WEAK-TO-STRONG DIFFUSION WITH REFLECTION

Lichen Bai
xLeaF Lab, HKUST(GZ)

Masashi Sugiyama
RIKEN AIP & UTokyo

Zeke Xie *
xLeaF Lab, HKUST(GZ)

Figure 1: The qualitative results of W2SD. Here we set Juggernaut-XL as strong model and SDXL as weak model. We present more cases in Appendix D.2.

ABSTRACT

The goal of diffusion generative models is to align the learned distribution with the real data distribution through gradient score matching. However, inherent limitations in training data quality, modeling strategies, and architectural design lead to inevitable gap between generated outputs and real data. To reduce this gap, we propose Weak-to-Strong Diffusion (**W2SD**), a novel framework that utilizes the estimated difference between existing weak and strong models (i.e., weak-to-strong difference) to bridge the gap between an ideal model and a strong model. By employing a reflective operation that alternates between denoising and inversion with weak-to-strong difference, we theoretically understand that W2SD steers latent variables along sampling trajectories toward regions of the real data distribution. W2SD is highly flexible and broadly applicable, enabling diverse improvements through the strategic selection of weak-to-strong model pairs (e.g., DreamShaper vs. SD1.5, good experts vs. bad experts in MoE). Extensive experiments demonstrate that W2SD significantly improves human preference, aesthetic quality, and prompt adherence, achieving SOTA performance across various modalities (e.g., image, video), architectures (e.g., UNet-based, DiT-based, MoE), and benchmarks. For example, Juggernaut-XL with W2SD can improve with the HPSv2 winning rate up to **90%** over the original results. Moreover, the performance gains achieved by W2SD markedly outweigh its additional computational overhead, while the cumulative improvements from different weak-to-strong difference further solidify its practical utility and deployability. The code is publicly available at github.com/xie-lab-ml/Weak-to-Strong-Diffusion-with-Reflection.

1 INTRODUCTION

In probabilistic modeling, the estimated density represents model-predicted distributions while the ground truth density reflects actual data distributions. Diffusion models (Song et al.), known for its

*xLeaF Lab, The Hong Kong University of Science and Technology (Guangzhou). Correspondence to: Zeke Xie **zekexie@hkust-gz.edu.cn**

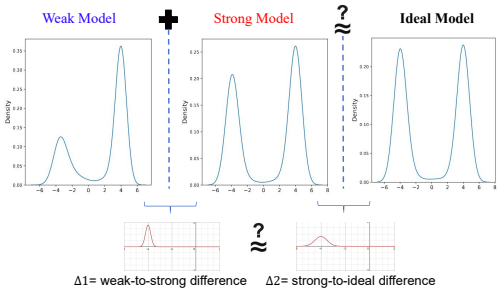

Figure 2: W2SD leverages the gap between weak and strong models to bridge the gap between strong and ideal models.

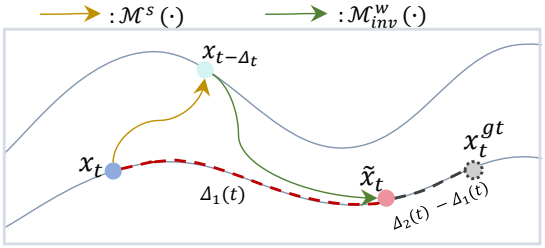

Figure 3: Visualizing the effectiveness of W2SD. When the weak-to-strong difference closely bridges the strong-to-ideal difference (e.g., $\Delta_2(t) - \Delta_1(t)$ is small), the refined latent variable $\tilde{x}_t$ converges to the ideal latent variable x_t^{gt}.

powerful generative capabilities and diversity, estimate the gradient of the log probability densities in perturbed data distributions by optimizing a score-based network, which has become a mainstream paradigm in many generative tasks (Podell et al.; Guo et al., 2023).

To bridge the gap between the learned and real data distributions in diffusion models, extensive research has focused on improving gradient alignment of log probability densities between estimated results and ground truth through various strategies (Karras et al., 2022; Song & Ermon, 2020; Ho et al., 2022). However, constrained by practical limitations including architectural design and dataset quality issues, existing diffusion models inevitably exhibit gradient estimation errors, leading to a gap between the learned and real data distributions. Depending on the magnitude of this gap, we can classify them as either "strong" or "weak" diffusion models. Specifically, due to the inaccessibility of the real data distribution, we cannot establish direct methodologies to measure this gap and effectively reduce it.

In this work, we demonstrate that the gap between the strong and weak models can be used to empirically bridge the gap between the ideal and strong models. In Figure 2, we propose to use the estimated weak-to-strong difference Δ_1 to bridge the inaccessible strong-to-ideal difference Δ_2, thereby enhancing the strong model toward the ideal model. We note this weak-to-strong concept has been widely applied in training processes of machine learning, with early instances such as AdaBoost (Schapire, 2013), where the ensembling of weak models enhances the performance of strong models. Also, Burns et al. (2023) have demonstrated that weak models can serve as supervisory signals to assist in the alignment during the training of large LLMs.

To empirically estimate the weak-to-strong difference, we draw upon reflective mechanisms, the process of modifying generated outputs based on prior states, which have been extensively studied in the field of LLMs (Gou et al.; Madaan et al., 2024; Shinn et al., 2024). Specifically, we propose Weak-to-Strong Diffusion (**W2SD**), a novel framework that leverages weak-to-strong difference to bridge the strong-to-ideal difference. By incorporating a reflective operation that alternates between denoising and inversion based on the weak-to-strong difference, we theoretically understand in Section 2.1 that W2SD guides latent variables along sampling trajectories, effectively steering them toward regions of the real data distribution. And we provide qualitative results in Figure 1.

We emphasize that W2SD can be generalized to diverse application scenarios. Notably, in Appendix F.2 we demonstrate existing inference enhancement methods including Re-Sampling (lug, 2022), Z-Sampling (Bai et al., 2024), FreeDom (Yu et al., 2023) and TFG (Ye et al., 2024) can be reinterpreted as specialized instances of W2SD. Users can define "weak-to-strong" model pairs based on their specific needs. Depending on the type of the model pair (e.g., DreamShaper vs. SD1.5, good experts vs. bad experts in MoE), different effects of improvements can be achieved. We provide further application analysis in Section 3 to demonstrate the broad applicability of W2SD.

The contributions can be summarized as follows.

First, we introduce the weak-to-strong concept into the inference enhancement of diffusion models, demonstrating that the gradient difference of estimated log probability densities between weak and strong diffusion models can approximate the difference between strong and ideal models, consequently bridging the gap between the learned and real data distribution.

Second, to estimate the weak-to-strong difference, we propose a novel inference framework called W2SD, with theoretical understanding that implicitly estimates the weak-to-strong difference through iterative reflection, effectively steering latent variables along sampling trajectories toward regions corresponding to the real data distribution.

Third, extensive experiments validate the effectiveness and broad applicability of W2SD across various generation tasks (e.g., image, video), architectures (e.g., UNet-based, DiT-based), and evaluation metrics. By defining various weak-to-strong model pairs, W2SD achieves diverse improvements effects, such as human preference, prompt adherence, and personalization. Importantly, these improvements effects can be cumulative in parallel, further enhancing the quality of the generation. In efficiency evaluations, W2SD's improvements surpass its overhead, maintaining superior quality over the baseline under equal time constraints, suggesting its strong versatility and applicability.

2 METHOD

In this section, we proposed Weak-to-Strong Diffusion and discuss its mechanism.

2.1 WEAK-TO-STRONG DIFFUSION

In this subsection, we integrate the weak-to-strong concept into diffusion model inference, introduce the W2S algorithm, and establish its theoretical understanding.

Gradient Difference of Log Probability Densities In diffusion models, the goal is to minimize the gradients difference of log probability densities between estimated results and ground truth. However, due to the real data distribution is inaccessible, this difference cannot be directly quantified. To tackle this issue, we consider models of varying capacities: a strong model (with its corresponding denoising process denoted as $\mathcal{M}^{\mathrm{s}}$ and the estimated density as p^{s}) and a weak model (similarly, $\mathcal{M}^{\mathrm{w}}$ and p^{w}).

As shown in Figure 2, we define the weak-to-strong difference as $\Delta_1 = \nabla \log p^{\mathrm{s}} - \nabla \log p^{\mathrm{w}}$, and the strong-to-ideal difference as $\Delta_2 = \nabla \log p^{\mathrm{gt}} - \nabla \log p^{\mathrm{s}}$. By approximating Δ_2 using the estimable Δ_1, we indirectly reduce the gap between existing diffusion models and the ideal model, bringing the learned distribution closer to the real data distribution. In the next subsection, we introduce how to achieve this approximation from a reflective perspective.

W2SD Consider a strong model $\mathcal{M}^{\mathrm{s}}$ and a weak model $\mathcal{M}^{\mathrm{w}}$, we can optimize the sampling trajectories using the reflection operator $\mathcal{M}^{\mathrm{w}}_{\mathrm{inv}}(\mathcal{M}^{\mathrm{s}}(\cdot))$, as outlined in Algorithm 1. Through the iterative integration of strong model denoising and weak model inversion, we achieve a step-by-step reflective process during sampling process, refining the latent variable x_t into an improved $\tilde{x}_t$.

