Response to Reviewer ktSz

We thank you for your detailed review and for acknowledging the strengths of our work, including:

- Introducing an effective approximate method for addressing multi-condition sampling challenges.
- Demonstrating strong experimental benchmarks that establish the performance benefits of TAG across various settings.

We address your concerns below.

[R1-1] Clarifiacation of Claims:

Thank you for the feedback. We want to clarify that early theoretical bounds and Figure 2a were intended primarily as motivation—illustrating *potential* issues like guidance contributing to divergence—not core claims. The bound specifically illustrated just one scenario where guidance *could* impact divergence even with perfect scores; TAG itself is not designed to address that specific bound, but rather the *general* off-manifold phenomenon arising from various sources. We recognize the limitations of such bounds and that Figure 2a is an extreme case visualization.

Our primary justification rests on **Lemma 3.3** (detailed in [R1-2]), which provides the core theoretical insight into *how* TAG addresses these general deviations via its adaptive temporal correction mechanism. We will revise the manuscript to clarify the limited scope of the initial illustrations, ensuring the main justification clearly rests on Lemma 3.3 and our compelling experimental evidence.

[R1-2] Guidance as a Source of Error:

Thank you for suggesting we highlight literature examples of guidance limitations. Indeed, while powerful, various guidance methods can contribute to off-manifold deviations:

- **Classifier Guidance/CFG:** These established methods can be very effective but often require costly labeled data and classifier training [1, 2] or complex conditional model training for CFG [4]. Furthermore, imperfect classifiers [1, 2] or CFG's approximations (especially at high guidance scales [3, 4]) can still lead to errors.
- **Training-Free Guidance (TFG):** TFG methods emerged partly to bypass the costs associated with CG/CFG by leveraging pre-trained networks. However, applying networks trained on clean data to noisy diffusion states can cause domain mismatch issues, potentially leading to adversarial gradients or off-manifold drift [1, 3].
- **Other Scenarios:** Multi-condition guidance faces challenges in combining signals accurately [5, 6], and few-step sampling introduces discretization errors [7, 8], both increasing off-manifold risk.

TAG aims to mitigate these general deviations across different guidance types. Crucially, it achieves this by adding only an auxiliary time predictor, whose training cost is substantially lower than training full conditional models for CFG or high-performance classifiers. Lemma 3.3 provides the theoretical mechanism, explaining how TAG's Time-Linked Score (TLS) $\nabla_x \log p(t \mid x)$ adaptively amplifies correction when a sample is far from the correct manifold $p_t(x)$, pushing it back towards the high-density region. To provide better intuition on this adaptive correction, we have prepared a new figure illustrating Lemma 3.3's dynamics, which we will add to the revised manuscript (https://xxxx.com/lemma3.3_viz). We will incorporate this discussion into our final manuscript.

[R1-3] Realism of Synthetic Experiments:

We agree the synthetic experiments (Fig. 2a/Sec 3.1, Sec 4.1) are extreme scenarios, primarily for illustrating the off-manifold phenomenon and TAG's corrective mechanism (supported by Lemma 3.3). Our main evidence stems from realistic settings (Sec 4.2, 4.3).

[R1-4] Additional Baselines:

Comparisons with exposure bias mitigation methods are indeed valuable. Techniques like Input Perturbation [9] modify training to improve robustness to inference-time errors, while methods like Epsilon Scaling [10] adjust sampling dynamics. TAG offers a distinct, complementary *inference-time* correction based on temporal consistency, actively rectifying deviations potentially caused by guidance or other factors. We conducted experiments comparing/integrating TAG with such approaches [9, 10, 11] and will include this comparison and discussion on their complementary nature in our final manuscript.

Method	Dataset	Task	Metric	Baseline	+TAG (Ours)	+[9] (Input Pert.)	+[10] (Epsilon Scaling)	+[11] (Shifte Steps)
Exposure Bias Methods	CIFAR-10	Unconditional	FID	TBD	TBD	TBD	TBD	TBD

(Table placeholder for new baseline comparisons)

[R1-5] FID Value:

Thank you for noting the discrepancy in FID scores. As explained in Appendix C.3, the higher FID values reported in some experiments resulted from using fewer evaluation samples (e.g., 512 for CIFAR-10). We initially used this smaller sample size because it was sufficient to observe the relative performance trends between methods and allowed for faster evaluation across the broad range of tasks we investigated, aligning with protocols used in related works such as TFG [14].

