← Go to **ICML 2024 Conference** homepage (/group?id=ICML.cc/2024/Conference)

DITTO: Diffusion Inference-Time T-Optimization for Music Generation

PDF (/pdf?id=z5Ux2u6t7U)

Zachary Novack (/profile?id=~Zachary_Novack1), Julian McAuley (/profile?id=~Julian_McAuley1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1)

Published: 01 May 2024, Last Modified: 01 May 2024 ICML 2024 Conference, Senior Area Chairs, Area Chairs, Reviewers, Publication Chairs, Authors Revisions (/revisions?id=z5Ux2u6t7U) BibTeX CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Verify Author List: I have double-checked the author list and understand that additions and removals will not be allowed after the submission deadline.

Keywords: music generation, controllable music generation, diffusion models, training-free control

Abstract:
We propose Diffusion Inference-Time T-Optimization (DITTO), a general-purpose framework for controlling pre-trained text-to-music diffusion models at inference-time via optimizing initial noise latents. Our method can be used to optimize through any differentiable feature matching loss to achieve a target (stylized) output and leverages gradient checkpointing for memory efficiency. We demonstrate a surprisingly wide-range of applications for music generation including inpainting, outpainting, and looping as well as intensity, melody, and musical structure control – all without ever fine-tuning the underlying model. When we compare our approach against related training, guidance, and optimization-based methods, we find DITTO achieves state-of-the-art performance on nearly all tasks, including outperforming comparable approaches on controllability, audio quality, and computational efficiency, thus opening the door for high-quality, flexible, training-free control of diffusion models. Sound examples can be found at https://icmlanon2024.github.io/web/ (https://icmlanon2024.github.io/web/).

Primary Area: Applications (computational biology, crowdsourcing, healthcare, neuroscience, social good, climate science, etc.)

Position Paper Track: No

Paper Checklist Guidelines: I certify that all co-authors of this work have read and commit to adhering to the Paper Checklist Guidelines, Call for Papers and Publication Ethics.

Financial Aid: novackze@gmail.com

Submission Number: 5107

Filter by reply type | Filter by author | Search keywords...

Sort: Newest First

- = ≡

Everyone | Program Chairs | Submission5107 Authors | Submission5107...

Submission5107 Area... | Submission5107... | Submission5107... | Submission5107...

Submission5107... | Submission5107... | ✖

20 / 23 replies shown

Add: **Withdrawal**

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 16:03 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=wGu3G36YSa)

Rebuttal:

We sincerely thank all the reviewers for their detailed and thoughtful feedback. We are happy to see positive reception of our work and are encouraged by the fact that reviewers find our work to be strong in terms of generality of the method (Reviewers zuhX, eD1b, 4G5f), breadth of experiments (Reviewer eD1b, SGiY, 4G5f, zuhX), methodological intuition (Review eD1b), and overall clarity (Reviewer 4G5f).

Following the constructive feedback and concerns brought up by all reviewers, we below address shared concerns and our responses:

Inclusion of Subjective Listening Study:

As suggested by Reviewers SGiY and zuhX, we have conducted a subjective listening evaluation for DITTO. We set up the listening test as a 3-part MUSHRA-style rating test: we generated 30 random examples across 3 different test settings (Intensity, Outpainting, and Melody, 10 examples for each) using the same text prompt and control condition for each control method. We compare DITTO with FreeDoM and Music ControlNet for intensity and melody control, and with FreeDoM and MD-50 for outpainting. For each triplet of outputs for the given controls, participants were asked to rate the overall quality of the generated music for each output on a 0-100 scale. We recruited 15 participants for the listening study. Below we show (for each control setting and comparison test) the number of wins for DITTO, as well as the t statistic, p value, whether we reject the null hypothesis that the quality of the compared methods are the same given a pairwise two-sided T-test. We additionally show the average score (and standard errors) for each experiment and method.

Control	**Comparison Test**	**# Wins**	**t Statistic**	**p value**	**H_0 Decision**
Intensity	DITTO vs. Music-ControlNet	97	5.69	6.51×10^{-8}	Reject
Intensity	DITTO vs. FreeDoM	107	7.99	3.33×10^{-13}	Reject
Outpainting	DITTO vs. MD-50	116	9.13	4.62×10^{-16}	Reject
Outpainting	DITTO vs. FreeDoM	120	11.23	1.35×10^{-21}	Reject
Melody	DITTO vs. Music-ControlNet	72	0.61	0.54	Fail to Reject
Melody	DITTO vs. FreeDoM	92	4.67	6.79×10^{-6}	Reject

(1/n)

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 16:11 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=TlE8BBsFva)

Rebuttal:

(continued from first part, showing average score from our subjective listening test)

Experiment	Method	Average Score	Standard Error
Intensity	DITTO	72.13	1.78
Intensity	Music-ControlNet	56.22	2.54
Intensity	FreeDoM	51.77	2.18
Outpainting	DITTO	76.85	1.62
Outpainting	MD-50	53.36	2.05
Outpainting	FreeDoM	50.67	1.93
Melody	DITTO	62.85	1.84
Melody	Music-ControlNet	61.45	1.89
Melody	FreeDoM	53.36	2.05

Notably, we find strong preference for DITTO against all baselines for both Intensity and Outpainting and preference over FreeDoM on Melody, while DITTO and Music-ControlNet were rated practically equivalent on Melody. Overall, this provides solid evidence that DITTO has superior or equal quality over SOTA controllable music generation methods.