Importantly, the choice of $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$ significantly affects the direction of improvements effects. We summarize some promising weak-to-strong model pairs in Table 7 of Appendix C.3. And in Section 3 we present extensive application analyses and experiments, highlighting the powerful capabilities and flexibility of the proposed framework.

In Theorem 1, we demonstrate that W2SD refines latent variable x_t toward the direction defined by the estimated weak-to-strong difference $\Delta_1(t)$. As shown in Figure 3, when the weak-to-strong difference closely bridges the strong-to-ideal difference (i.e., $\Delta_2(t) - \Delta_1(t)$ is small), the reflection mechanism of W2SD drives x_t closer to the ideal x_t^{gt}. The visualization analysis in Section 2.2 validates the correctness of our theory, and we provide a detailed proof in Appendix E.1.

Algorithm 1 W2SD

Input: Strong Model $\mathcal{M}^{\mathrm{s}}$, Weak Model $\mathcal{M}^{\mathrm{w}}$, Total Inference Steps: T, optimization steps: λ
Output: Clean Data x_0
Sample Gaussian noise x_T
for $t = T$ **to** 1 **do**
  if $t > T - \lambda$ **then**
    #W2SD with Reflection
    $\tilde{x}_t = \mathcal{M}^{\mathrm{w}}_{\mathrm{inv}}(\mathcal{M}^{\mathrm{s}}(x_t, t), t)$
  end if
  $x_{t-1} = M^{\mathrm{s}}(\tilde{x}_t, t)$
end for

Theorem 1 (Theoretical Understanding of W2SD) *Suppose x_t is the latent variable at time t, let p_t^{s} and p_t^{w} denote the probability density estimates derived from $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$, respectively. The reflective operator $\mathcal{M}_{\mathrm{inv}}^{\mathrm{w}}(\mathcal{M}^{\mathrm{s}}(\cdot))$ refines x_t to $\tilde{x}_t$ as*

$$\tilde{x}_t = x_t + \sigma^{2t}\Delta t(\nabla_{x_t}\log p_t^{\mathrm{s}}(x_t) - \nabla_{x_t}\log p_t^{\mathrm{w}}(x_t)), \tag{1}$$

where $\Delta_1(t) = \nabla_{x_t}\log p_t^{\mathrm{s}}(x_t) - \nabla_{x_t}\log p_t^{\mathrm{w}}(x_t)$ represents the weak-to-strong difference between $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$ at time t.

It is noted that Karras et al. (2024) proposed Auto-guidance, which employs a simplistic interpolation approach using a degraded model version to refine latent variables. While Equation (1) shares a similar high-level concept, W2SD introduces fundamentally distinct mechanisms, delivering significantly stronger performance and broader applicability. Comprehensive quantitative and qualitative analyses are provided in Appendix F.3.

2.2 Visualization and Explanation in Various Settings

In this subsection, we validate the theory presented in Section 2.1 using both synthetic Gaussian mixture data and real-world image data, providing intuitive visual evidence.

1-D Gaussian Mixture Data We begin by analyzing the 1-D Gaussian data scenario. In Figure 2, the diffusion model is designed to generate data with two distinct peaks at "-4" and "4". Adjusting the proportion of these two peaks in the training dataset, we obtain $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$. Although both models effectively generate samples near the right peak, their performance differs significantly for the left peak.

In Figure 4, we visualize the denoising trajectories under three different settings (weak model, strong model, W2SD). Through the reflective operator $\mathbf{M}_{\mathrm{inv}}^{\mathrm{w}}(\mathcal{M}^{\mathrm{s}}(\cdot))$, the latent variable is progressively drawn closer to the left peak. In contrast, for both strong and weak models, the generated samples are predominantly concentrated around the right peak.

2-D Gaussian Mixture Data We also visualize the mechanism of W2SD on 2D scenario. In Figure 5 (first row), we modulate the proportion of the training data to obtain $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$. W2SD balances the distribution by exploiting the discrepancy between $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$ (the region in the bottom-left corner, denoted as Δ_1) to bridge the unattainable strong-to-ideal difference Δ_2, and boost the chances of sampling toward the bottom-left region. In Figure 5 (second row), we visualize the denoising trajectories under different settings, further validating the effectiveness of W2SD.

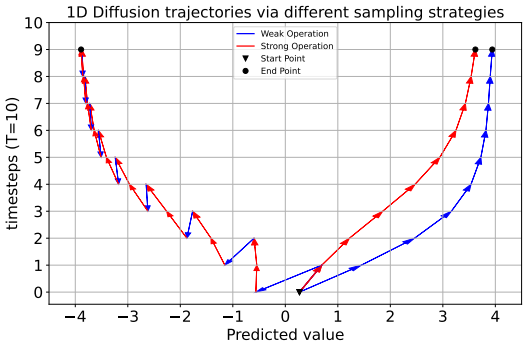

Figure 4: Denoising trajectories across different settings (1-D Gauss). The weak model (blue) generates only right-peak data due to missing left-peak training samples, while the strong model (red) produces data between both peaks. W2SD balances the distribution by leveraging the reflective operator $\mathcal{M}_{\mathrm{inv}}^{\mathrm{w}}(\mathcal{M}^{\mathrm{s}}(\cdot))$.

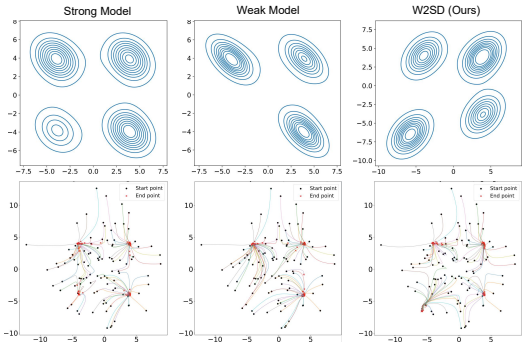

Figure 5: Probability contour plot and denoising trajectories across different settings (2-D Gauss). W2SD balances the learned distribution, bringing it closer to the real data distribution.

Real Image Data Furthermore, we investigate the performance of W2SD on real image data. For ease of analysis, we select two classes from CIFAR-10 (Krizhevsky et al., 2009): **car** and **horse**. Specifically, we train the weak and strong diffusion models on distinct datasets. For $\mathcal{M}^{\mathrm{s}}$, the dataset consists of 5,000 horse images and 2,500 car images. For $\mathcal{M}^{\mathrm{w}}$, it comprises 5,000 horse images and 500 car images. In this scenario, due to the imbalance in training dataset, both $\mathcal{M}^{\mathrm{w}}$ and $\mathcal{M}^{\mathrm{s}}$ are more inclined to generate horses. Moreover, since $\mathcal{M}^{\mathrm{w}}$ is trained on a limited number of car images, it rarely generates cars.

In Figure 6, we note that W2SD can help balance the image categories, enhancing inference process to increase the probability of generating cars. In Figure 7, we also perform a t-SNE dimensionality reduction (Van der Maaten & Hinton, 2008) on the CLIP features of the generated images. W2SD effectively disentangles the representations of "car" and "horse" in the 2D space. Notably, the ratio of "horse" to "car" under W2SD approaches 1:1. In contrast, applying the negative reflection operator $\mathcal{M}^{\mathrm{s}}_{\mathrm{inv}}(\mathcal{M}^{\mathrm{w}}(\cdot))$ worsens the data imbalance, validating the effectiveness of our method.

Figure 6: Qualitative results of W2SD based on dataset differences (CIFAR-10). Our method enhances the probability of generating "cars" and promote a more balanced generation distribution.

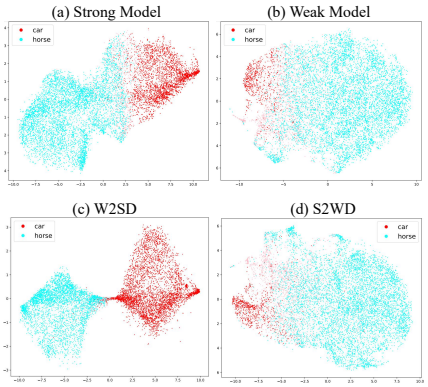

Figure 7: The CLIP feature corresponding to the generated image ($32\times32\times3$) is projected into a 2D space.

3 Empirical Analysis

In this section, we justify our design choices in Section 2 and illustrate the wide applicability of W2SD across diverse combinations of $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$.

Due to page limitations, here we validate the effectiveness of W2SD using Pick-a-Pic Dataset (Kirstain et al., 2023). In Appendix C, we provide a detailed description of the experimental settings. And in Appendix D, we conduct more extensive quantitative and qualitative results across diverse benchmarks (e.g., Drawbench (Saharia et al., 2022)) and modalities (e.g., video generation), demonstrating the broad applicability of our method.

3.1 Weight Difference

The capability differences between models can be directly captured by their parameter weights. And W2SD leverages this weight difference to empirically estimate the weak-to-strong difference, enabling effective reflective operations.