However, we understand the importance of reporting scores based on standard, larger sample sizes for broader comparability. We are currently conducting more extensive evaluations using standard sample counts (e.g., 50,000 for CIFAR-10) across all relevant tasks. We expect these results will yield FID scores more consistent with conventional benchmarks and will incorporate these updated results into our final manuscript.

Dataset	Task	Method	Samples	FID (Original 512)	FID (Target 50k)
CIFAR-10	Unconditional	DDPM	512	~11.8	TBD
CIFAR-10	Few-Step (NFE=5)	DDPM+TAG	512	~97.4	TBD
...	...	...	...	...	...

(Table placeholder for updated FID scores with standard sample counts)

[R1-6] Relation to Prior Work:

We appreciate the suggestions and will incorporate detailed comparisons into the revised manuscript:

- **Exposure Bias Methods:** While methods like Input Perturbation [9] address this by modifying *training* data for robustness, others like Epsilon Scaling [10] (output scaling) and Time-Shift Samplers [11] (adjusting time steps) modify *inference* dynamics, similar to TAG operating at inference time. However, their *mechanisms* differ significantly. Epsilon Scaling/Time-Shifting apply heuristics or simple adjustments to existing sampling dynamics to better match training conditions or mitigate mismatch. In contrast, TAG introduces a *distinct, learned guidance term*—the Time-Linked Score (TLS) $\nabla_x \log p(t|x)$—derived from a time predictor. This TLS provides an *active corrective force* specifically aimed at pulling samples back towards the manifold associated with the predicted temporal state t. Lemma 3.3 further details how this correction adaptively strengthens when temporal inconsistency is high. Thus, while addressing related error accumulation issues, TAG employs a unique mechanism based on learned temporal consistency, suggesting potential complementarity with methods focused on training robustness or different inference adjustments.

- **Classification Diffusion Models (CDM):** CDM [12] utilizes a classifier, but one trained to predict the *noise level* of a sample, fundamentally linking it to density ratio estimation and enabling tasks like likelihood evaluation. TAG also uses a classifier (our time predictor), but it's trained to predict the *timestep* t. Furthermore, TAG uses the *gradient* of this predictor's log-probability (the TLS) as a *guidance signal* for sample correction, focusing on temporal alignment. CDM uses its noise-level classifier's output differently, often within its density estimation framework. Thus, despite both

leveraging classifiers, the type of information predicted (noise level vs. timestep) and how that information is utilized in the generative process (density ratio estimation vs. corrective guidance gradient) are substantially different.

- **Energy Diffusion / NCSN:** These score-based methods [13] learn or approximate the score function $\nabla_x \log p_t(x)$. The reverse diffusion process inherently follows this score, and some techniques add Langevin correction steps, which iteratively refine samples using this same score gradient to move them towards higher density regions of $p_t(x)$. TAG's approach differs: it introduces an *additional*, separate correction step using the TLS gradient, $\nabla_x \log p(t|x)$, which is distinct from the primary score $\nabla_x \log p_t(x)$. Theoretically, as detailed in Lemma 3.3, the TLS relies on *relative* probabilities across different timesteps ($p_k(x)/p_{tot}(x)$). This makes TAG's corrective force *adaptive*—it naturally strengthens when a sample is far off the manifold for time t (i.e., $p_t(x)$ is low relative to some $p_k(x)$). This adaptive nature may offer robustness advantages over standard Langevin correction, especially in low-density regions where the primary score $\nabla_x \log p_t(x)$ might be weak or unreliable, but where temporal inconsistency (high $p_k(x)$ for $k \neq t$) could still be detected by the time predictor.

[R1-7] Comparative Experiment on Energy Diffusion / NCSN

Compared to Langevin dynamics per timestep [13], which refines samples using $\nabla_x \log p_t(x)$, TAG uses $\nabla_x \log p(t|x)$. As suggested by analysis of score-based methods, Langevin steps might struggle if the sample is far off-manifold where score estimates are unreliable. TAG's adaptive correction (Lemma 3.3), driven by temporal consistency, is potentially more robust for pulling samples back *towards* the correct manifold from low-density regions. Our experiments support this. We will incorporate this comparative discussion and results into our final manuscript.

