Efficient Weighted Sampling via Score-based Generative Models

Heasung Kim* | Taekyun Lee* | Hyeji Kim | Gustavo de Veciana

The University of Texas at Austin | 6G@UT / WNCG | *Equal contribution

Problem Setting

$q(\mathbf{x}) \propto w(\mathbf{x})\, p(\mathbf{x}), \qquad \nabla_{\mathbf{x}} \log q_t(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \tilde{g}(\mathbf{x}, t).$

Weighted sampling reweights a base distribution using a task-specific weight function. The goal is to reuse a pre-trained score model for p and sample from q without re-training.

- Prior training-free methods often rely on Hessians or resampling-heavy inference.
- We want the weighted target distribution while preserving low inference cost on large diffusion models.

Core Idea: LAGS

LAGS uses a Hessian-free first-order correction with time-adaptive guidance strength.

- **Hessian-free correction:**

$$\nabla_{\mathbf{x}} \log q_t(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \tau(t)\, \tilde{g}^{(1)}(\mathbf{x}, t)$$

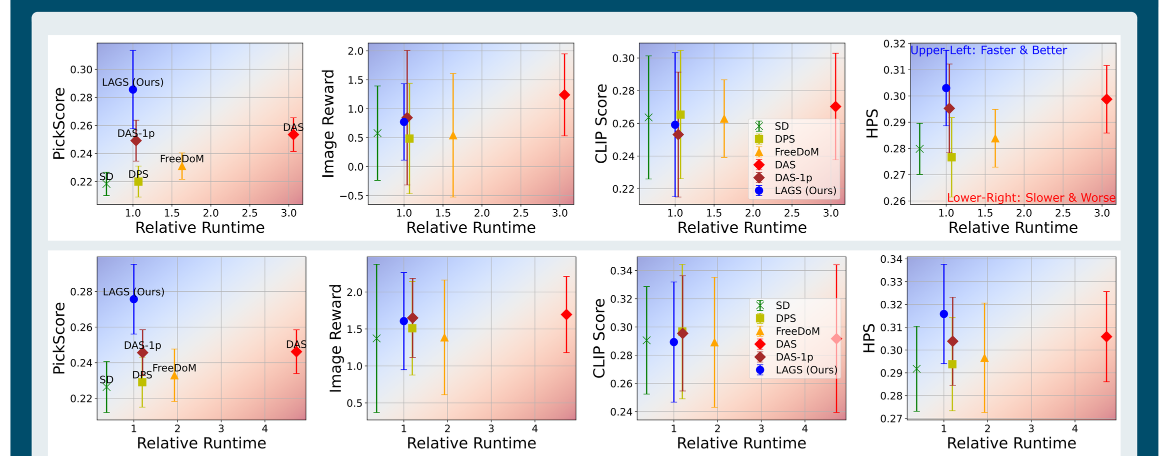
- **One extra score call instead of Hessian-vector products:**

$$\tilde{g}^{(1)}(\mathbf{x}, t) = \frac{\nabla_{\tilde{\mathbf{x}}'_{0|\mathbf{x},t}} \log w(\tilde{\mathbf{x}}'_{0|\mathbf{x},t})}{\sqrt{\bar{\alpha}(t)}} + \frac{\nabla_{\mathbf{x}} \log p_t(\mathbf{x} + \epsilon \nabla \log w) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x})}{\epsilon (1 - \bar{\alpha}(t))^{-1} \sqrt{\bar{\alpha}(t)}}.$$

$$\tau(t) \approx \left(1 + c\frac{(1 - \bar{\alpha}(t))^2}{\bar{\alpha}(t)^2}\right)^{-1}$$

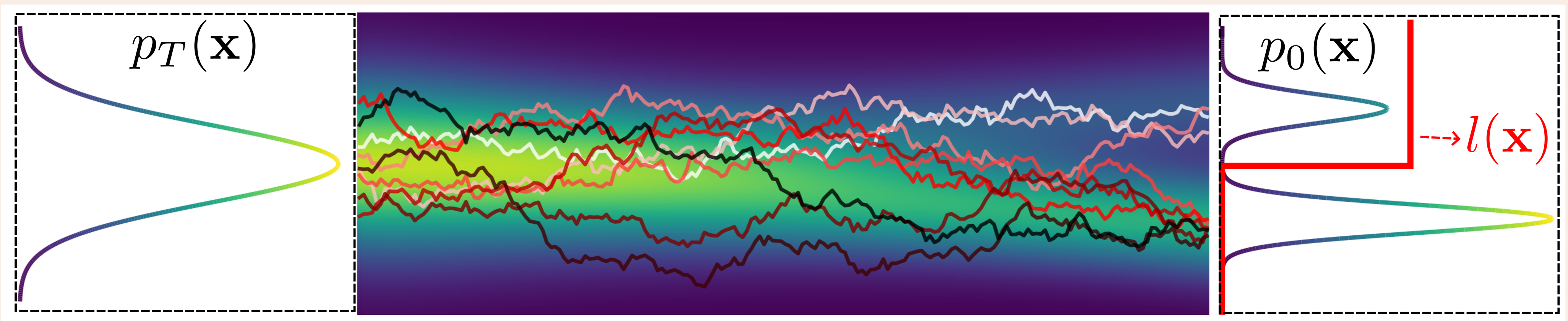
Key difference: instead of Hessian-vector products, LAGS uses one extra score evaluation to form a finite-difference correction, then increases its weight near the end of reverse sampling.