Full Parameter Fine-tuning We first select the full parameter fine-tuned models (e.g., DreamShaper, Juggernaut-XL) that align more closely with human preferences as $\mathcal{M}^{\mathrm{s}}$, and the corresponding standard models as $\mathcal{M}^{\mathrm{w}}$ (e.g., SD1.5 (Rombach et al., 2022), SDXL (Podell et al.)). We evaluate our method with various metrics, as shown in Table 1, W2SD based on weight difference shows significant improvements in human preference metrics such as HPS v2 (Wu et al., 2023b) and PickScore (Kirstain et al., 2023).

LoRA Mechanism Furthermore, W2SD is also applicable to efficiently fine-tuned model. we select the personalized models derived through the LoRA mechanism (Hu et al.) as $\mathcal{M}^{\mathrm{s}}$, and employ

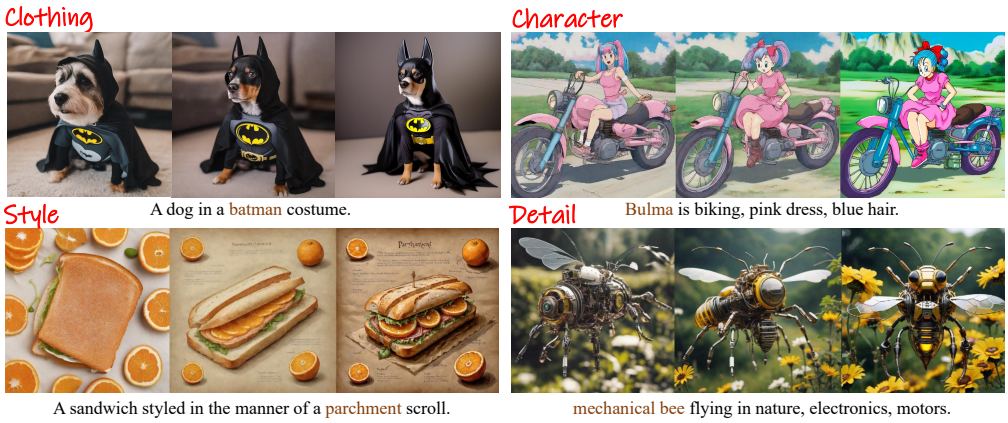

Figure 8: Qualitative comparisons with weak model (**left**), strong model (**middle**) and W2SD based on weight difference (**right**). Our method utilizes the differences between chosen strong and weak models (e.g., high-detail LoRA vs. standard model) to deliver improvements in various dimensions. We provide detail settings in Appendix D.2 (weight difference).

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SD1.5	24.9558	5.5003	20.1368	-
DreamShaper	30.1477	6.1155	21.5035	46.8705
W2SD	**30.4924**	**6.2478**	**21.5727**	**53.1304**
SDXL	29.8701	**6.0939**	21.6487	-
Juggernaut-XL	31.6412	5.9790	22.1903	45.7397
W2SD	**32.0992**	6.0712	**22.2434**	**54.2634**

Table 1: Quantitative results of W2SD based on a full parameter fine-tuning strategy. Our method generates results better aligned with human preferences. Datasets: Pick-a-Pic.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
SD1.5	27.47	52.08	20.14
Personalized LoRA	48.03	64.37	25.99
W2SD	**51.58**	**68.04**	**27.66**

Table 2: Quantitative results of W2SD based on personalized LoRA model. Here, the weight difference between $\mathcal{M}^w$ (SD1.5) and $\mathcal{M}^s$ (SD1.5 +LoRA) biases the generated results towards a more personalized direction.

W2SD to attain a more robust and customized personalization effect. We test 20 LoRA checkpoints across object, person, animal, and style categories, with details in Appendix C.3. Table 2 demonstrates that W2SD results in significant improvements across multiple personalization metrics (e.g., Clip-T and Clip-I). And we present the qualitative results in Figure 8, with more visual cases provided in Appendix D.2.

Method	IS ↑	FiD ↓	AES ↑	HPS v2 ↑
DiT-MoE-S	45.4437	15.1032	4.4755	20.0486
W2SD	**55.5341**	**9.1001**	**4.5053**	**22.3225**

Table 3: Quantitative results of W2SD based on MoE Mechanism. Datasets: ImageNet 50K.

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SD1.5	24.9558	5.5003	20.1368	42.1101
W2SD	**25.5069**	**5.5073**	**20.2443**	**57.8903**
SDXL	29.8701	6.0939	21.6487	43.9425
W2SD	**31.2020**	**6.0970**	**21.7980**	**56.0608**

Table 4: Quantitative results of W2SD based on guidance difference. Datasets: Pick-a-Pic.

MoE Mechanism We also note that a weak-to-strong difference can be induced by controlling the expert selection strategy within the MoE mechanism. Specifically, we focus on DiT-MoE (Fei et al., 2024), a novel architecture that integrates multiple experts and selects the two highest-performing experts during each denoising step to optimize generation quality. Based on this framework, we use DiT-MoE-S as $\mathcal{M}^{\mathrm{s}}$ and define $\mathcal{M}^{\mathrm{w}}$ as the two lowest-performing experts in each denoising step, thereby establishing a quantifiable weight difference.

In Table 3, we show that W2SD reduces FiD (Seitzer, 2020) from 15.1032 to **9.1001**, aligning the learned distribution closer to the real data distribution. Additionally, as shown in Figure 15, although the limited capacity of MoE-DiT-S (with only 71M active parameters) often results in image degradation and distortion (the first row in Figure 15), the application of W2SD significantly enhances the image quality.

3.2 CONDITION DIFFERENCE

Here we demonstrate that the weak-to-strong difference applies not only across different weights but also within the same diffusion model under the weak-to-strong conditions.

Semantics of Prompts We demonstrate that the semantic differences within the condition prompts themselves can be leveraged to enhance generation quality through the reflection process.

Prompt Type	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
Raw Prompt	25.3897	5.4454	20.7144	-
Refined Prompt	28.5698	5.7714	21.6350	45.7719
W2SD	**29.4023**	**5.8812**	**21.8053**	**54.2275**

Table 5: Quantitative results of W2SD based on semantic differences between prompts. Model: SDXL. Datasets: GenEval.

Weight Difference	Condition Difference	HPS v2 ↑	Winning Rate ↑
×	×	31.6412	-
×	✓	32.8217	84%
✓	×	32.0992	76%
✓	✓	**32.9623**	**90%**

Table 6: The improvements effects from different model differences can be cumulative. Datasets: Pick-a-Pic.

Specifically, we select a total of 533 text prompts from the GenEval (Ghosh et al., 2024) for $\mathcal{M}^{\mathrm{w}}$, with an average length ranging from 5 to 6 words and minimal contextual complexity. Additionally, we leverage LLM (here we use Qwen-Plus (Yang et al., 2024)) to enrich and refine these prompts with additional details for $\mathcal{M}_{\mathrm{s}}$. resulting in weak-to-strong prompt pairs.

We report the quantitative results in Table 5, which relies solely on the semantic differences between prompts, achieves improvements across all metrics. And in Figure 16 of Appendix D.2 we illustrate that W2SD is capable of more precisely capturing the intricate details based on the differences in prompt pairs. Notably, Table 4 demonstrates that the classifier-free guidance mechanism can also be applied in W2SD, with detailed information provided in the Appendix D.1.

3.3 SAMPLING PIPELINE DIFFERENCE

Extensive works (Si et al., 2024; Chefer et al., 2023; Zhang et al., 2023) have been devoted to design advanced inference pipelines to improve the quality of diffusion generation results. We demonstrate that W2SD can enhance the inference process by leveraging the difference derived from these powerful existing pipelines. Here we define those enhanced sampling methods as $\mathcal{M}^{\mathrm{s}}$, while $\mathcal{M}^{\mathrm{w}}$ represents the standard sampling (Song et al., 2020; Lu et al., 2022).

We first select ControlNet (Zhang et al., 2023) as $\mathcal{M}^{\mathrm{s}}$, which incorporates additional network structures into the pipeline. By utilizing reference images (e.g., edge maps), it facilitates controllable image generation. As shown in Figure 9, the images generated by W2Sd exhibit a stronger alignment with the provided edge maps. We also present visualization cases of other pipelines (e.g., Ip-adapter (Ye et al., 2023)) as strong models in Figure 17 of Appendix D.2.

3.4 CUMULATIVE EFFECTS OF DIFFERENT MODEL DIFFERENCES

Finally, it is important to note that the improvements effects of W2SD, derived from different type of differences, can be even cumulative. Here we apply the weight difference and the condition difference simultaneously, where $\mathcal{M}^{\mathrm{s}}$ represents the fine-tuned model Juggernaut-XL with high guidance scale at 5.5, while $\mathcal{M}^{\mathrm{w}}$ represents the standard model SDXL with low guidance scale at zero. As shown in Table 6, the combination of guidance difference and condition difference leads to a substantial improvement in Pick-a-Pic Dataset compared to $\mathcal{M}^{\mathrm{s}}$, achieving a winning rate of up to 90% on HPS v2.

