(\theta, x, c) \nabla_{\theta} \log \pi_{\theta}(c | x)}_{\text{Term 1: CoT Exploration}} + \underbrace{\nabla_{\theta} r(\theta, x, c)}_{\text{Term 2: Prediction Refinement}} \right]$$

- **Term 1:** policy gradient $\times$ regression reward $\rightarrow$ explores reasoning trajectories.
 - **Term 2:** backprop through $\hat{y}_{\theta} \rightarrow$ refines the numeric score.
- $\Rightarrow$ **Jointly optimizes reasoning quality and numerical accuracy.**

REAL — Results & Takeaways

Empirical results (8B $\rightarrow$ 32B).

- Outperforms regression-aware SFT *and* standard RL.
- Strong **out-of-domain** generalization.



Remark: Handling Inconsistent Human Feedback

For a given response, humans may provide different ratings $y_1, \dots, y_n$. REAL's objective naturally handles this: $(\hat{y}_\theta - \bar{y})^2$, where $\bar{y} = \frac{1}{n} \sum_i y_i$. At inference, RAIL's expected-value decoding $\hat{y}_\theta = \sum_k k \cdot \pi_\theta(k \mid x, c)$ also reflects the **expectation** of human feedback.

Key Takeaway

In regression-aware RL, the **probability** assigned to the answer token offers a richer, more informative reward signal than **binary accuracy** alone.

PART III

Generating

How to learn generative models
when clean data is exhausted?

Score Distillation Beyond Acceleration:

Generative Modeling from Corrupted Data. ICLR 2026.

Flow Priors for Linear Inverse Problems via

Iterative Corrupted Trajectory Matching. NeurIPS 2024.

Restoration Score Distillation — Motivation

Clean data is running out

Frontier models have consumed most available high-quality data. Can we use:

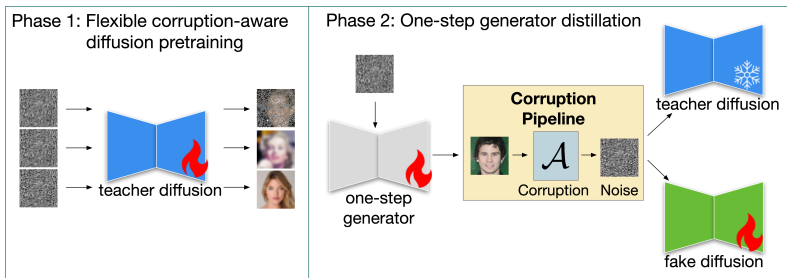
- **Synthetic data?** Risks model collapse.
- **Real-world data?** Inherently corrupted.
- **Scientific domains?** Clean ground truth never existed or costly to collect.

⇒ **Can we learn generative models from corruption?**

Our key insight

Score distillation *goes beyond acceleration* and is surprisingly effective at learning generative models from corruption.

Restoration Score Distillation (RSD)



Teacher diffusion: ϕ , One-step generator (student): θ , Fake diffusion: ψ

Phase I — Flexible Corruption-Aware Diffusion Pretraining:

$$\mathcal{L}_{\text{teacher}}(\phi) = \mathbb{E}_{y, \sigma_t, \epsilon} \left[\lambda(t) \|f_{\phi}(y_t, t) - y\|_2^2 \right] \quad \# \text{ Teacher learns corrupted distribution } p_Y$$

Phase II — One-Step Generator Distillation (Fisher Divergence):

$$\mathcal{L}_{\text{distill}}(\theta) = \mathbb{E}_{\sigma_t, z} \mathbb{E}_{x_{\theta} = G_{\theta}(z)} \left[w(t) \|f_{\phi}(y_t, t) - f_{\psi}(y_t, t)\|_2^2 \right], \quad y_t = \text{CorruptionPipeline}(x_{\theta})$$

Student generates x_{θ} , corrupts it, matches teacher distribution

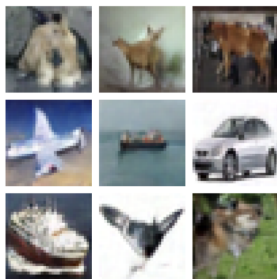
Restoration Score Distillation (RSD) — Results

Training data



$\sigma = 0.2$

One-step generation



FID: 4.77

CIFAR-10 (32×32) FID ↓

DDPM (clean data) 4.04

Teacher diffusion 12.21

RSD (ours) 4.77

Efficiency (30× speedup)

Time to generate 50k images:

Teacher diffusion: ~10 min

RSD (one-step): ~20 sec

on 4× NVIDIA RTX A6000

Note: RSD requires **no clean data** during training.

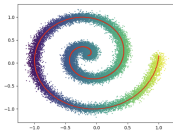
Restoration Score Distillation — Theory

Theorem 1 Student Beats Teacher

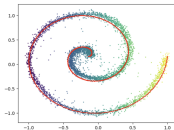
Under linear Gaussian setting, minor assumptions on perfect score estimation and a bounded noise schedule, the global minimizers θ_σ^* of the score distillation loss satisfy

$$\underbrace{W_2^2(p_{G_{\theta_\sigma^*}}, p_X)}_{\text{Student Distance}} = \underbrace{W_2^2(p_{Y,\sigma}, p_X)}_{\text{Teacher Distance}} - \underbrace{(d-r)\sigma^2}_{>0} < W_2^2(p_{Y,\sigma}, p_X).$$

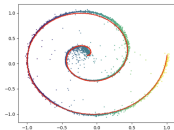
where $W_2(\cdot, \cdot)$ denotes the Wasserstein-2 distance between two distributions.