Method	Dataset	Task	Metric	Score
Original (Alg 1)	CIFAR-10	Corrupted ($\sigma = 0.2$)	FID	TBD
Langevin Dynamics	CIFAR-10	Corrupted ($\sigma = 0.2$)	FID	TBD

(Table placeholder for Langevin Dynamics comparison)

[R1-7] Update Order in Algorithm 1:

Applying TAG's gradient $\nabla_x \log p_\phi(t \mid x_t)$ as part of the update step to obtain x_{t-1} (as shown in Algorithm 1) is theoretically justified. This guidance term is derived analogously to standard conditional scores (like Eq. 4) using Bayes' rule, formally defined in our work (Def. 3.4), and incorporating such conditional score components within the reverse step calculation is a common practice in the literature.

However, as you insightfully noted, an alternative approach exists: one could first update the current sample x_t using the TAG gradient ($x_t \rightarrow \tilde{x}_t$) and *then* apply the standard reverse diffusion step to obtain x_{t-1} from $\tilde{x}_t$. Intuitively, this might seem appealing as it

aligns the sample with the correct temporal manifold *before* executing the main reverse process step. We tested this alternative experimentally and found minimal performance difference compared to the order presented in Algorithm 1.

Method	Dataset	Task	Metric	Score
Original (Alg 1)	CIFAR-10	Corrupted ($\sigma = 0.2$)	FID	TBD
Alternative Order	CIFAR-10	Corrupted ($\sigma = 0.2$)	FID	TBD

(Table placeholder for Algorithm 1 update order comparison)

References:

[1] Han, D., et al. (2024). Understanding Training-free Diffusion Guidance: Mechanisms and Limitations. *arXiv:2403.12404*.
[2] Kim, J. Google Scholar profile. (Context for classifier limitations).
[3] He, Y., et al. (2024). CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. *arXiv:2406.08070*.
[4] Lin, C. H., et al. (2025). Diffusion Models without Classifier-free Guidance. *arXiv:2502.12154*.
[5] Shi, J., et al. (2023). Language-driven Scene Synthesis using Multi-conditional Diffusion Model. *arXiv:2310.15948*.
[6] Schneuing, A., et al. (2024). Inverse Molecular Design with Multi-Conditional Diffusion Guidance. *arXiv:2401.13858*.
[7] Inference-Time Diffusion Model Distillation. *arXiv:2412.08871*.
[8] Li, J., et al. (2023). On Error Propagation of Diffusion Models. *arXiv:2308.05021*.
[9] Ning, Z., Li, W., He, D., & Zhang, L. (2023). Input Perturbation Reduces Exposure Bias in Diffusion Models. *ICML*. (arXiv:2301.11706)
[10] Ning, Z., Li, W., He, D., & Zhang, L. (2024). Elucidating the Exposure Bias in Diffusion Models. *ICLR*. (arXiv:2308.15321)
[11] Li, Z., Liu, J., & Zhang, L. (2024). Alleviating exposure bias in diffusion models through sampling with shifted time steps. *Advances in Neural Information Processing Systems*.
[12] Le, N. A. K., Nguyen, T., & Tran, A. T. (2024). Classification Diffusion Models: Revitalizing Density Ratio Estimation. *NeurIPS*. (arXiv:2402.10095)
[13] Song, Y., & Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. *NeurIPS*. (arXiv:1907.05600)
[14] Ye, H., et al. (2024). TFG: Unified training-free guidance for diffusion models. arXiv:2409.15761.

Response to Reviewer fuXE

Thank you for your constructive comments and for recognizing the strengths of our work:

- Presenting TAG as a novel classifier-based guidance method to mitigate off-manifold errors.
- Delivering a comprehensive experimental evaluation across multiple models and tasks.
- Acknowledging the significance of studying the role of timestep in diffusion sampling.
- Noting the paper is well-written and generally easy to follow.

We address your concerns below.

[R2-1] Theoretical Claim:

Thank you for your careful reading and pointing out the potential misinterpretation of Theorem 3.1's conclusion. We agree that stating the divergence *is* large based solely on the upper bound is inaccurate. An upper bound doesn't necessitate a large value, only that it *could* be influenced by the guidance term v.

Our *intention* with this theorem was primarily illustrative and motivational, similar to our clarification in response [R1-1]. It was meant to depict *one potential scenario*—where external guidance might contribute to divergence even assuming perfect scores—which falls under the broader category of off-manifold phenomena TAG addresses. TAG itself is not designed specifically for this bound but rather targets the *general* issue of off-manifold deviations arising from various sources.