4 MORE DETAILED ANALYSIS

In this section, we perform more studies on W2SD, to validate the effectiveness of our theory.

Due to page limitations, further details are provided in: Appendix F.1 analysis the impact of inversion approximation errors on performance improvement; Appendix F.2 analyzes the connections between W2SD, Re-Sampling (lug, 2022), and other inference methods; Appendix F.4 demonstrates that W2SD consistently outperforms standard sampling in terms of time cost, offering high efficiency.

Figure 9: Qualitative results of W2SD based on pipeline difference. We set ControlNet as $\mathcal{M}^{\mathrm{s}}$, DDIM as $\mathcal{M}^{\mathrm{w}}$. W2SD improves alignment with reference images.

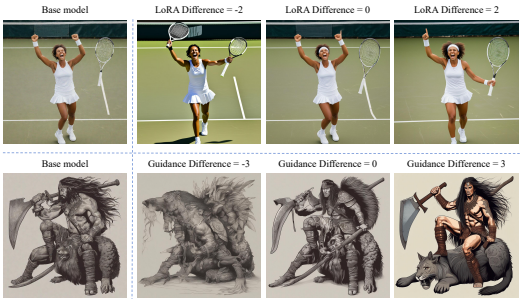

Figure 10: When the weak-to-strong difference is greater than 0, W2SD yields positive gains. When it equals 0, the process degenerates into standard sampling. When it is less than 0, negative gains occurs, resulting in poor image quality.

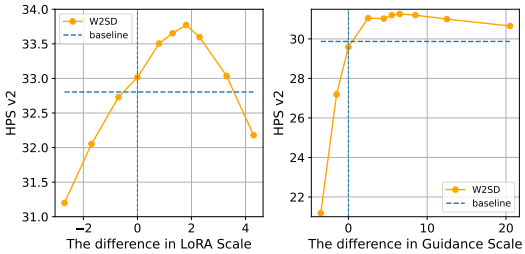

Figure 11: The magnitude of weak-to-strong difference is a key factor impacting the effects of improvements. The horizontal axis shows the magnitude of the weak-to-strong difference, while the vertical axis shows the average HPS v2 on the Pick-a-Pic. When $\mathcal{M}^{\mathrm{s}}$ is weaker than $\mathcal{M}^{\mathrm{w}}$, W2SD results in negative gains.

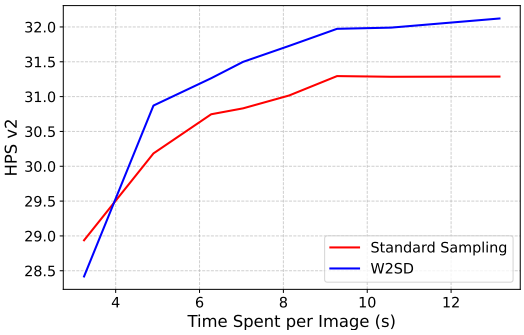

Figure 12: W2SD outperforms standard sampling with identical time costs. The horizontal axis denotes the average generation time per image, the vertical axis represents the HPS v2 on Pick-a-Pic. We provide more details in Appendix F.4.

The Magnitude of Weak-to-Strong Difference Here we explore how the magnitude of weak-to-strong difference affects the improvements effects. To quantify the difference between $\mathcal{M}^{\mathrm{s}}$ and $\mathcal{M}^{\mathrm{w}}$, we first consider the LoRA-based W2SD approach. We fix the LoRA scale of $\mathcal{M}^{\mathrm{s}}$ at 0.8 and adjust the LoRA scale of $\mathcal{M}^{\mathrm{w}}$ (e.g., from -3.5 to 3.5) to observe its impact on the effects of improvements. As shown in Figure 11 (left), when the LoRA scale of $\mathcal{M}^{\mathrm{s}}$ exceeds that of $\mathcal{M}^{\mathrm{w}}$, W2SD improves performance. However, when $\mathcal{M}^{\mathrm{s}}$ is weaker than $\mathcal{M}^{\mathrm{w}}$ (i.e., the LoRA scale difference is negative), the improvements effects diminish or even result in the negative gains. In Figure 10, we present a qualitative analysis showing that the magnitude of the weak-to-strong difference is a key factor that influences the quality of generation results.

5 Conclusion

To the best of our knowledge, this work is the first to systematically integrate the weak-to-strong mechanism into the inference enhancement of diffusion models. We theoretically and empirically understand that the estimated weak-to-strong difference can effectively bridge the strong-to-ideal difference, enhancing the alignment between the learned distributions from existing diffusion models and the real data distribution. Building on this concept, we propose W2SD, which utilizes the estimated difference in density gradients to optimize sampling trajectories via reflective operations. W2SD demonstrates its effectiveness as a general-purpose framework through its cumulative performance improvements, flexible definition of weak-to-strong model pairs, and efficient performance gains with minimal computational overhead.

REFERENCES

Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024.

Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. *arXiv preprint arXiv:2412.10891*, 2024.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2025.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *Forty-first International Conference on Machine Learning*.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Kasper Green Larsen and Martin Ritzert. Optimal weak to strong learning. *Advances in Neural Information Processing Systems*, 35:32830–32841, 2022.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. *arXiv preprint arXiv:2311.16424*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Mikael Møller Høgsgaard, Kasper Green Larsen, and Martin Ritzert. Adaboost is not an optimal weak to strong learner. In *International Conference on Machine Learning*, pp. 13118–13140. PMLR, 2023.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024.

Aaron Lou and Stefano Ermon. Reflected diffusion models. In *International Conference on Machine Learning*, pp. 22675–22701. PMLR, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Robert E Schapire. Explaining adaboost. In *Empirical inference: festschrift in honor of vladimir N. Vapnik*, pp. 37–52. Springer, 2013.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, August 2020. Version 0.3.0.

Shitong Shao, Zikai Zhou, Lichen Bai, Haoyi Xiong, and Zeke Xie. Iv-mixed sampler: Leveraging image diffusion models for enhanced video synthesis. *arXiv preprint arXiv:2410.04171*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Masashi Sugiyama, Takafumi Kanamori, Taiji Suzuki, Marthinus Christoffel Du Plessis, Song Liu, and Ichiro Takeuchi. Density-difference estimation. *Neural Computation*, 25(10):2734–2775, 2013.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023a.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *arXiv preprint arXiv:2409.15761*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024.

Zikai Zhou, Shitong Shao, Lichen Bai, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. *arXiv preprint arXiv:2411.09502*, 2024.

A Related Work

In this section, we review existing works relevant to W2SD.

Weak-to-Strong Mechanism The concept of improving weak models into strong models originates from the AdaBoost (Høgsgaard et al., 2023), which constructs a more accurate classifier by aggregating multiple weak classifiers. Building upon this, Green Larsen & Ritzert (2022); Høgsgaard et al. (2023) introduced a provably optimal weak-to-strong learner, establishing a robust theoretical foundation for this weak-to strong paradigm. In the field of LLMs, several studies (Chen et al.; Burns et al., 2023) have utilized weak models as supervisory signals to facilitate the alignment of LLMs. This paradigm of weak-to-strong generation during training has similarly been investigated in the context of diffusion model training (Chen et al., 2025). And Sugiyama et al. (2013) recommended directly estimating the density difference between weak and strong models instead of separate estimations. In this work, we propose W2SD, a general framework that, for the first time, implements weak-to-strong enhancement in diffusion inference through a reflective mechanism.

Diffusion Inference Enhancement The study of inference scaling laws in diffusion models has recently become a prominent focus within the research community (Ma et al., 2025; Ye et al.; Liu et al., 2024). This line of work can be traced back to Re-Sampling (lug, 2022), which iteratively refines latent variables by injecting random Gaussian noise, effectively reverting the noise level to a previous scale. This iterative paradigm has been utilized in subsequent works, including universal conditional control (Bansal et al., 2023), video generation (Wu et al., 2023a), and protein design (Jumper et al., 2021), to enhance inference performance. However, it has primarily been treated as a heuristic trick, with its underlying mechanisms remaining underexplored. Z-Sampling (Bai et al., 2024) extended this paradigm by replacing random noise injection with inversion operations and identified the guidance difference between denoising and inversion as a critical factor. This phenomenon has also been validated in subsequent studies (Zhou et al., 2024; Shao et al., 2024; Ahn et al., 2024). In our work, we systematically unify these inference enhancement methods, demonstrating that their essence lies in approximating the strong-to-ideal difference via the weak-to-strong difference, and integrate them into a unified reflective framework, W2SD, through theoretical and empirical analysis.

B Preliminaries

In this section, we present preliminaries about denoising and inversion operation in diffusion models (Song et al.; Ho et al., 2020). Due to page limitations, we introduce the related work in Appendix A.

Given the random Gaussian noise z_t, we denote the forward process of diffusion models as $x_t = x_{t-\Delta t} + \sigma^t \sqrt{\Delta t} z_t$, where $t \in [0, 1]$, and σ represents the predefined variance. We denote the ground truth density of real data distribution as p_0^{gt}, After noise addition at time t, the resulting density is represented as p_t^{gt}.