Discussion of speed costs of controllable music generation methods:

We would like to take the opportunity to clarify concerns about the relative speed of DITTO compared to other methods (aside from our discussion in Section 6.3 against DOODL). Specifically, below we show the time for the main control methods (here for intensity) in terms of the number of GPU hours needed to finetune the base TTM model for controllability and the generation speed (i.e. how long it takes to generate a single sample in seconds).

Method	Finetuning Cost (GPU Hours)	Generation Speed
Base TTM	-	0.612
ControlNet	576	1.456
FreeDoM	0	2.867
DITTO	0	82.192
DOODL	0	206.897

We recognize that DITTO, as an inference-time optimization based method, is slower than training-based methods like ControlNet as well as inference-time guidance methods (which only incur the cost of 1 gradient at each sampling step). ControlNet-like approaches, however, require 100s of GPU hours to finetune for each type of new control target. When we compare against within-class inference-time optimization methods like DOODL, we are still over 2x faster. Furthermore, we believe for many creative music editing and control tasks, our compute speed is reasonable, and note that DITTO will accelerate as diffusion models are able to sample with fewer steps, which we are actively investigating as follow-up work.

(2/n)

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 16:31 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=ZOdWaZ9ybg)

Rebuttal:

(continued from previous part)

Multi-Objective Optimization Metrics:

Given the suggestions by reviewers eD1b and SGiY, we have included formal metrics for the Multi-Objective DITTO experiment (originally mentioned in Appendix H) along with additional baselines. Here, we add the Intensity+Melody experiment (with $\lambda = 1/4$ for Intensity), and compare against FreeDoM for Intensity+Outpainting and Intensity+Structure and both FreeDoM and Music-ControlNet for Intensity+Melody. Results are shown below (here we use the updated FAD backbone as described in the next section):

	Intensity-Outpainting			**Intensity-Structure**				**Intensity-Melody**			
Method	Intensity MSE (↓)	FAD (↓)	CLAP (↑)	Intensity MSE (↓)	Structure MSE (↓)	FAD (↓)	CLAP (↑)	Intensity MSE (↓)	Melody Acc (↑)	FAD (↓)	CLAP (↑)
DITTO	**5.783**	**0.699**	**0.506**	**6.802**	**0.092**	**0.661**	0.432	**7.833**	0.436	0.680	0.405
FreeDoM	23.945	0.705	0.502	21.033	0.304	0.669	**0.490**	21.185	0.198	0.683	**0.494**
Music-ControlNet	-	-	-	-	-	-	-	37.841	**0.452**	**0.604**	0.347

We find that FreeDoM tends to struggle on all controls, while DITTO is able to much more effectively balance the multiple objectives (albeit with some control adherence degradation overall). On the Intensity+Melody task, we find that all methods seem to struggle on matching the reference melodies, which we leave for future work.

(3/n)

Replying to Rebuttal by Authors

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 16:47 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=n8ryNWU0v2)

Rebuttal:

Improving FAD backbone:

Based of Reviewer SGiY's comment that FAD with the standard VGGish backbone does not correlate well with human perception of musical quality [1], we have added FAD using the LAION-CLAP (music) [2] backbone to our evaluation suite, with the updated values shown below:

Baseline TTM: 0.707

Outpaint/Looping:

Method	$o = 1$	$o = 2$	$o = 3$	**Looping**
DOODL	0.719	0.707	0.700	0.750
Naive	0.722	0.716	0.712	0.753
MD	0.733	0.716	0.710	0.749

Method	$o = 1$	$o = 2$	$o = 3$	Looping
MD-50	0.718	0.714	0.705	0.752
GG	0.754	0.738	0.719	0.774
FreeDoM	0.726	0.723	0.715	0.758
DITTO	**0.716**	**0.703**	**0.698**	**0.746**

Inpainting:

Method	gap = 2	gap = 3	gap = 4
DOODL	0.688	0.693	0.696
Naive	0.697	0.705	0.707
MD	0.690	0.694	0.701
MD-50	0.701	0.708	0.711
GG	0.7	0.709	0.717
FreeDoM	0.704	0.709	0.719
DITTO	**0.686**	**0.688**	**0.690**

Intensity/Melody/Structure:

Method	Intensity	Melody	Structure
DOODL	0.695	0.715	0.653
ControlNet	**0.637**	**0.545**	-
FreeDoM	0.673	0.706	0.668
DITTO	0.682	0.699	**0.632**

Notably, with the exception of the FAD improving for Music-ControlNet on Intensity/Melody control and DITTO on Structure control (surpassing FreeDoM), all other relative differences between methods remain consistent with our original results in showing DITTO's ability for both control and editing tasks.

[1] https://arxiv.org/abs/2311.01616 (https://arxiv.org/abs/2311.01616) [2] https://arxiv.org/abs/2211.06687 (https://arxiv.org/abs/2211.06687)

(4/4)

Official Review of Submission5107 by Reviewer eD1b

Official Review Reviewer eD1b 15 Mar 2024, 16:37 (modified: 02 Apr 2024, 22:04)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer eD1b

Revisions (/revisions?id=74GXX7R7Uk)

Summary:

The authors propose a method to impose target attributes on generated audio by optimizing the initial noise latent of an audio diffusion model guided by a differentiable feature-matching loss. By moving through the latent space of a pretrained diffusion model at inference time, the proposed method allows for flexible control of the output without the need for introducing specialized controls at training time. The authors show that the proposed method achieves strong performance in guided music generation while improving on the efficiency of existing inference-time optimization methods.