Text-to-Image Alignment with Human Preference Scores



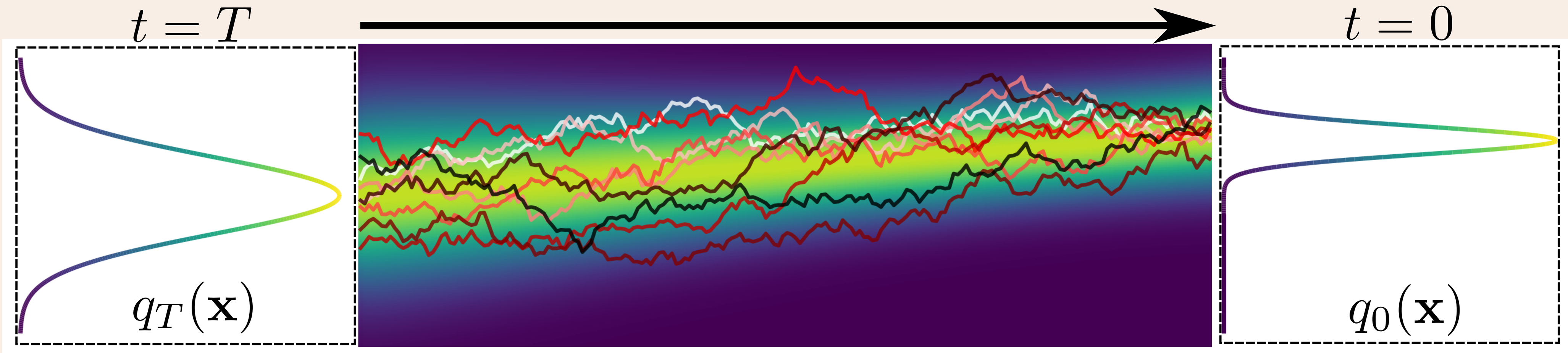
- Highest PickScore with the lowest runtime.

Method Overview



$$\mathrm{d}\mathbf{X}'_t = \left(f(\mathbf{X}'_t, t) - \sigma(t)^2 \nabla_{\mathbf{X}'_t} \log p_t(\mathbf{X}'_t)\right) \mathrm{d}t + \sigma(t)\, \mathrm{d}\tilde{\mathbf{W}}_t$$