Following Song et al. (2020), we can obtain the denoised results $x_{t-\Delta t}$ from noisy data x_t through the process of an ordinary differential equation as

$$x_{t-\Delta t} = \mathcal{M}(x_t, t) \tag{2}$$

$$= x_t + \sigma^{2t} s_\theta(x_t, t)\Delta t, \quad t \in [0, 1]. \tag{3}$$

where $s_\theta(\cdot, \cdot)$ represents the trained score network, utilized to predict the score at the time t. Similarly, we can invert Equation (3) to transform $x_{t-\Delta t}$ back to a new $\tilde{x}_t$ as

$$\tilde{x}_t = \mathcal{M}_{\mathrm{inv}}(x_{t-\Delta t}, t), \tag{4}$$

$$= x_{t-\Delta t} - \sigma^{2t} s_\theta(x_t, t)\Delta t, \tag{5}$$

$$\approx x_{t-\Delta t} - \sigma^{2t} s_\theta(x_{t-\Delta t}, t)\Delta t, \quad t \in [0, 1]. \tag{6}$$

In practice, we often approximate the score value predicted at time t with time $t - \Delta t$ along the inversion process, i.e., $s_\theta(x_t, t) \approx s_\theta(x_{t-\Delta t}, t)$ in Equation (6). Given that the approximation

error is negligible and the same socre netowrk s_θ, $\mathcal{M}$ and $\mathcal{M}_{\text{inv}}$ can be treated as mutually inverse mappings, thereby satisfying $\mathcal{M}_{\text{inv}}(\mathcal{M}(x_t, t), t) = x_t$. In Appendix F.1, we conduct a detailed analysis of the impact caused by this approximation error.

C EXPERIMENT SETTINGS

In this section, we introduce the details of hyperparameters, metrics and datasets used in the experiments.

C.1 HYPERPARAMETERS

Weight Difference For the W2SD based on weight difference, we consider full-parameter tuning, LoRA-based efficient tuning, and the MoE mechanism.

For the full-parameter tuning, we first set SD1.5 (Rombach et al., 2022) as the weak model and the fine-tuned DreamShaper v8 as the strong model, which demonstrates superior performance in terms of human preference and achieves high-quality generation results. Similarly, we also use SDXL (Podell et al.) as the weak model and Juggernaut-XL as the strong model to further evaluate W2SD's performance under weight differences.

For the efficient tuning, we distinguish strong and weak models by adjusting the LoRA scale. First, we use the xlMoreArtFullV1 LoRA checkpoint to test W2SD's ability to enhance overall image quality. Additionally, to validate the ability of W2SD to improve personalization, we select a series of personalized LoRAs, as detailed in Appendix C.3. For the strong model, the LoRA scale is set to 0.8, while for the weak model, it is set to -1.5.

Following the default settings (Bai et al., 2024), we set the denoising steps $T = 50$, and the guidance scale to 5.5. Reflection operations steps $\lambda = T - 1$. Notably, to eliminate influence from guidance differences, the guidance scale of M^{s} and M^{w} are both set to 1.0 during reflection, ensuring the guidance difference is zero.

For the MoE (Mixture of Experts) mechanism, we select DiT-MoE-S (Fei et al., 2024) as the strong model, which routes the top 2 optimal experts out of 8 during the inference process. The weak model, in contrast, is configured to route the 2 least optimal experts out of 8. Following the default settings, the denoising steps $T = 50$, and the guidance scale is set to 1.5.

Condition Difference In the W2SD research based on condition differences, we focus on analyzing the differences caused by two mechanisms: guidance scale and prompt semantics.

By adjusting the guidance scale to distinguish the strong and weak model, we adapt the same settings as Z-Sampling (Bai et al., 2024): the guidance scale of $\mathcal{M}^{\text{s}}$ was set to 5.5, and the guidance scale of $\mathcal{M}^{\text{w}}$ is set to 0. The diffusion model used is SDXL.

For semantic differences, we set $\mathcal{M}^{\text{w}}$ to use the GenEval prompt, which is short (4-5 words), ambiguous, and coarse, often resulting in uncontrolled outputs. In contrast, $\mathcal{M}^{\text{s}}$ uses refined prompts enhanced by QWen-Plus (Yang et al., 2024), providing greater detail and semantic richness. During the reflection process, the guidance scale was set to 1.0 to eliminate the influence of guidance differences on the results.

Similar to the weight difference setup, we set $T = 50$, $\lambda = T - 1$, and the denoising guidance scale to 5.5.

Sampling Pipeline Difference We demonstrate that W2SD can also generalize to capability differences across different diffusion pipelines. Specifically, we select ControlNet (Zhang et al., 2023) as $\mathcal{M}^{\text{s}}$, with the control scale set to the default value of 1.0. The standard sampling pipeline (Song et al., 2020) is chosen as the weak model. Consistent with the weight difference setup, we configure $T = 50$, $\lambda = T - 1$, and guidance scale of 5.5. During reflection operation, the guidance scale for both $\mathcal{M}^{\text{s}}$ and $\mathcal{M}^{\text{w}}$ are set to 0.5 to eliminate the influence of guidance differences.

C.2 METRICS

AES. AES (Schuhmann et al., 2022) is an evaluation metric that assesses the visual quality of generated images by analyzing key aesthetic attributes such as contrast, composition, color, and detail, thereby measuring their alignment with human aesthetic standards.

PickScore. PickScore (Kirstain et al., 2023) is a CLIP-based metric model trained on a comprehensive open dataset containing text-to-image prompts and corresponding real user preferences for generated images, specifically designed for predicting human aesthetic preferences.

HPS v2. Building upon the Human Preference Dataset v2 (HPD v2), a comprehensive collection of 798,090 human preference judgments across 433,760 image pairs, (Wu et al., 2023b) developed HPS v2 (Human Preference Score v2) through CLIP fine-tuning, establishing a more accurate predictive model for human preferences in generated images.

MPS. MPS (Zhang et al., 2024) is a metric for text-to-image model evaluation, trained on the MHP dataset containing 918,315 human preference annotations across 607,541 images. This novel metric demonstrates superior performance by effectively capturing human judgments across four critical dimensions: aesthetic quality, semantic alignment, detail fidelity, and overall assessment.

C.3 DATASETS

Pick-a-Pic. The Pick-a-Pic dataset (Kirstain et al., 2023), collected through user interactions with a dedicated web application for text-to-image generation, systematically records each comparison with a prompt, two generated images, and a preference label (indicating either a preferred image or a tie when no significant preference exists). Following Bai et al. (2024), we utilize the initial 100 prompts as a representative test set, which provides adequate coverage to assess model performance.

Drawbench. DrawBench (Saharia et al., 2022) is a comprehensive evaluation benchmark for text-to-image models, featuring approximately 200 text prompts across 11 distinct categories that assess critical capabilities including color rendering, object counting, and text generation.

GenEval. Geneval (Ghosh et al., 2024) is an object-focused framework that evaluates image composition through object co-occurrence, position, count, and color. Using 553 prompts, it achieves 83% agreement with human judgments on image correctness.

VBench VBench (Huang et al., 2024) is a comprehensive benchmark for video generation models, featuring a hierarchical evaluation framework across multiple quality dimensions. It supports both automatic and human assessment, with VBench++ extending to text-to-video and image-to-video tasks while incorporating trustworthiness evaluation.

Peronalization Dataset To evaluate the performance gains of W2SD in personalized generation, we selected 20 LoRA checkpoints from the Civitai platform, covering a diverse range of categories, including persons (e.g. Anne Hathaway, Scarlett Johansson), animals (e.g., Scottish Fold cat, prehistoric dinosaur), styles (e.g., Disney style, parchment style), anime characters (e.g. Sun Wukong, Bulma) and objects (e.g. cars).

D SUPPLEMENTARY EXPERIMENTAL RESULTS

In this section, we present more quantitative and qualitative results of W2SD.

D.1 QUANTITATIVE RESULTS

Results of W2SD in other benchmarks To further validate the effectiveness of W2SD, we also conducted experiments on Drawbench (Saharia et al., 2022). For clarity, we have systematically organized and summarized these experiments in Table 7. In Drawbench, we report results based on

Table 7: Different types of model differences lead to improvements effects in different directions.

Model Difference	$\mathcal{M}^{\mathrm{s}}$	$\mathcal{M}^{\mathrm{w}}$	Results
Weight Difference	Finetune Mechanism	SDXL/SD1.5	Tables 1 and 8
	LoRA Mechanism	SDXL/SD1.5	Tables 2, 10 and 11
	Strong Experts (MoE)	Weak experts (MoE)	Table 3
Condition Difference	High CFG	Low CFG	Tables 4 and 9
	Refined Prompt	Raw Prompt	Table 5
Sampling Pipeline Difference	ControlNet	Standard Pipeline (DDIM)	Figure 9
	IP-Adapter	Standard Pipeline (DDIM)	Figure 17

Table 8: Quantitative results of W2SD based on a full parameter fine-tuning strategy. Our method generates results better aligned with human preferences. Datasets: Drawbench.