Strengths And Weaknesses:
Strengths:

- The proposed method is simple and intuitive — the use of gradient checkpointing seems to allow for a pretty straightforward optimization procedure.
- The method is very general, and suggests a number of promising directions for future work
- The use of self-similarity matrices to enforce structural constraints is clever
- The experiments are thorough and cover a strong set of baselines
- An average GPU memory consumption of 5GB during optimization (according to Table 4) seems very reasonable, given that this includes both the diffusion model and vocoder

Weaknesses:

- While I understand that space is limited, I feel that multi-objective optimization is underexplored. In practice, musicians often want to impose constraints along multiple attributes (e.g. key and tempo). It would be nice to see a full experiment on multi-objective optimization rather than just the two examples provided in figures 8-9, both to see the degree to which multi-objective optimization degrades adherence to individual objectives, and to compare DITTO with the baseline methods.
- Based on Table 4, DITTO appears to be very slow, requiring 40+ optimization steps of 1-2s each on an A100 in order to generate 6 seconds of audio. While this may be an improvement over DOODL, it is still seems far from practical for many creative workflows, and at odds with the authors' claim on line 431 that DITTO is "reasonable for practical use."

Questions:
Questions:

- Does the number of sampling steps used during optimization have to match the number of steps used for sampling the final optimized latent? E.g. can DITTO be optimized with 20 sampling steps, and then the final latent sampled with 1000 steps to improve quality? Or does this result in much lower adherence to the constraints?

I also have two suggestions, both of which may fall into the realm of future work:

- DITTO directly optimizes the initial noise latent. Previously, Wen et al. [1] proposed to watermark the outputs of an image diffusion model by adding perturbations to the Fourier-domain representation of the initial noise latent, which provided a number of advantages in the watermarking setting that may or may not be relevant to music generation (e.g. robust recovery of the perturbation upon inverting the sampled output). Have the authors considered optimizing the Fourier-domain representation of the initial noise latent? Because the Fourier transform and its inverse are differentiable, this may be easy to implement.
- For multi-objective optimization, gradient-balancing the losses as proposed in [2] might be worth exploring as an alternative to putting scalar multipliers on the individual losses. This requires only the final gradient with respect to the optimized parameters (i.e. the noise latent), and so should be unaffected by the checkpointing scheme.

[1] https://arxiv.org/abs/2305.20030 (https://arxiv.org/abs/2305.20030) [2] https://arxiv.org/abs/2210.13438 (https://arxiv.org/abs/2210.13438)

Limitations:
I think the authors do not adequately address the limitations imposed by DITTO's runtime -- if I understand Table 4 correctly, DITTO requires ~1.5 minutes to generate 6 seconds of audio on an A100, which seems very impractical.

Ethics Flag: No
Soundness: 3: good
Presentation: 4: excellent
Contribution: 3: good
Rating: 8: Strong Accept: Technically strong paper, with novel ideas, excellent impact on at least one area, or high-to-excellent impact on multiple areas, with excellent evaluation, resources, and reproducibility, and no unaddressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 17:12 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=xngEWPOwY9)

Rebuttal:

The authors thank the reviewer for the detailed feedback, and are happy to hear their positive view on the paper as a whole. Below, we address concerns brought up in the review:

> **”It would be nice to see a full experiment on multi-objective optimization...to see the degree to which multi-objective optimization degrades adherence to individual objectives, and to compare DITTO with the baseline methods.”**

We agree with the reviewer that multi-objective optimization is an important use case for controllable music generation (albeit not the main focus of the present work), and thus have included an additional experiment on this. Full details and discussion can be found in the general rebuttal above.

> **”Based on Table 4, DITTO appears to be very slow...While this may be an improvement over DOODL, it is still seems far from practical for many creative workflows, and at odds with the authors' claim on line 431 that DITTO is "reasonable for practical use."”**

The authors recognize the reviewer’s point about the lack of clarity in our discussion of DITTO’s speed, and thus we have drawn explicit attention to DITTO’s runtime limitations among the other baselines in our current updated draft of the paper. In general, DITTO takes 82.192 seconds on average to generate an example, compared with the base model (0.612s), Music-ControlNet (1.456s), FreeDoM (2.867s), and DOODL (206.897s). We direct the reviewer to the general rebuttal above for the full table.

In this way, we find that DITTO represents a functional middle ground total speed-wise between the guidance-based methods (training-free and fast yet can lack expressivity) and the training-based methods (which are fast at inference but require over 500 GPU hours for finetuning). Additionally, as DITTO’s speed is tied primarily to the number of sampling steps used, there are clear ways to accelerate DITTO through using the growing line of work for fast diffusion samplers [1-3], which we are actively pursuing as future research. Based on your suggestion in the next comment, we present an early first step in accelerating DITTO for faster creative workflows.