The core theoretical justification for *how* TAG functions and why it is effective lies in **Lemma 3.3**, which details the adaptive correction mechanism via the Time-Linked Score (TLS), $\nabla_{\mathrm{x}} \log \mathrm{p}(\mathrm{t}|\mathrm{x})$. This mechanism, supported by strong empirical results, forms the primary basis for our claims. We will revise the manuscript to clarify Theorem 3.1's limited, illustrative scope and ensure our justification centers on Lemma 3.3 and the experiments. We will also include a visualization of Lemma 3.3's dynamics (https://xxxx.com/lemma3.3_viz) for better intuition.

[R2-2] Comparison with Related Methods (TCS, TSG, SG):

We appreciate the references to TCS [1], TSG [2], and SG [3]. Based on literature descriptions:

- **TCS [1]:** TCS focuses on adjusting the *perceived* time by reassigning samples to timestep $\widetilde{\mathrm{t}}$ based on the predictor's output. TAG, instead, uses the gradient $\nabla_{\mathrm{x}} \log \mathrm{p}(\mathrm{t}|\mathrm{x}_{\mathrm{t}})$ (TLS) to *directly modify* the sample via additive guidance, actively pushing it towards the manifold $\mathrm{p}_{\mathrm{t}}(\mathrm{x})$.
- **TSG [2] & SG [3]:** These methods perturb the time *input* t to the diffusion model [2, 3]. TAG does *not* alter the time input for score prediction $\nabla_{\mathrm{x}} \log \mathrm{p}_{\mathrm{t}}(\mathrm{x})$; it adds the separate TLS guidance term. This provides a theoretically motivated (Lemma 3.3) soft constraint based on temporal consistency, distinct from altering the time conditioning itself.

We will incorporate this detailed discussion comparing TAG's unique gradient-based, sample-correcting mechanism against timestep reassignment (TCS) and time input perturbation (TSG/SG) into the revised manuscript, along with the following experimental comparisons:

Method	Dataset	Task	Metric	Score
DPS	QM9	Property (α)	AS	TBD
DPS + TCS [1]	QM9	Property (α)	AS	TBD
DPS + TAG	QM9	Property (α)	AS	TBD
TFG	CIFAR-10	Label Cond.	FID	TBD
TFG + TSG [2]	CIFAR-10	Label Cond.	FID	TBD
TFG + SG [3]	CIFAR-10	Label Cond.	FID	TBD
TFG + TAG	CIFAR-10	Label Cond.	FID	TBD

(Table placeholders for TCS/TSG/SG comparisons)

[R2-3] Modest Empirical Result on Image Tasks:

TAG mitigates guidance side-effects (like off-manifold drift), not enhance guidance capability itself. Our image baseline, optimized TFG [4], exhibits less drift, hence TAG's *additional* gain seems modest. Contrast this with molecule generation (Table 2), where TFG often fails (NaNs), indicating severe drift. TAG rescues these failures, showing high impact when baseline guidance struggles. This aligns with Lemma 3.3: TAG's adaptive correction intensifies when needed most. Thus, TAG's visible impact depends on the severity of off-manifold issues in the baseline setup.

[R2-4] Discrete Classifier with Continuous-Time Samplers & Inference Steps:

Regarding your question about using a discretely trained time classifier with potentially continuous-time samplers: while many modern diffusion models are formulated with continuous time, their numerical solvers (e.g., for ODEs/SDEs) invariably discretize the integration path into a finite number of steps. Our approach trains the time classifier using these same standard discretizations (e.g., the 1000 steps in a typical DDPM schedule) by minimizing the cross-entropy loss against the discrete timestep indices. Because the classifier operates on the noisy states encountered during this discretized solving process, training it on the corresponding discrete indices is sufficient and aligns well with conventional practices.

For clarity on our experiments, all results reported in the main paper used 100 inference steps, based on models trained with 1000 discrete timesteps. To further validate robustness, we conducted additional experiments varying the number of inference function evaluations (NFE). TAG consistently demonstrated benefits across different

sampling densities (50, 100, 250, and 1000 steps), confirming its effectiveness is not limited to a specific number of steps. We will include detailed results in the revised manuscript.

NFE	Dataset	Task	Metric	Baseline	+TAG (Ours)
50	CIFAR-10	Label Conditional	FID	TBD	TBD
100	CIFAR-10	Label Conditional	FID	TBD	TBD
250	CIFAR-10	Label Conditional	FID	TBD	TBD
1000	CIFAR-10	Label Conditional	FID	TBD	TBD
...	...	...	...	...	...