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SD1.5	25.3601	5.2023	21.0519	-
DreamShaper	28.7845	5.7047	21.8522	47.8813
W2SD	**28.7901**	**5.7847**	**21.9057**	**52.1192**
SDXL	28.5536	**5.4946**	22.2464	-
Juggernaut-XL	28.9085	5.3455	22.4906	47.5648
W2SD	**29.3246**	5.4261	**22.5803**	**52.4358**

Table 9: Quantitative results of W2SD based on guidance difference. Model: SDXL. Datasets: DrawBench.

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SD1.5	25.3601	5.2023	21.0519	47.7075
W2SD	**25.8234**	**5.2157**	**21.2079**	**52.2934**
SDXL	28.5536	**5.4946**	22.2464	40.9590
W2SD	**30.1426**	**5.6600**	**22.4434**	**59.0415**

Table 10: Quantitative results of W2SD based on human preference LoRA model. Our method generates results better aligned with human preferences. Datasets: Pick-a-Pic.

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SD1.5	24.9558	5.5003	20.1368	-
Dpo-Lora	25.5678	5.5804	20.3514	44.2889
W2SD	**26.0825**	**5.6567**	**20.5096**	**55.7106**
SDXL	29.8701	6.0939	21.6487	-
xlMoreArtFullV1	32.8040	6.1176	22.3259	48.2224
W2SD	**33.5959**	**6.2252**	**22.3644**	**51.7770**

Table 11: Quantitative results of W2SD based on human preference LoRA model. Our method generates results better aligned with human preferences. Datasets: DrawBench.

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SD1.5	25.3601	5.2023	21.0519	-
Dpo-Lora	25.8896	5.2895	21.2308	49.3617
W2SD	**25.9431**	**5.3553**	**21.2589**	**50.6399**
SDXL	28.5536	5.4946	22.2464	-
xlMoreArtFullV1	31.2727	5.5487	22.7721	47.0396
W2SD	**32.34857**	**5.7595**	**22.8301**	**52.9588**

weight difference (see Tables 8 and 11) and guidance difference (see Table 9). Additionally, in Pick-a-Pic, we report results based on weight difference (see Table 10). Notably, W2SD demonstrates consistent improvements across all evaluated metrics.

Results of W2SD in Video Generation Task We also validate the performance of W2SD on video generation task to demonstrate its broad applicability. We randomly select 200 prompts from VBench (Huang et al., 2024) as test prompt cases and focus on analyzing the AnimateDiff (Guo et al.) as video generation model.

For the strong model $\mathcal{M}^{\mathrm{s}}$, we employ AnimateDiff with Juggernaut-XL using guidance scale of 3.0, while the weak model $\mathcal{M}^{\mathrm{w}}$ utilizes AnimateDiff with SDXL at guidance scale of 1.0. This setup introduces both weight and guidance differences, meeting the conditions required by W2SD.

In Table 12, W2SD achieves significant improvements across different dimensions such as Subject Consistency, Background Consistency, Aesthetic Quality and mage Quality, confirming its effectiveness in video generation task. Notably, in the Motion Smoothness dimension, W2SD exhibits

Table 12: Quantitative results of W2SD on the video generation task. Model: AnimateDiff. Datasets: VBench.

Method	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Image Quality ↑
AnimateDiff (SDXL)	91.4152%	95.8491%	94.1647%	**45.0000%**	55.3108%	58.4293%
AnimateDiff (Juggernaut-XL)	96.0820%	97.5463%	**96.8834%**	13.0653%	57.7097%	64.1674%
W2SD	**97.1398%**	**97.9386%**	96.8706%	8.0000%	**59.3736%**	**64.6987%**

a performance degradation due to Juggernaut-XL's inherent limitations compared to SDXL. This observation further validates our theory in Section 2.1.

Results of W2SD based on Guidance Scale The classifier-free guidance mechanism (Ho & Salimans, 2021) extrapolates between conditional and unconditional noise predictions, adjusting the guidance scale to control the semantic attributes of the generated images. To apply W2SD, diffusion models under high guidance scale can be viewed as $\mathcal{M}^{\mathrm{s}}$, those under low or even negative guidance scale can be treated as $\mathcal{M}^{\mathrm{w}}$. This guidance scale discrepancy constructs a weak-to-strong difference that aligns the learned generated distribution more closely with the semantic conditions of the prompt.

We note that Z-Sampling (Bai et al., 2024) is a specific instance of W2SD under the guidance difference. In Table 4, our method significantly enhances human preference (e.g., HPS v2) and aesthetic characteristics (e.g., AES) in mainstream models such as SDXL and SD1.5.

D.2 Qualitative Results

Weight Difference In Figure 13, we present visualization results of W2SD based on weight difference. Specifically, to enhance human preference for generated images, we select xlMoreArtFullV1 as the strong model and SDXL as the weak model.

Additionally, by setting the LoRA-based personalized model as strong model and the un-finetuned base model as weak model, Figure 14 showcases the improvement of W2SD in personalized generation effectiveness.

Finally, in Figure 8 we set SDXL as weak model, and we provide the links to the personalized LoRAs mentioned in Figure 8, which are as follows:

- Clothing: Batman 1989.
- Character: Bulma.
- Style: Parchartxl.
- Detail: xlMoreArtFullV1.

Condition Difference In Figure 16, we present the visualization results of W2SD based on condition differences. When the strong model utilizes detailed and semantically rich prompts, while the weak model relies on simple prompts (containing only 4–5 words), W2SD effectively captures fine-grained conditional features. Notably, Z-Sampling (Bai et al., 2024) is a special case of W2SD based on guidance differences, with extensive visual evidence already provided; thus, we omit additional visualizations here.

Sampling Pipeline Difference In Figure 17, we present the visualization results of W2SD based on the differences in the sampling pipeline. By incorporating reference image information through Ip-adapter during the denoising process and employing standard DDIM for inversion, W2SD ensures that the generated image adheres more closely to the given stylistic conditions, resulting in higher-quality results.

Figure 13: Qualitative results of W2SD based on weight differences (human preference). Here we select xlMoreArtFullV1 as the strong model and SDXL as the weak model. W2SD can effectively enhance the performance of human preference.

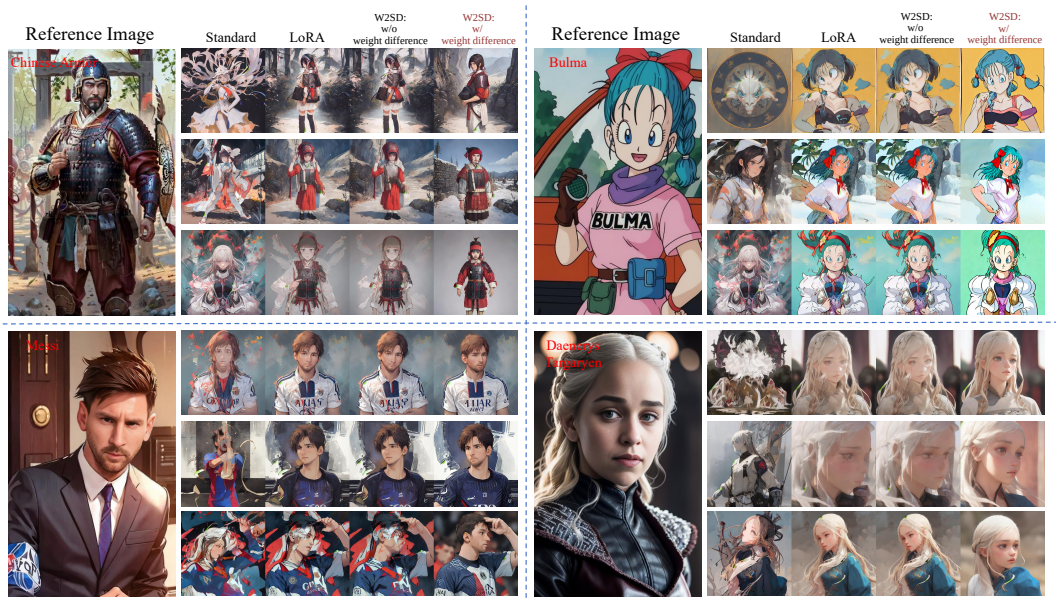

Figure 14: Qualitative results of W2SD based on weight differences (personalization). Here we set LoRA-based personalized model as strong model and the standard model as weak model

Figure 15: Quantitative Results of W2SD Based on the MoE Mechanism. The **first row** shows the results for DiT-MoE-S, while the **second row** presents W2SD. W2SD achieves significant improvements, even with small models featuring 71M activated parameters.