[1] https://arxiv.org/abs/2211.01095 (https://arxiv.org/abs/2211.01095) [2] https://arxiv.org/abs/2310.04378 (https://arxiv.org/abs/2310.04378) [3] https://arxiv.org/abs/2310.02279 (https://arxiv.org/abs/2310.02279)

(1/n)

Replying to Rebuttal by Authors

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 17:21 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=OBl7yOmu3U)

Rebuttal:

(continued from previous part)

> **”Does the number of sampling steps used during optimization have to match the number of steps used for sampling the final optimized latent?...does this result in much lower adherence to the constraints?”**

We appreciate the reviewer’s question here, and have conducted a first experiment to investigate this idea. Namely, we ran DITTO (for Intensity Control) using 8 sampling steps with DDIM (as this was the fewest steps we could use that still made coherent, though low-quality, outputs), which results in a ~2x speed up per optimization step (45.8s). Then, we finally decoded from the optimized latent with the standard 20 DDIM steps. We show results for both the 8 step and 20 step final output (note that FAD results are using the new CLAP backbone as requested by Reviewer SGiY):

Optimization Steps	Decoding Steps	MSE (↓)	FAD (↓)	CLAP (↑)
8	8	7.098	0.7024	0.44
8	20	22.777	0.6902	0.441

Notably, we do find that there is a drop in control adherence moving from the 8-step to the 20-step generation. We attribute part of this to the inability of our base model to produce quality results in few steps, and leave further accelerations for future work to improve the few-step optimization process.

> **” Have the authors considered optimizing the Fourier-domain representation of the initial noise latent?”**

We thank the reviewer for this interesting idea (as the loss landscape may be easier to navigate in the fourier domain), and are excited about pursuing this thoroughly in future work.

> **”For multi-objective optimization, gradient-balancing the losses as proposed in [2] might be worth exploring as an alternative to putting scalar multipliers on the individual losses.”**

The authors appreciate the reviewer’s idea, and plan to investigate this idea further in hopes that it might speed up and help stabilize the multi-objective optimization process.

(2/2)

Replying to Rebuttal by Authors

Response to Authors

Official Comment Reviewer eD1b 02 Apr 2024, 22:03

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I thank the authors for their detailed reply. The multi-objective and optimization-vs.-decoding results are interesting. I think the paper will benefit from the inclusion of the additional experiments, subjective evaluation, and expanded discussion of inference costs. I have raised my score accordingly.

Official Review of Submission5107 by Reviewer SGiY

Official Review Reviewer SGiY 15 Mar 2024, 02:44 (modified: 03 Apr 2024, 01:00)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer SGiY

Revisions (/revisions?id=WVsgEdTMMO)

Summary:

This paper introduces DITTO, a novel framework for inference-time control of pre-trained text-to-music diffusion models. DITTO allows the manipulation of initial noise latents through backpropagation using any differentiable loss function and employs gradient checkpointing. The method's simplicity and applicability ranges across many different applications such as inpainting, outpainting, looping, and control over intensity, melody, and musical structure. The authors provide extensive experimental results and audio examples online.

Strengths And Weaknesses:

Strengths

- The authors have conducted a broad array of experiments, detailing a significant amount of information. This extensive experimentation underlines the versatility and robustness of DITTO across various music generation tasks, although this somewhat blur the paper’s primary focus and contributions.
- Providing audio examples online for qualitative evaluation is highly appreciated, allowing for a direct assessment of the generated audio's quality and the effectiveness of the control mechanisms.

Weaknesses

- **Lack of Subjective Evaluation:** The absence of subjective evaluation is a critical and the reviewer’s main oversight of this paper. Reliance solely on objective measures (mainly FAD), does not adequately capture the qualitative aspects of audio generation, particularly in terms of maintaining or improving quality through inference-time control.
- **Objective Measure Concerns - FAD**
  - The manuscript should include the generation time for each application, as it significantly impacts the interpretation of outpainting and inpainting results. For instance, if the generation duration is 6 seconds and $o = 3$ for the outpainting task, does this imply the actual generation time is 3 seconds ($6 - o$)? If so, FAD results on different o could be reasoned due to the duration of the reference audio instead of the generation quality. Nonetheless, FAD cannot accurately represent the consistency between the reference and generated audio.
  - From a recent work [1], VGGish turns out to not correlate well with human perception regarding musical quality. The reviewer strongly recommends exploring alternative deep embeddings for FAD or incorporating other evaluation methods, such as listening tests.
  - The rationale behind FAD outperforming the default TTM under control scenarios for *Melody* and *Structure* (Table 3) is unclear. Shouldn't the Default TTM serve as the upper benchmark since it undergoes no interventions? This raises questions about the suitability of MusicCaps as a reference dataset distribution for FAD computation due to its degraded music quality.
- Subsection 6.2. – Isn’t Music ControlNet also capable of gaining control on the *Structure*? The authors could have trained the model for this to report the performance.
- The purpose behind presenting numerous baselines remains unclear. Are they meant to provide insights, or are they merely listed to show worse performance than the proposed approach?
  - Which “baseline” is used for Figure 4?
- The paper does not present objective results for simultaneous control over multiple musical factors, as seen in the Music-ControlNet paper.
- There are no audio samples of other baselines or the original TTM which restricts qualitative comparative evaluation for the reviewer.

Minor Issues

- The reviewer suggests putting “Eq.” front of equation numbers when referring inside the manuscript. In subsection 3.3., it gets confusing with the combination of 1) 2) and (1) (2) being written together.

[1] Gui, Azalea, et al. "Adapting Frechet audio distance for generative music evaluation." ICASSP 2024.

After the rebuttal, I changed the rating from 3. Reject to 7. Accept.