(Table placeholder for results across varying inference steps)

[R2-5] Relationship Between Classifier Robustness and TAG Performance:

이거 고칠 필요가 있겠는데: 10k로도 충분하다?
Yes, there is a relationship. The performance of TAG improves with a better classifier, as it provides a more accurate estimate of the true Time-Linked Score (TLS). We conducted experiments using different training checkpoints (10k, 20k, 30k training steps) for both a simple CNN predictor and a larger UNet-like architecture (details can be found in Appendix C.4). Our findings indicate that performance generally improved with more training steps. The best results were achieved with the 30,000-step checkpoint using the simple CNN, and importantly, the much larger UNet-like architecture yielded similar performance at 30,000 steps. This suggests that while a reasonably well-trained predictor is beneficial, TAG does not necessarily require an overly complex or large predictor architecture, highlighting a favorable accuracy-cost trade-off. We will include a table detailing these ablation results in the revised manuscript.

Predictor Arch	Training Steps	Dataset	Task	Metric	Score
SimpleCNN	10k	CIFAR-10	Corrupted	FID	TBD
SimpleCNN	20k	CIFAR-10	Corrupted	FID	TBD
SimpleCNN	30k	CIFAR-10	Corrupted	FID	TBD
UNet-like	30k	CIFAR-10	Corrupted	FID	TBD

(Table placeholder for classifier robustness ablation)

[R2-6] High FID Value:

We understand your concern regarding the high FID values reported in some experiments (see also our responses [R1-3], [R4-3]). There are two main contexts for this:

1. **Illustrative Experiments:** In the synthetic examples designed to clearly show the off-manifold phenomenon (Sec 3.1, 4.1), we used extreme conditions (like added noise or repulsive guidance). High absolute FID scores are expected in these scenarios; the key result here is TAG's significant *relative* improvement over the baseline, demonstrating its corrective capability even under stress.
2. **TFG Sample Size:** For some Training-Free Guidance experiments (e.g., Table 2), the reported FIDs are higher than typical benchmarks because we initially used fewer evaluation samples (e.g., 512 for CIFAR-10), following protocols in related TFG literature like Ye et al. [1]. This smaller sample size, while sufficient for observing relative performance trends and enabling faster evaluation across many tasks, can inflate FID scores.

Crucially, across the realistic TFG and few-step generation tasks (Tables 2, 3, 4, 5), TAG *consistently* improves generation quality relative to the respective baselines, as measured by FID/KID and task-specific validity metrics. This consistent relative improvement is the primary indicator of TAG's effectiveness in mitigating off-manifold issues.

To provide scores aligned with standard benchmarks, we are currently running evaluations with larger sample sizes (e.g., 50,000 for CIFAR-10) across all relevant tasks. We will incorporate these updated results, along with qualitative visualizations (https://xxxx.com/qualitative_results), into the final manuscript.

[R2-7] Typos

Thank you for pointing out. We will revise in our final manuscript.

[R2-8] Negative KID value.

As stated in Appendix C.3, the KID values are expressed on a log scale, consistent with the settings in [4]. Therefore, the negative values are correct.

References:

[1] Jung, H., Park, Y., Schmid, L., Jo, J., Lee, D., Kim, B., Yun, S.-Y., & Shin, J. (2024). Conditional synthesis of 3D molecules with time correction sampler. *Advances in Neural Information Processing Systems*, 37.
[2] Sadat, S., Kansy, M., Hilliges, O., & Weber, R. M. (2024). No training, no problem: Rethinking classifier-free guidance for diffusion models. *arXiv preprint arXiv:2407.02687*.
[3] Li, T., Luo, W., Chen, Z., Ma, L., & Qi, G. J. (2024). Self-Guidance: Boosting Flow and Diffusion Generation on Their Own. *arXiv preprint arXiv:2412.05827*.
[4] Ye, H., et al. (2024). TFG: Unified training-free guidance for diffusion models. arXiv:2409.15761.

Response to Reviewer jywh

We thank you for your review and for recognizing the strengths of our work:

- Presenting a simple approach to mitigate the off-manifold phenomenon.
- Acknowledging that our experiments show evidence supporting TAG's effectiveness in various sampling scenarios.

We address your concerns and questions below.

[R3-1] Time Gap Metric Clarification:

It was primarily chosen for convenience.

Our time predictor outputs logits corresponding to the discretized steps, so using the `argmax` index is direct. While we could rescale this index to a normalized time value, assuming standard uniform steps, this wouldn't change the relative meaning of the metric—it's just a difference in scale. We opted for the simpler integer index representation.

Crucially, TAG itself uses the predictor's *gradient*, not the index value, so its mechanism is independent of how we scale the Time Gap metric for reporting.