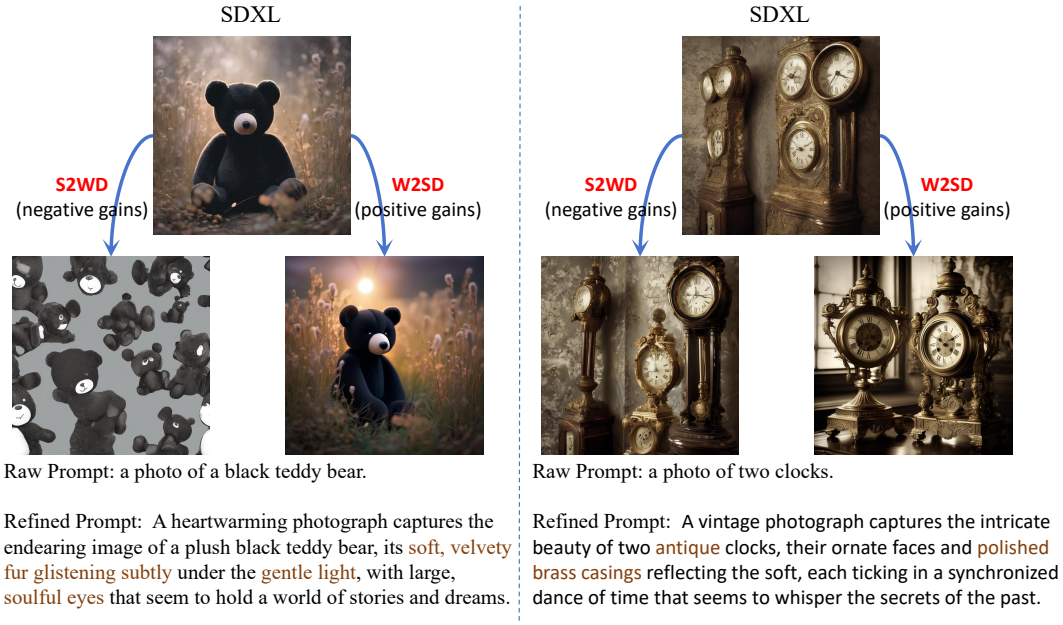

Figure 16: Qualitative results of W2SD based on semantic differences between prompts, which refines the generation process by placing greater emphasis on the fine-grained details.

Figure 17: Qualitative results of W2SD based on sampling pipeline difference. When the strong model employs Ip-adapter and the weak model utilizes DDIM , W2SD can enhance the alignment of the generated results with the reference image (e.g., style).

E PROOFS

In this section, we provide the proofs for Theorem 1 presented in this work.

E.1 PROOF OF THEOREM 1

We first establish the relationship between the latent variable x_t and the refined latent variable $\tilde{x}_t$ through the reflection operation, proving that the core mechanism of W2SD is to optimize x_t in the direction of the weak-to-strong difference.

Proof 1 *At a given time t, the W2SD reflection operation applies the operator $\mathcal{M}_{inv}^{w}\mathcal{M}^{s}(\cdot)$ to the latent variable x_t.*

Specifically, we first denoise x_t using the strong model $\mathcal{M}^{\mathrm{s}}$ according to Equation (3)*, obtaining $x_{t-\Delta t}$ as*

$$x_{t-\Delta t} = x_t + \sigma^{2t} s_\theta^{\mathrm{s}}(x_t, t)\Delta t, \tag{7}$$

where s_θ^{s} denotes the score predicted by $\mathcal{M}^{\mathrm{s}}$, which is equivalent to the gradient of the log probability density $\log p_t^{\mathrm{s}}$, so, Equation (7) *can be rewritten as*

$$x_{t-\Delta t} = x_t + \sigma^{2t} \nabla_{x_t} \log p_t^{\mathrm{s}}(x_t)\Delta t. \tag{8}$$

After obtaining $x_{t-\Delta t}$, we apply the weak model $\mathcal{M}^{\mathrm{w}}$ to invert $x_{t-\Delta t}$ back to the noise level at time t according Equation (6)*, thereby completing the reflection process as*

$$\tilde{x}_t \approx x_{t-\Delta t} - \sigma^{2t} s_\theta^{\mathrm{w}}(x_{t-\Delta t}, t)\Delta t, \tag{9}$$

where s_θ^{w} denotes the score predicted by $\mathcal{M}^{\mathrm{w}}$. Since Δt is typically small, we neglect the approximation error in the diffusion inversion process which implies that $s_\theta^{\mathrm{w}}(x_{t-\Delta t}, t)$ is equivalent to the gradient of the log probability density $\log p_t^{\mathrm{w}}$. Hence we can reformulate Equation (9) *as*

$$\tilde{x}_t = x_{t-\Delta t} - \sigma^{2t} \nabla_{x_t} \log p_t^{\mathrm{w}}(x_t)\Delta t. \tag{10}$$

Combining Equation (8) *and Equation* (10)*, we obtain $\tilde{x}_t$ as*

$$\begin{aligned}
\tilde{x}_t &= x_{t-\Delta t} - \sigma^{2t} \nabla_{x_t} \log p_t^{\mathrm{w}}(x_t)\Delta t && (11)\\
&= x_t + \sigma^{2t} \nabla_{x_t} \log p_t^{s}(x_t)\Delta t - \sigma^{2t} \nabla_{x_t} \log p_t^{\mathrm{w}}(x_t)\Delta t && (12)\\
&= x_t + \sigma^{2t} \Delta t \underbrace{(\nabla_{x_t} \log p_t^{s}(x_t) - \nabla_{x_t} \log p_t^{\mathrm{w}}(x_t))}_{\textit{weak-to-strong difference}\,\Delta_1(t)}. && (13)
\end{aligned}$$

From Equation (13)*, we observe that for the latent variable x_t, the reflection operator $\mathcal{M}_{\mathrm{inv}}^{\mathrm{w}}(\mathcal{M}^{\mathrm{s}}(\cdot))$ in W2SD perturbs x_t along the direction of $\Delta_1(t)$, producing the refined variable $\tilde{x}_t$. When the weak-to-strong difference Δ_1 closely bridges the unattainable strong-to-ideal difference $\Delta_2(t)$, the resulting $\tilde{x}_t$ becomes more aligned with the ground truth ideal distribution.*

F FURTHER ANALYTICAL EXPLORATION

F.1 THE IMPACT OF APPROXIMATION ERROR IN THE INVERSION PROCESS

In Appendix B, we assume that the inversion and denoising process are reversible, meaning that for the same generative model (including the same guidance scale, etc.), $\mathcal{M}_{\mathrm{inv}}(\mathcal{M}(x_t, t), t) = x_t$. However, in practice, the inversion process inevitably introduces errors. Specifically, for x_t at time t, the inversion process actually used in the algorithm implementation is as

$$\begin{aligned}
\tilde{x}_t &= x_{t-\Delta t} - \sigma^{2t} s_\theta(x_{t-\Delta t}, t)\Delta t && (14)\\
&= x_{t-\Delta t} - \sigma^{2t} (\underbrace{s_\theta(x_{t-\Delta t}, t) - s_\theta(x_t, t)}_{\text{Inversion Error } \mathrm{E}_t} + s_\theta(x_t, t))\Delta t, && (15)
\end{aligned}$$

Table 13: As the magnitude of the approximation error in inversion process increases, the gains from W2SD diminish. Model :xlMoreArtFullV1. Datasets: Pick-a-Pic. The type of W2SD: weight difference.

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SDXL	29.8701	6.0939	21.6487	-
xlMoreArtFullV1	32.8040	6.1176	22.3259	48.2224
W2SD (k=0)	**33.5959**	6.2252	**22.3644**	**51.7770**
W2SD (k=0.005)	32.9341	**6.3228**	22.3221	46.7130
W2SD (k=0.010)	19.5277	4.3949	17.9267	1.3440

Table 14: By selecting Gaussian noise during the Re-Sampling process, the advanced Algorithm 3 achieves superior performance in the sampling process, demonstrating that Re-Sampling is a specific instance of W2SD. Model: SDXL. Datasets: Pick-a-Pic. The type of W2SD: guidance difference.

Method	HPS v2 ↑	AES ↑	PickScore ↑	MPS ↑
SDXL	29.8701	6.0939	21.6487	-
Re-Sampling (sim score<0)	30.2797	6.0744	21.6894	48.4793
Re-Sampling (sim score>0)	30.7844	6.0555	**21.8620**	51.5210
W2SD	**31.2020**	**6.0970**	21.7980	**56.0608**

And the effect of W2S on the latent variable in Theorem 1 can also be expressed as

$$\tilde{x}_t = x_t + \sigma^{2t}\Delta t(\nabla_{x_t}\log p_t^{\mathrm{s}}(x_t) - \nabla_{x_t}\log p_t^{\mathrm{w}}(x_t) - \mathrm{E}_t), \tag{16}$$

where Inversion Error E_t represents the approximation error in the inversion process at time t. Since Δt is very small, $s_\theta(x_t, t)$ and $s_\theta(x_{t-\Delta t}, t)$ are assumed to be approximately equal by default, i.e., $\mathrm{E}_t \approx 0$.

To further analyze the impact of E_t on the W2SD, we simplify the analysis by replacing E_t with Gaussian noise of controllable magnitude as

$$\tilde{x}_t = x_t + \sigma^{2t}\Delta t(\nabla_{x_t}\log p_t^{\mathrm{s}}(x_t) - \nabla_{x_t}\log p_t^{\mathrm{w}}(x_t) - k\epsilon). \tag{17}$$

By varying the value of k, we adjust the magnitude of the error E_t to study its effect on the generated results. We analyze W2SD based on weight differences, with the strong model using xlMoreArtFullV1 and the weak model using SDXL.