Questions:

- How long does it take to perform inference-time control on a single sample (6 seconds long)? This information is crucial to justify the practicality of this method as an “inference-time” control.
- Is gaining controllability on Intensity really a significant application to showcase? What do the authors aim to demonstrate with this application, given that Intensity can be adjusted through simple audio post-processing?
- Do the authors plan to open-source the pre-trained TTM model or the source code of DITTO? Making the model weights available to the public (similar to MusicGen) could significantly enhance its impact on the community.

Limitations:
The authors have stated potential impact in the section Broader Impacts.

Ethics Flag: No
Soundness: 2: fair
Presentation: 2: fair
Contribution: 3: good
Rating: 7: Accept: Technically solid paper, with high impact on at least one sub-area, or moderate-to-high impact on more than one areas, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

- ═ ≡

Rebuttal by Authors

Rebuttal

✎ Authors (👁 Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

📅 28 Mar 2024, 17:35 (modified: 29 Mar 2024, 05:50)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

📄 Revisions (/revisions?id=xBbcenNBNP)

Rebuttal:
Thank you for your detailed response! Below, we address concerns brought up:

> **”...this somewhat blur the paper’s primary focus and contributions.”**

Thank you for sharing your concerns regarding our primary focus and contributions. We outline our contributions at the end of Section 1 primarily focusing on a novel, training-free framework for controlling pre-trained TTM diffusion models that optimizes the initial noise latents to control the model outputs. Our breadth of experiments are done to support our claims of a “framework” that is applicable to a wide range of tasks. We will adjust our prose in the introduction and Section 3 to clarify.

> **”Lack of Subjective Evaluation”**

The authors recognize the reviewer’s concern about the lack of a subjective listening test in our original draft, and thus have conducted a subjective listening test (please see the overall rebuttal for a detailed description of the study and our results). Overall, we found that DITTO is strongly preferred to other baselines on outpainting and intensity and rated similarly to ControlNet on melody, providing qualitative evidence of DITTO’s strong quality against SOTA baselines.

> **”The manuscript should include the generation time for each application... if the generation duration is 6 seconds and o=3 for the outpainting task, does this imply the actual generation time is 3 seconds (6-o)? If so, FAD results on different o could be reasoned due to the duration of the reference audio instead of the generation quality.”**

We would like to clarify a point of confusion here (which we will elucidate thoroughly in our updated draft). Regarding generation time, all methods produce a total output of 6 seconds, save for outpainting which produces between 9-11s (given the 6s reference + 6s generation with overlaps of 1-3s) and looping which produces 20s. Notably, FAD metrics are *not* calculated between the reference audio and the generated outpainting (as this would not capture consistency as the reviewer noted), but rather between the *entire* output (i.e. reference+outpainting) against the reference distribution. This way, we are able to assess the consistency of the generation, as inconsistent generations relative to reference audio negatively affect overall quality (and thus FAD). While FAD results for different overlaps may be due to the difference in the entire audio, each column in Table 1 is still internally consistent in measuring quality.

(1/n)

-
═
≡

Replying to Rebuttal by Authors

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 17:42 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=ycD6swmZHe)

Rebuttal:

> **”The reviewer strongly recommends exploring alternative deep embeddings for FAD or incorporating other evaluation methods, such as listening tests.”**

We sincerely thank the reviewer for bringing this reference to our attention, and have since added FAD with the LAION-CLAP (music) [1] backbone to our evaluation set, which was shown to correlate much better with quality than VGGish [2]. We direct the reviewer to the overall rebuttal to see the updated FAD tables, and note that in general the relative performance of DITTO against the baselines remained unchanged with DITTO wining on all painting tasks, with the only changes being the FAD improving for Music ControlNet on melody and intensity and DITTO on structure.

> **”The rationale behind FAD outperforming the default TTM under control scenarios...is unclear. Shouldn't the Default TTM serve as the upper benchmark?”**

We would like to note that the trend of controllability methods beating baseline FAD has been observed in past work [3]. Additionally, it is not fully surprising that such behavior could arise: if the controls used are reasonably close to the distribution of implicit controls present in the reference distribution (MusicCaps), then the active inclusion of such controls in the generation process may bring the generation outputs closer to the reference distribution. Investigating this trend further is an interesting avenue for future work, and welcome any suggestions from the reviewer on better reference distributions to compare against.

> **”Isn’t Music ControlNet also capable of gaining control on the Structure?”**

Thank you for raising this important question. To our knowledge, neither Music-ControlNet nor other controllability methods have proposed structure control as we have described it (we claim this as our contribution). If we applied our proposed contribution to Music ControlNet (which would require another >500 GPU hours to finetune), we note that the Music ControlNet conditioning mechanism is focused on pixel-based tensor maps that have approx. time-frequency conditioning power (see Section IIIB of Music-ControlNet). This approx. pixel-wise conditioning is quite different from conditioning on a completely global similarity matrix, even if the self-similarity matrix has the same time resolution. Thus, our hypothesis is that structure control (conditioning on global similarity features) with a mechanism designed for time-frequency localized conditioning may perform poorly.