[R3-2] Novelty Relative to Prior Methods and Comparison with TCS:

We appreciate the comparison with TCS [1]. While both use a time predictor, TAG's novelty is significant:

1. **Mechanism:** TCS as adjusts the *perceived* noise level by *reassigning* samples to timestep $\widetilde{t}$. TAG, uniquely, introduces *gradient-based guidance* using $\nabla_x \log p(t \mid x_t)$ (TLS) to *actively modify* the sample itself, pushing it towards the manifold $p_t(x)$. This direct sample correction is fundamentally different.
2. **Theoretical Grounding (Lemma 3.3):** TAG's TLS gradient provides an adaptive corrective force, strengthening when the sample is further off-manifold. This mechanism based on score differences across time distinguishes TAG from TCS's reassignment strategy.

We will incorporate this detailed discussion contrasting TAG's gradient-based sample correction with TCS's timestep reassignment into the revision, along with empirical comparisons:

Method	Dataset	Task	Metric	Score
DPS	QM9	Property (α)	AS	TBD
DPS + TCS [1]	QM9	Property (α)	AS	TBD
DPS + TAG	QM9	Property (α)	AS	TBD

(Table placeholder for TCS comparison)

Please also see response [R2-2] for comparisons with TSG/SG.

[R3-3] Typos:

Thank you! The caption will be corrected to "DPS vs. DPS+TAG". We will also correct the typos (Line 21, 87, 299).

[R3-4] Comparison with Simple Correction Steps:

This relates to methods like Langevin correction using the score $\nabla_x \log p_t(x)$, mentioned in work like Song et al. [2]. Such steps refine samples near the manifold but can struggle far off-manifold where score estimates might be poor. TAG's TLS gradient $\nabla_x \log p(t|x)$ leverages temporal consistency (Lemma 3.3) and is potentially more robust for pulling samples back *towards* the correct manifold from such low-density regions. Our experiments comparing TAG with Langevin correction under strong guidance support this. We will include this comparative discussion and results in the revised manuscript.

Method	Dataset	Task	Metric	Score
DPS (Strong Guidance)	CIFAR-10	Label Cond.	FID	TBD
DPS + Langevin Corr. [2]	CIFAR-10	Label Cond.	FID	TBD
DPS + TAG (Ours)	CIFAR-10	Label Cond.	FID	TBD

(Table placeholder for Correction Step comparison)

[R3-5] Consistency of Initial Noise Samples:

Yes, all comparisons used identical initial noise samples x_T (via fixed seeds). We will state this explicitly in the revised manuscript.

[R3-6] Visual Comparisons for the Toy CIFAR Experiment (Sec 4.1):

Agreed. We will add qualitative sample grids for this illustrative experiment to the supplementary material (https://xxxx.com/corrupted_cifar_viz) to provide visual context.

[R3-7] Impact of the Time Classifier Network Size:

The predictor size in the toy experiment (Appendix B.1) was not critical; TAG works well with smaller predictors there too. In our main image experiments (Appendix C.4), the effective SimpleCNN predictor is much *smaller* than the UNet diffusion backbone, demonstrating TAG's efficiency. (See [R2-5] for extensive experiments.) We will add an ablation on predictor size for the toy experiment and clarify this point in the revision.

Predictor Size (Toy Exp)	Score Net Size (Toy Exp)	Metric	TAG Result
5-layer MLP (Original)	3-layer MLP	MSE	TBD
2-layer MLP (Smaller)	3-layer MLP	MSE	TBD

(Table placeholder for toy experiment predictor size ablation)

References:

[1] Jung, H., Park, Y., Schmid, L., Jo, J., Lee, D., Kim, B., Yun, S.-Y., & Shin, J. (2024). Conditional synthesis of 3D molecules with time correction sampler. *Advances in Neural Information Processing Systems*, 37.
[2] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*.

Response to Reviewer WSfF

We thank you for your detailed review and appreciate your positive remarks highlighting:

- TAG as an innovative approach using a time predictor–based gradient correction.
- The detailed algorithmic descriptions and thorough experimental analyses.

We address your concerns below.

[R4-1] Realism of the Off-Manifold Phenomenon and Synthetic Experiments:

We appreciate your points regarding the realism of certain experimental setups. We want to clarify the distinct roles of our synthetic/illustrative experiments versus our main results demonstrating practical relevance.