In Table 13, as k increases, indicating a larger approximation error E_t, the gains from W2SD diminish. This finding is consistent with Bai et al. (2024)'s results, demonstrating that this phenomenon occurs in both guidance difference and weight difference scenarios, indicating that minimizing the approximation error in inversion process is crucial.

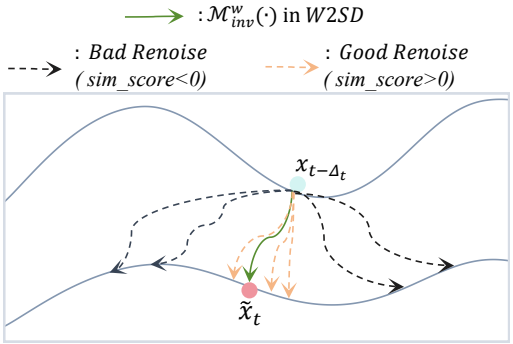

Figure 18: Re-Sampling, which can be considered a specific instance of W2SD, demonstrates improved performance when the randomly sampled Gaussian noise aligns closely with the perturbation vector introduced by the W2SD reflection mechanism (e.g., cosine similarity > 0)

Figure 19: Qualitative results of advanced Re-Sampling demonstrate that the improvements effects vary depending on the strategy used to select Gaussian noise for Re-Sampling. It can be considered a specific instance of W2SD.

F.2 Relationship with Re-Sampling

Re-Sampling (lug, 2022) is the earliest and most classic iterative algorithm of its kind proposed in diffusion models. As illustrated in Algorithm 2, it introduces randomly sampled Gaussian noise into the latent variable x_t, followed by repeated denoising processes.

We confirm that Re-Sampling is generally a specific instance of W2SD, which can be interpreted within the framework of our theory in Theorem 1. In Figure 18, we note that Re-Sampling performs better when the randomly sampled Gaussian noise in Re-Sampling is similar to the perturbation vector introduced by the W2SD reflection mechanism (e.g., cosine similarity > 0). Additionally, we propose an advanced version of Re-Sampling (see Algorithm 3), which evaluates whether the similarity score, **sim_score**$(\epsilon_{\text{w2s}}, \epsilon)$, exceeds 0 to determine whether the Gaussian noise ϵ should be accepted for the ReNoise operation.

We validate the effectiveness of Algorithm 3 in Table 14. When consistently selecting favorable random noise (i.e., $\epsilon_{\text{w2s}}, \epsilon) > 0$), advanced Re-Sampling demonstrates improved performance. Conversely, when consistently selecting unfavorable random noise (i.e., **sim_score**$(\epsilon_{\text{w2s}}, \epsilon) < 0$), the performance of Re-Sampling deteriorates. We also present the visualization results in Figure 19, further demonstrating that Re-Sampling can be incorporated into the W2SD framework, validating the correctness of our theory in Theorem 1.

Similarly, many subsequent research, such as FreeDoM (Yu et al., 2023), UGD (Bansal et al., 2023), MPGD (He et al., 2023), and TFG (Ye et al., 2024), follow the same approach as Re-Sampling by utilizing random Gaussian noise for iterative optimization. Therefore, these inference enhancement methods can be regarded as specific instances of W2SD. The primary distinction among these methods lies in the specific strong model $\mathcal{M}^{\text{s}}$ they employ. For instance, TFG utilizes a more refined parameter search mechanism, resulting in a strong model that exhibits greater robustness and performance compared to algorithms such as FreeDoM and UGD. As a consequence, under the condition of the same weak model (i.e., the addition of random Gaussian noise), TFG demonstrates significantly enhanced performance. However, these studies collectively overlook a fundamental aspect: the weak-to-strong difference constitutes the core principle that fundamentally drives the efficacy of such algorithms.

Algorithm 2 Vanilla Re-Sampling

Input: Strong Model $\mathcal{M}^{\text{s}}$, Total Inference Steps: T, Optimization Steps: λ
Output: Clean Data x_0
Sample Gaussian noise x_T
for $t = T$ **to** 1 **do**
  if $t > T - \lambda$ **then**
    $x_{t-1} = \mathcal{M}^{\text{s}}(x_t, t)$
    Initialize $\epsilon \sim \mathcal{N}(0, 1)$
    $x_t^{\text{Re}} = \texttt{Add_Noise}(x_t, \epsilon, t)$
  end if
  $x_{t-1} = \mathcal{M}^{\text{s}}(x_t^{\text{Re}}, t)$
end for

Algorithm 3 Advanced Re-Sampling

Input: Strong Model $\mathcal{M}^{\text{s}}$, Weak Model $\mathcal{M}^{\text{w}}$, Total Inference Steps: T, Optimization Steps: λ
Output: Clean Data x_0
Sample Gaussian noise x_T
for $t = T$ **to** 1 **do**
  if $t > T - \lambda$ **then**
    $x_{t-1} = \mathcal{M}^{\text{s}}(x_t, t)$
    $\tilde{x}_t = \mathcal{M}^{\text{w}}_{\text{inv}}(x_{t-1}, t)$
    $\epsilon_{\text{w2s}} = \tilde{x}_t - x_t$
    Calculate ϵ_{w2s} based on Equation (1)
    #Select Optimal ReNoise
    Initialize $\epsilon \sim \mathcal{N}(0, 1)$
    while similarity_score $(\epsilon_{\text{w2s}}, \epsilon) < 0$ **do**
      Initialize $\epsilon \sim \mathcal{N}(0, 1)$
    end while
    $x_t^{\text{Re}} = \texttt{Add_Noise}(x_t, \epsilon, t)$
  end if
  $x_{t-1} = \mathcal{M}^{\text{s}}(x_t^{\text{Re}}, t)$
end for

F.3 Relationship with Auto-guidance

We note that in the weak-to-strong framework, Auto-guidance (Karras et al., 2024) employs a pre-trained diffusion model along with a corrupted version of it (typically achieved by adding perturbations or reducing training iterations through training from scratch). It directly enhances performance by interpolating in the latent space. And here we clarify the contributions of W2SD in relation to it.

Figure 20: Compared with Auto-guidance, W2SD achieves superior performance with enhanced robustness, effectively addressing critical issues such as oversaturation and optimization failure.

Different mechanisms W2SD employs an reflection mechanism, while Auto-guidance utilizes an interpolation-based method. For comparison with W2SD, we set w=1 in Auto-guidance. When using Strong Model (human preference model) vs Weak Model (SDXL), in Table 15, we show that W2SD achieves notably higher scores on human preference metrics including PickScore and HPS v2. However, in this setting, direct interpolation in Auto-guidance leads to performance degradation in certain metrics, manifesting as oversaturation and artifacts. Actually, the auto-guidance mechanism similarly utilizes the following operations as

$$x_t \to x_{t-1}^{good} \tag{18}$$

$$x_t \to x_{t-1}^{bad} \tag{19}$$

$$x_{t-1}^{new} = x_{t-1}^{good} + w \cdot (x_{t-1}^{good} - x_{t-1}^{bad}). \tag{20}$$

When w is too large, applying Equation (20) roughly refines the distribution of the latent variables, leading to an unnatural shift in the data distribution (Sadat et al., 2024; Lou & Ermon, 2023). When w is too small, it fails to produce sufficient refinement.

Table 15: Performance comparison between Auto-guidance's interpolation and W2SD's reflection mechanism in latent variable refinement.

Method	HPS v2 ↑	AES ↑	PickScore ↑
SDXL	29.8701	6.0939	21.6487
xlMoreArtFullV1	32.8040	6.1176	22.3259
Auto-guidance	32.1650	6.1187	22.0177
W2SD	**33.5959**	**6.2252**	**22.3644**

In W2SD's process, $x_t \to x_{t-1} \to x_t$, the refinement operation (marked in red) is implicitly performed by score network's internal transformation which avoids common artifacts like distortion and over-saturation, please see Figure 20.

On the other hand, we note W2SD generalizes the concept of weak/strong model pairs—where the "weak" model is not limited to underperforming variants created through reduced capacity or training strategies (e.g., data corruptions or degradations as in Auto-guidance). We propose that differences in semantic interpretation of prompts or sampling methodologies (e.g., MoE routers, ControlNet adaptations) can equally constitute valid weak-strong pairings. This expanded paradigm demonstrates significantly greater practical utility, as it accommodates real-world deployment scenarios where model capabilities vary along multiple axes.

F.4 TIME EFFICIENCY COMPARISON

Assuming standard sampling and W2SD require T_{std} and T_{w2s} denoising steps, respectively, the score predictions needed are T_{std} and $T_{\text{w2s}}+2\lambda$. To ensure a fair comparison, we set $T_{\text{w2s}} = \lfloor \frac{1}{2} T_{\text{std}} \rfloor$ and $\lambda = \lfloor \frac{1}{2} T_{\text{w2s}} \rfloor$, matching the runtime for generating the same image. In this setting, as shown in Figure 12, W2SD consistently outperforms standard sampling, demonstrating that the gains from reflection operations far outweigh the additional computational cost, validating the time efficiency of our method.