(2/n)

-
═
≡

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 17:47 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=7ccwciBSeo)

Rebuttal:

> **"The purpose behind presenting numerous baselines remains unclear. Are they meant to provide insights, or are they merely listed to show worse performance than the proposed approach?"**

In this work, we compared against a wide range of baselines to cover the growing space of training [3] and training-free [4-7] control methods for diffusion models in comparison with DITTO. Presenting these baselines also serve to highlight how these methods work for controllable music generation, such as the audible seam issue in some outpainting methods (Section 6.1 paragraph 2), how guidance-based methods like FreeDoM struggle on complex feature extractions (Section 6.2 paragraph 1), and reward hacking with DOODL (Section 6.2 paragraph 2). We thank the reviewer for posing this question, and will make the reasoning for our presentation of numerous baselines more explicit to avoid any confusion in the final draft.

> **"Which "baseline" is used for Figure 4?"**

For this figure, the "baseline" method denotes the MultiDiffusion outpainting approach.

> **"The paper does not present objective results for simultaneous control over multiple musical factors, as seen in the Music-ControlNet paper."**

The authors recognize the reviewer's concern over the need of quantitative metrics for multi-objective optimization (as we only provide the high-level formulation and some examples in Appendix H), and have such included a full experiment on multi-objective optimization. Please see the overall rebuttal for our results and discussion in depth.

> **"There are no audio samples of other baselines or the original TTM which restricts qualitative comparative evaluation for the reviewer."**

We recognize the reviewer's request for baseline audio samples, and have included a set of audio examples between DITTO, Music-ControlNet, FreeDoM, and MD-50 at the following link: https://github.com/anonicml2024/rebuttal-examples (https://github.com/anonicml2024/rebuttal-examples)

> **"Do the authors plan to open-source the pre-trained TTM model or the source code of DITTO?"**

Upon acceptance, the authors plan to release an open source framework for the DITTO control process such that it can be worked into open source diffusion models in the music domain and beyond such as AudioLDM and Stable Diffusion.

(3/n)

Replying to Rebuttal by Authors

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 17:56 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=9QkAHqeTqG)

Rebuttal:

> **”How long does it take to perform inference-time control on a single sample (6 seconds long)? This information is crucial to justify the practicality of this method as an “inference-time” control.”**

In general, DITTO takes 82.192 seconds on average to generate an example, compared with the base model (0.612s), Music-ControlNet (1.456s), FreeDoM (2.867s), and DOODL (206.897s). We direct the reviewer to the general rebuttal above for the full table and a deeper discussion on DITTO’s speed. We find that DITTO represents a functional middle ground total speed-wise between the guidance-based methods (training-free and fast yet can lack expressivity) and the training-based methods (which are fast at inference but require over 500 GPU hours for finetuning). Additionally, as DITTO’s speed is tied primarily to the number of sampling steps used, there are clear ways to accelerate DITTO through using the growing line of work for fast diffusion samplers [8-10], which we are actively pursuing as future research.

> **”Is gaining controllability on Intensity really a significant application…given that Intensity can be adjusted through simple audio post-processing?”**

We would like to clarify the use case of intensity control (first proposed in [3]). The reviewer is correct that simple volume can be easily adjusted in post processing, yet like in prior work, intensity control has the effect of controlling much more than just raw volume. Specifically, as intensity curves in the data are often correlated with changes in harmonic and rhythmic density, intensity control has the added effect of modifying the underlying musical content to grow and become busier as the volume grows. Additionally, using a *smoothed* volume curve for intensity effectively ignores fast changing dynamic content (such as the kick drum in electronic music), and thus controlling the intensity within these genres has the qualitative effect of reducing the number of instruments playing with the drums during the song (we point the reviewer to our main webpage for examples of each: https://icmlanon2024.github.io/web/ (https://icmlanon2024.github.io/web/)).

[1] https://arxiv.org/abs/2211.06687 (https://arxiv.org/abs/2211.06687) [2] https://arxiv.org/abs/2311.01616 (https://arxiv.org/abs/2311.01616) [3] https://arxiv.org/abs/2311.07069 (https://arxiv.org/abs/2311.07069) [4] https://arxiv.org/abs/2302.08113 (https://arxiv.org/abs/2302.08113) [5] https://arxiv.org/abs/2303.09833 (https://arxiv.org/abs/2303.09833) [6] https://arxiv.org/abs/2311.00613 (https://arxiv.org/abs/2311.00613) [7] https://arxiv.org/abs/2303.13703 (https://arxiv.org/abs/2303.13703) [8] https://arxiv.org/abs/2211.01095 (https://arxiv.org/abs/2211.01095) [9] https://arxiv.org/abs/2310.04378 (https://arxiv.org/abs/2310.04378) [10] https://arxiv.org/abs/2310.02279 (https://arxiv.org/abs/2310.02279)

(4/4)

Replying to Rebuttal by Authors

Official Comment by Reviewer SGiY

Official Comment Reviewer SGiY 03 Apr 2024, 01:00

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I applaud the authors for their detailed response and the additional evaluations conducted during the rebuttal session. The authors have addressed the primary concern regarding subjective evaluation and have also carried out additional assessments (generation speed, FAD with LAION-CLAP, and multi-task control).

While the proposed method does have some potential limitations, such as generation time and the need for more comprehensive evaluation across applications, I believe these are inevitable challenges when introducing a new task; perhaps raising further questions about a new task can be seen as a positive development that encourages future research.

Given these considerations, I change my initial rating towards acceptance. I trust that the authors will revise the final version of the paper to address the major issues discussed during the rebuttal session. Although these revisions might require significant changes to the paper, I am confident that the authors recognize their importance and will implement the necessary modifications, as demonstrated during the rebuttal.