As discussed in our responses to other reviewers ([R1-1], [R1-3], [R2-6]), the toy experiment with repulsive guidance (Sec 3.1) and the corrupted CIFAR-10 experiment with added noise (Sec 4.1) were intentionally designed as illustrative scenarios. Their primary purpose is pedagogical: to clearly demonstrate TAG's corrective mechanism (theoretically supported by Lemma 3.3) under controlled, even extreme, off-manifold conditions. High absolute FID scores are expected in these stress tests; the significant *relative* improvement shown by TAG in these settings highlights its corrective capability.

While these specific *experiments* are illustrative, the underlying *off-manifold phenomenon* they highlight—the deviation of generated samples from the expected temporal manifold $p_t(x)$—is indeed relevant in standard, practical applications. This is because, as discussed in the literature, inherent limitations in guidance and sampling methods can cause such deviations: imperfect classifiers [1], CFG approximations [2], TFG domain mismatch [1], challenges in combining multiple conditions [3, 4], and discretization errors in few-step sampling [5, 6] can all cause generated samples to deviate from the expected manifold $p_t(x)$. (See [R1-2] for further discussion.) TAG addresses these practical deviations.

Our main evidence for TAG's practical effectiveness and relevance lies in **Sections 4.2 and 4.3**. These sections evaluate TAG within standard application scenarios prone to such issues (TFG across various tasks, multi-condition generation, few-step sampling) using established benchmarks, demonstrating consistent improvements. These results confirm TAG's utility in addressing off-manifold problems where they manifest in conventional use

cases. We will incorporate this detailed discussion on practical relevance and add a figure visualizing Lemma 3.3's mechanism (https://xxxx.com/lemma3.3_viz) into the revised manuscript.

[R4-2] Fine-Tuning vs. Training a Time Predictor:

This is a pertinent comparison. We position TAG distinctly from task-specific fine-tuning approaches like ControlNet [7]:

- **Motivation & Applicability:** Training-free guidance methods, where TAG is often applied, were partly motivated by scenarios with unlabelled data where fine-tuning frameworks like ControlNet are not feasible [8, 9]. TAG provides a way to improve guidance robustness in such settings.
- **Cost:** Even when fine-tuning is possible, TAG offers a significant cost advantage. Training our proposed time predictor (a simple CNN architecture suffices, see Appendix C.4) is substantially faster and cheaper than training large ControlNet modules or fine-tuning the entire diffusion backbone [cf. 11, 12].
- **Generalizability & Mechanism:** TAG's potential for broader generalizability, especially to out-of-distribution (OOD) data, stems from its core mechanism. The time predictor primarily learns to map the noise level or variance characteristics of the input sample x_t to the corresponding timestep t. This noise variance (e.g., related to $1 - \bar{\alpha}_\mathrm{t}$ in DDPM) is a fundamental property of the predefined diffusion process itself, largely independent of the specific data content. Consequently, the time prediction task generalizes well. In contrast, ControlNet learns task-specific spatial representations tightly coupled to the training data distribution [7], which may limit its robustness to OOD inputs or conditions [10]. TAG leverages general *temporal information* inherent in the diffusion process, whereas ControlNet relies on learned *spatial conditioning*.
- **Complementarity:** TAG can potentially be integrated with ControlNet. Since TAG focuses on temporal consistency based on noise variance, it could help stabilize ControlNet-guided generation, particularly in challenging OOD scenarios where ControlNet's learned representations might falter, while TAG's variance-based temporal check remains robust.

Therefore, TAG serves as a complementary, low-cost tool offering wider generalizability due to its reliance on fundamental diffusion properties (temporal consistency linked to noise variance) rather than task-specific learned representations. We will include this detailed comparison discussing motivation, cost, mechanism, and generalizability in the revised manuscript.

[R4-3] High FID Scores:

As detailed in [R1-3] and [R2-6], FIDs > 100 occurred mainly in illustrative synthetic tests or due to low evaluation sample counts (following [9]). We are updating results using standard sample counts (e.g., 50k), yielding typical FID scores. The key finding is TAG's consistent *relative improvement* over baselines (Tables 2, 4, 5). We will incorporate

updated tables and qualitative results (https://xxxx.com/qualitative_results) into the revision.

[R4-4] Related Prior works.

Thank you for suggesting MultiDiffusion [14]. MultiDiffusion achieves multi-conditional control, particularly spatial consistency, by generating and then *fusing multiple diffusion paths*. Our approach differs: we handle multiple standard guidance terms and focus on mitigating the increased off-manifold drift using TAG. For efficiency, we propose approximations (Sec 3.3, Appx A.3) using single or unconditional time predictors to apply TAG's corrective temporal gradient ($\nabla_x \log p(t|x)$). Thus, MultiDiffusion uses path fusion primarily for spatial control, while our method uses approximated temporal alignment guidance to maintain manifold adherence under combined guidance signals. We will add this discussion revise this into our final manuscript.