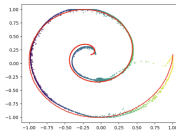
(a) Noisy dataset $\sigma = 0.05$



(b) Teacher-Full



(c) Teacher-Truncated



(d) RSD (ours one-step)

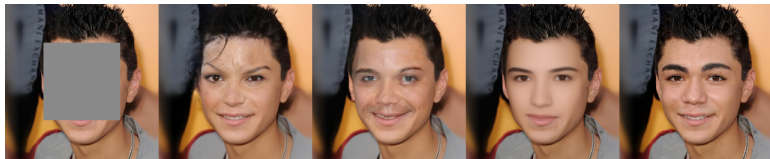
Restoration Score Distillation — Takeaways

1. Score distillation serves as a **principled method** for learning generative models from corruption—going beyond mere acceleration.
2. Achieves **one-step generation (30× speedup)** and **state-of-the-art FID** across denoising, deblurring, inpainting, super-resolution, and multi-coil MRI.
3. Provides **theoretical guarantees** for the implicit regularization of restoration score distillation (student provably beats teacher).
4. **Downstream:** the learned clean prior powers **inverse problems**—*Flow Priors* (NeurIPS 2024).

Downstream: Flow Priors for Linear Inverse Problems



(b) Inpainting (random)



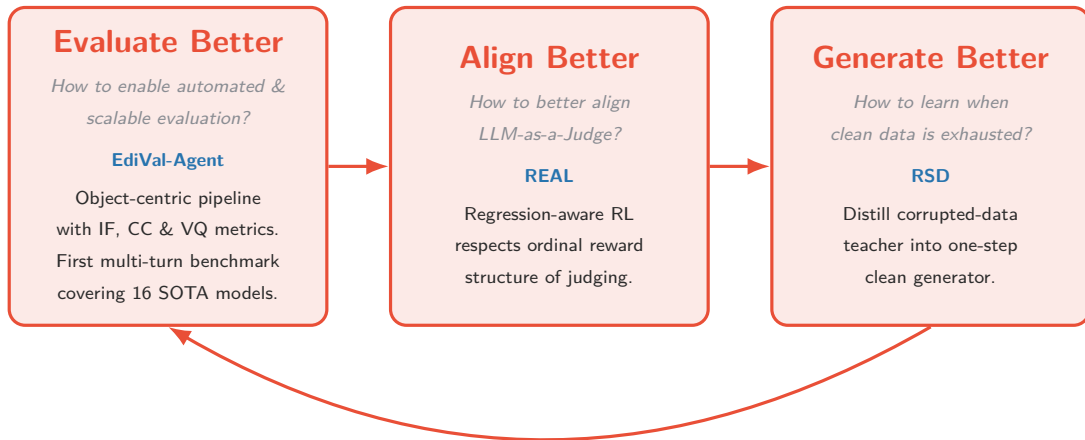
(d) Inpainting (box)

Key Takeaway

RSD: corrupted distribution $\rightarrow$ clean distribution / prior.

Flow Priors: corrupted image $\rightarrow$ clean image (with the help of the learned clean prior).

A Virtuous Closed Loop — Answered



Thank You!

Questions are welcome.

References

- 1 Tianyu Chen*, **Yasi Zhang***, et al.
EdiVal-Agent: An Object-Centric Framework for Automated, Fine-Grained Evaluation of Multi-Turn Editing. [ICLR 2026](#).
arxiv.org/abs/2509.13399
- 2 **Yasi Zhang***, Tianyu Chen*, Zhendong Wang, Ying Nian Wu, Mingyuan Zhou, Oscar Leong.
Score Distillation Beyond Acceleration: Generative Modeling from Corrupted Data. [ICLR 2026](#).
arxiv.org/abs/2505.13377
- 3 **Yasi Zhang**, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, Oscar Leong.
Flow Priors for Linear Inverse Problems via Iterative Corrupted Trajectory Matching. [NeurIPS 2024](#).
arxiv.org/abs/2405.18816
- 4 **Yasi Zhang***, Tianyu Chen*, Mingyuan Zhou, Oscar Leong, Ying Nian Wu, Michal Lukasik.
REAL: Regression-Aware Reinforcement Learning for LLM-as-a-Judge. [ICML 2026](#).
arxiv.org/abs/2603.17145
- 5 **Yasi Zhang**, Peiyu Yu, Ying Nian Wu.
Object-Conditioned Energy-Based Attention Map Alignment in Text-to-Image Diffusion Models. [ECCV 2024](#).
arxiv.org/abs/2404.07389
- 6 Jiahui Huang*, **Yasi Zhang***, et al.
MT-EditFlow: Reinforcement Learning for Multi-Turn Image Editing with Flow Matching. [In submission](#).
arxiv.org/abs/2606.01985

* Equal contribution