Official Review of Submission5107 by Reviewer 4G5f

Official Review Reviewer 4G5f 11 Mar 2024, 10:36 (modified: 21 Mar 2024, 05:24)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 4G5f

Revisions (/revisions?id=MnRZX5dq2u)

Summary:
This paper proposes a novel method to control a diffusion model at inference time using any differentiable loss defined on the data space. It consists in optimizing the initial noise in the diffusion sampling process directly, meaning that ones at to differentiate through the sample path. This is made possible via the use of checkpointing. If the proposed method is general to be applicable to any diffusion model trained on raw data, the authors focus on music generation and showcase impressive results on many musically-meaningful control tasks (inpainting, outpainting, volume control, melody control, structure control) in the accompanying website. It is worth noting that tasks like structure control are highly non-trivial. The objective metrics demonstrate that the proposed method leads to faster optimization times, better memory usage and better adherence to the controls.

Strengths And Weaknesses:
The paper is clear, detailed and interesting from start to finish. The related works section and the explanation on how the proposed method differ from previous approaches are exhaustive and insightful. It is an important achievement that showcases the practical usability of differentiating through the sampler. The flexibility of the approach and the relevance of the musical applications make this article an important contribution that may be of interest for a wide community.

No real weakness can be pinpointed.

Questions:

- Availability of the code
- MOS in Table 4 may be misleading as we immediately read Mean Opinion Score instead of Mean Optimization Speed.

Limitations:
It would be interesting to indicate if the code will be shared as the optimization procedure may be complex to implement: this could hinder the wide adoption of this method.

Ethics Flag: No
Soundness: 4: excellent
Presentation: 4: excellent
Contribution: 4: excellent
Rating: 9: Very Strong Accept: Technically flawless paper with groundbreaking impact on at least one area of AI/ML and excellent impact on multiple areas of AI/ML, with flawless evaluation, resources, and reproducibility, and no unaddressed ethical considerations.
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 17:58 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=SmH1Sb2BCB)

Rebuttal:

We sincerely thank the reviewer for their response and appreciate their positive review of our work! Below, we address some of the questions brought up:

> **”It would be interesting to indicate if the code will be shared as the optimization procedure may be complex to implement: this could hinder the wide adoption of this method.”**

Upon acceptance, the authors plan to release an open source framework for the DITTO control process such that it can be worked into open source diffusion models in the music domain and beyond such as AudioLDM and Stable Diffusion. The authors would like to note that from the engineering side, implementing DITTO is actually quite simple (as it is mostly occurring outside of the main sampling logic besides the model checkpointing), and hope that our future open-source implementation can be of wide use for the community.

> **”MOS in Table 4 may be misleading as we immediately read Mean Opinion Score instead of Mean Optimization Speed.”**

Thank you for bringing this source of confusion to our attention. We will update to “Average Optimization Speed (AOS)” and will incorporate our updated speed metrics addressed above.

Official Review of Submission5107 by Reviewer zuhX

Official Review Reviewer zuhX 07 Mar 2024, 03:08 (modified: 02 Apr 2024, 23:57)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer zuhX

Revisions (/revisions?id=qbXV5P3CBB)

Summary:

This paper introduces DITTO, an optimization strategy of the initial noise (xT) in the reverse process that enables editing or controlling generation of a pre-trained diffusion model. DITTO was compared to a variety of possible diffusion controlling algorithms in the context of text-to-music generation, and showed remarkable editing ability and controllability as a training-free framework.

Strengths And Weaknesses:

Strengths:

- **Generality**: The proposed DITTO could extend any pre-trained diffusion model to editing and controlling tasks with improved performance. The authors demonstrated its practical values on text-to-music generation task. In principle, DITTO is also applicable to other audio diffusion models, or even on other modalities.
- **Demonstrated on a wide range of tasks**: The authors tested DITTO on a series of music generation/edition tasks. As far as I have listened, the in-painting and out-painting performances are nice.

Weakness:

- **Presentation**: I found some flaws in the main paper that hinders readability, but these are not fatal.
    - I find some notations in the Figure 2 confusing. For example, what is the meaning of dashed lines? I can only see that it is used to separate two forms of computing (presumably, for denoting forward and backward process?). Similarly, what does ∇ mean in this Figure? Clearer presentations are needed, to prevent the readers from guessing "what that means".
    - At first glance, I had also been confused about the values on Table 1, 2 & 4, though I later spotted "FAD" on the first column of header. In an usual setting, the header should depict the category of the respective column. Yet, "FAD" is obviously not describing the first column, which consists of different approaches. For a better presentation, I suggest to explicitly describe the metric (FADs) in the caption, and changing "FAD" to "Method" or sth. equivalent.

- **Insufficient discussion of time costs**: The authors spend much effort to compare DITTO against DOODL in terms of memory and speed. However, as mentioned by the authors, there are also other methods that supposedly run much faster than DITTO, e.g., ControlNet. The authors should expose all limitations of this method instead of only emphasizing the strengths.
- **Insufficient references**: There are two seminal TTM diffusion works that I think this paper should also refer to: [1] and [2].
  - [1] Musika! Fast Infinite Waveform Music Generation
  - [2] Efficient Neural Music Generation
- **Invalid evaluation scheme**: Last but not least, I concern about a fatal issue in this paper -- **the absence of subjective evaluation**. The authors' statements regarding the superiority of DITTO heavily rely on a premise that "higher FAD corresponds to higher quality and better musicality". Yet, in the context of music generation, it remains skeptical whether FAD could really represent musicality (as discussed in the paper of MusicLM and Noise2Music). As a result, I strongly recommend the authors to conduct subjective evaluation for fair and justifiable comparisons. I will increase my rating if this can be done.