[R4-5] Discrete vs. Continuous Timesteps:

As discussed in [R2-4], TAG's discrete predictor can be used with continuous solvers employing discretization. The gradient $\nabla_x \log p_\phi(t|x)$ is evaluated at these steps.

[R4-6] Integration with RB-Modulation:

Thank you for this excellent question. RB-Modulation [13] provides training-free style personalization and can be viewed as a specific instance within the broader Training-Free Guidance (TFG) framework, akin to related energy-guided or manifold-preserving approaches such as FreeDoM [15] and MPGD [16].

Since TAG's core function is to mitigate general guidance-induced drift by enforcing temporal consistency, it is indeed complementary to methods like RB-Modulation. Integrating TAG could enhance such approaches by ensuring generated samples adhere to the correct data manifold throughout the diffusion process, even under strong style (or other external) guidance, potentially improving overall fidelity and stability.

Our preliminary experiments integrating TAG specifically with an RB-Modulation-like setup have shown consistent improvements, supporting this synergy. These results demonstrate that TAG aligns well with and can enhance TFG approaches like RB-Modulation. This positive interaction, along with the detailed results (see table placeholder below), will be included and discussed further in the final manuscript.

Method	Dataset	Task	Metric	Score
RB-Mod [13]	CelebA	Style Personalization	FID	TBD
RB-Mod + TAG	CelebA	Style Personalization	FID	TBD

(Table placeholder for RB-Modulation integration results)

[R4-7] Typos:

Thank you for pointing out. We will revise in our final manuscript.

References:

[1] Han, D., et al. (2024). Understanding Training-free Diffusion Guidance: Mechanisms and Limitations. *arXiv:2403.12404*.
[2] He, Y., et al. (2024). CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. *arXiv:2406.08070*.
[3] Shi, J., et al. (2023). Language-driven Scene Synthesis using Multi-conditional Diffusion Model. *arXiv:2310.15948*.
[4] Schneuing, A., et al. (2024). Inverse Molecular Design with Multi-Conditional Diffusion Guidance. *arXiv:2401.13858*.
[5] Inference-Time Diffusion Model Distillation. *arXiv:2412.08871*.
[6] Li, J., et al. (2023). On Error Propagation of Diffusion Models. *arXiv:2308.05021*.
[7] Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *ICCV*.
[8] Chung, H., et al. (2022). Diffusion posterior sampling for general noisy inverse problems. *arXiv:2209.14687*.
[9] Ye, H., et al. (2024). TFG: Unified training-free guidance for diffusion models. *arXiv:2409.15761*.
[10] Context for ControlNet generalization limitations (e.g., LooseControl [arXiv:2312.03079], Meta ControlNet [OpenReview]).
[11] Context for training costs - Stable Diffusion API Pricing.
[12] Context for training costs - ControlNet training discussions/estimates.
[13] Rout, L., et al. (2025). RB-Modulation: Training-Free Personalization of Diffusion Models using Stochastic Optimal Control. *ICLR*. (arXiv:2405.17401)
[14] Bar-Tal, O., et al. (2023). MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *ICML*.
[15] Yu, J., Wang, Y., Zhao, C., Ghanem, B., & Zhang, J. (2023). FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model. *CVPR*.
[16] He, Y., Murata, N., Lai, C. H., Takida, Y., Uesaka, T., Kim, D., Liao, W. H., Mitsufuji, Y., Kolter, J Z., Salakhutdinov, R., & Ermon, S. (2024). Manifold Preserving Guided Diffusion (MPGD). *ICLR*.

Table1

Method	Deblur		Super-resolution		CIFAR10		ImageNet	
	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	Acc. ↑	FID ↓	Acc. ↑
DPS	139.7	0.613	139.0	0.614	217.1	57.5	196.9	24.5

DPS + TAG (ours)	**128.9**	**0.570**	**128.3**	**0.572**	**190.4**	**63.2**	**192.2**	22.9
TFG	64.2	0.154	65.5	0.187	114.1	55.8	231.0	14.3
TFG + TAG (ours)	**62.7**	**0.151**	**64.7**	**0.175**	**102.7**	**61.5**	**219.4**	**17.8**
TCS								
Time Step Guidance								
Self-Guidance								