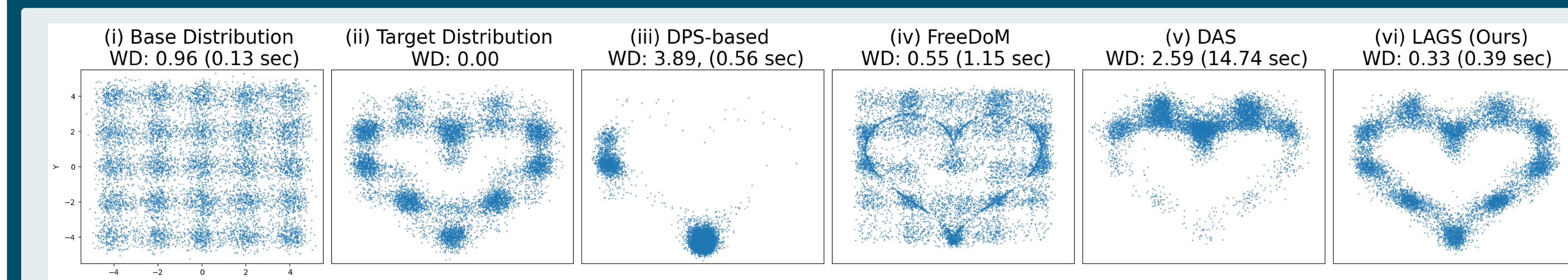
$t = T \longrightarrow t = 0$

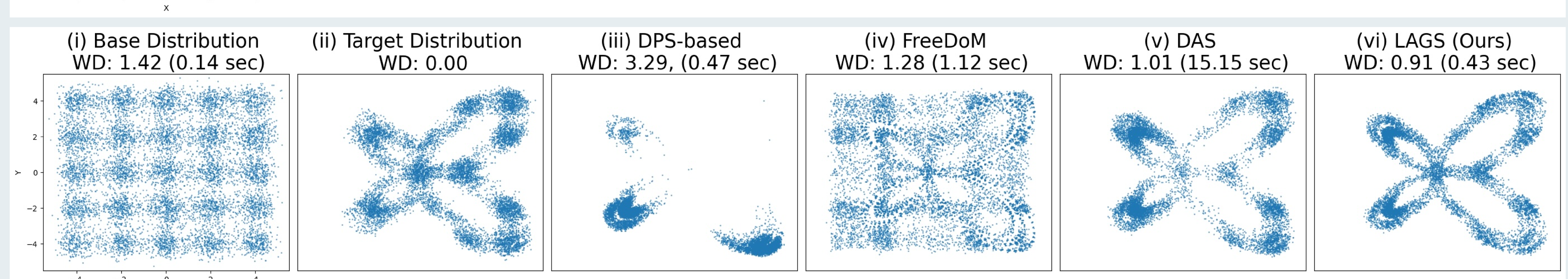


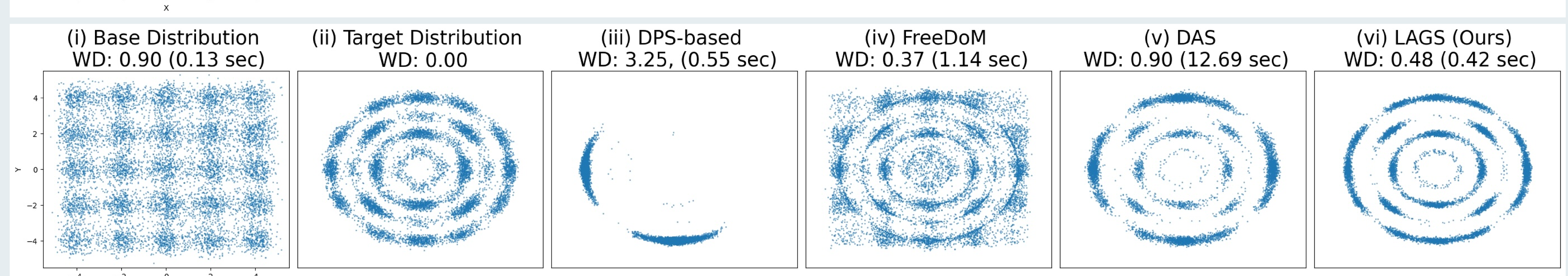
$$\mathrm{d}\mathbf{X}_t = \left(f(\mathbf{X}_t, t) - \sigma(t)^2 \nabla_{\mathbf{X}_t} \log q_t(\mathbf{X}_t)\right) \mathrm{d}t + \sigma(t)\, \mathrm{d}\tilde{\mathbf{W}}_t$$

LAGS augments the pretrained base score with a lightweight guidance term derived from the weight function, then scales that guidance according to diffusion-time uncertainty. No retraining, no Hessians, no resampling-heavy inference.

2D Multimodal Weighted Sampling







Across the heart, butterfly, and rings targets, LAGSconsistently matches structured multimodal targets while remaining fast.

- **Lowest Wasserstein distance** and **fastest runtime** on the main 2D task.
- 10^4 samples in **0.33** s.

Qualitative SDXL Results

Highest PickScore Samples

Lowest PickScore Samples

Across these prompts, the highest PickScore samples preserve object count, spatial relations, and style more faithfully. The lower-PickScore samples show the typical failure modes that the proposed guidance helps avoid.

- Better semantic fidelity to prompts such as object count, relative placement, and style.

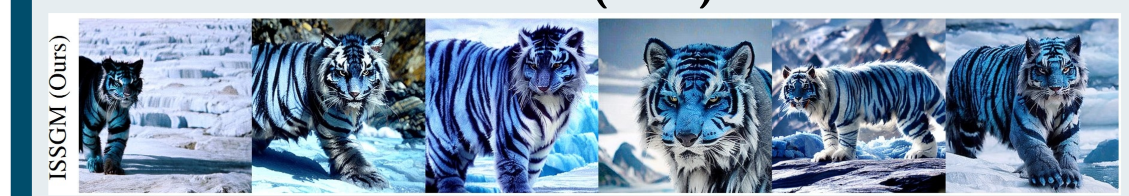
More Qualitative Results

Prompt: "A blue-furred tiger on a glacier."

SD (base)

DAS

LAGS (ours)



Stacking the rows makes the qualitative gap easier to read: the proposed method better matches the requested blue fur while preserving the glacier scene.

Takeaways

- **Training-free:** adapt a pretrained score model to a new weighted target without fine-tuning.
- **Efficient:** avoid Hessians and resampling while preserving strong target performance.
- **Scalable:** gains become more pronounced on larger generators such as SDXL.