Questions:
The questions have been raised above.

Limitations:
The limitations have been mentioned above.

Ethics Flag: No
Soundness: 3: good
Presentation: 3: good
Contribution: 4: excellent
Rating: 7: Accept: Technically solid paper, with high impact on at least one sub-area, or moderate-to-high impact on more than one areas, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 18:04 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=ENbZLrKXZ3)

Rebuttal:
The authors thank the reviewer for their detailed and insightful comments, and have addressed concerns brought up below:

> **”Presentation: I found some flaws in the main paper that hinders readability...I find some notations in the Figure 2 confusing. For example, what is the meaning of dashed lines?...Similarly, what does ∇mean in this Figure?... For a better presentation, I suggest to explicitly describe the metric (FADs) in the caption, and changing "FAD" to "Method" or sth. equivalent.”**

The authors appreciate the reviewers comments and will take their ideas into full account in our updated draft. The dashed lines (along with the ∇’s) were meant to denote the backwards pass of DITTO, as the gradients backpropagate from the feature matching loss to the initial latents. We will provide an explicit legend and better explanation in the caption in order to elucidate this. Additionally, for tables 1,2, and 4, we will change the placement of “FAD” in the table and have a more expressive caption to explain better.

> **”Insufficient discussion of time costs:”**

The authors recognize the reviewer’s point about the lack of discussion on the speed of DITTO, and thus we have drawn explicit attention to DITTO’s limitations in terms of runtime among the other baselines in our current updated draft of the paper. In general, DITTO takes 82.192 seconds on average to generate an example, compared with the base model (0.612s), Music-ControlNet (1.456s), FreeDoM (2.867s), and DOODL (206.897s). We direct the reviewer to the general rebuttal above for the full table and a deeper discussion on DITTO’s speed. We find that DITTO represents a functional middle ground total speed-wise between the guidance-based methods (training-free and fast yet can lack expressivity) and the training-based methods (which are fast at inference but require over 500 GPU hours for finetuning). Additionally, as DITTO’s speed is tied primarily to the number of sampling steps used, there are clear ways to accelerate DITTO through using the growing line of work for fast diffusion samplers [1-3], which we are actively pursuing as future research.

> **”Insufficient references:”**

The authors thank the reviewer for noticing this oversight on our end and will adjust to include these papers in our related works in the updated draft.

(1/n)

Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Nicholas J. Bryan (/profile?id=~Nicholas_J._Bryan1), Zachary Novack (/profile?id=~Zachary_Novack1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1))

28 Mar 2024, 18:08 (modified: 29 Mar 2024, 05:50)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=ozlrjpS9Il)

Rebuttal:

> **”Invalid evaluation scheme:...the absence of subjective evaluation. The authors' statements regarding the superiority of DITTO heavily rely on a premise that "higher FAD corresponds to higher quality and better musicality". Yet, in the context of music generation, it remains skeptical whether FAD could really represent musicality...I strongly recommend the authors to conduct subjective evaluation for fair and justifiable comparisons. I will increase my rating if this can be done.”**

The authors recognize the reviewer’s concern about the lack of subjective evaluation in our original draft, and thus have taken the time to conduct a subjective listening test (please see the overall rebuttal for a detailed description of the study and our results). Overall, we found that DITTO is strongly preferred to other baselines on outpainting and intensity and rated similarly to ControlNet on melody, providing qualitative evidence of DITTO’s strong quality against SOTA baselines.

> **”Ethics Flag: Yes”**

Given no specific comments were made regarding ethics concerns, we believe the reviewer might have selected the checkbox by mistake. If so, please kindly adjust.

If there are real concerns, thank you for sharing. We take ethics issues extremely seriously and address our perspective in our Broader Impacts Section. We note, however, there is no specific comments or further information regarding this concern and as a result we cannot properly respond. We would kindly request you expand on this issue so we can address concerns and update accordingly.

[1] https://arxiv.org/abs/2211.01095 (https://arxiv.org/abs/2211.01095) [2] https://arxiv.org/abs/2310.04378 (https://arxiv.org/abs/2310.04378) [3] https://arxiv.org/abs/2310.02279 (https://arxiv.org/abs/2310.02279)

(2/2)

Replying to Rebuttal by Authors

≡

Official Comment by Reviewer zuhX

Official Comment ✎ Reviewer zuhX 📅 02 Apr 2024, 23:56

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thanks for the detailed response. I can see most of my concerns were addressed. I have also read all the rebuttal and the feedbacks from other reviewers. In general, I agree that the contribution of this work is significant for its generality and superiority reported in extensive experiments. Therefore, I decide to raise my score to 7.

About OpenReview (/about)
Hosting a Venue (/group?id=OpenReview.net/Support)
All Venues (/venues)
Sponsors (/sponsors)

Frequently Asked Questions (https://docs.openreview.net/getting-started/frequently-asked-questions)
Contact (/contact)
Feedback
Terms of Use (/legal/terms)
Privacy Policy (/legal/privacy)

OpenReview (/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through Code for Science & Society (https://codeforscience.org/). We gratefully acknowledge the support of the OpenReview Sponsors (/sponsors). © 2024 OpenReview