I read the authors' response and am largely satisfied, but I have a few minor remaining suggestions.

We thank the reviewer again for his constructive suggestions. We believe that we have addressed the additional concerns by adding a more comprehensive overview of global evaluation metrics and by conducting another large set of experiments that investigate the role of the pre-processing transformations for precipitation in the performance of generative models. Please find point-by-point responses in the following.

Basic metrics / sanity checks: Even though global metrics like RMSE are not very informative for climate downscaling, they are still useful as basic sanity checks. Please report RMSE (and possibly a couple of simple metrics like bias) for all baselines and for your model, so readers can see a coarse quantitative comparison.

We thank the reviewer for this suggestion. While Mean Absolute Bias and Pearson Correlation were already integrated into our analysis, we have now added a comprehensive performance summary, that now also includes the RMSE, for all model climatologies in Table S1 of the Supplementary Information.
As expected for global metrics, Quantile Mapping performs just slightly better in terms of Mean Absolute Bias and RMSE. This is because QM is mathematically designed to minimize local grid-point distribution errors. However, these metrics do not account for spatial structure and variability. As shown in our PSD analysis (Fig. 5A) and individual samples (Fig. 3), the Diffusion Model significantly outperforms QM in restoring physically realistic spatial patterns and high-frequency details which are smoothed out by traditional statistical methods

Conditioning variables: It would be helpful to clarify whether you experimented with additional conditioning variables before deciding to use precipitation only. Modern diffusion based downscaling work often naturally explores richer conditioning (for example, temperature, circulation variables, or large-scale predictors), so briefly reporting any such experiments - even if they did not help - would be informative for the community.

We thank the reviewer for this point regarding multivariate conditioning. In this study, we focused exclusively on precipitation to demonstrate that our conditional diffusion framework can handle highly non-Gaussian, heavy-tailed statistics in an unpaired setting. In this work we did not experiment with adding additional input variables in our experiments and have clarified this in our revised manuscript.

GAN baselines: I share your concerns about some GAN baselines, but it should still be feasible to include at least one reasonably strong GAN baseline. While GANs can be trickier to stabilize, diffusion models are typically more computationally demanding; given that you already train a diffusion model, adding a tuned GAN baseline would strengthen the empirical comparison. That said, it's a non-critical issue.

Thank you, we indeed appreciate the suggestion. As noted in the manuscript, we prioritized stability and mode coverage, which motivated our choice of Diffusion Models over GANs. To provide a strong baseline against other generative deep-learning methods, we compared our method against a VQ-VAE (Vector Quantized Variational Autoencoder), a state-of-the-art Consistency Model (CM) and another diffusion model formulation (EDM) (Section 3, Figs. S2, S5, S6, S22, S23). Crucially, the key conceptual novelty of our study is the embedding space transformation, which allows for the application of any supervised learning method. We have already included a comprehensive set of baselines and comparisons, covering statistical methods, U-Net and Transformer-based architectures, VQ-VAE, CM and an alternative diffusion formulation (EDM). Given that the reviewer notes this addition is non-critical, we believe the current comparison sufficiently contextualizes the performance of our method against other (generative) approaches. We think that an even more exhaustive comparison of generative model architectures lies beyond the primary scope of this work and would be more appropriate for a dedicated machine learning publication.

Data transformations and ablations: Your response on transformations is generic and restates standard practice. Now that it is clear only precipitation is modeled, the chosen pipeline (e.g., log transform + z-scaling + rescaling to [-1,1]) makes more sense, but I still wonder about the effect of different choices, such as z-scaling only, log + z-scaling, and log + z-scaling + rescaling to [-1,1]. Once precipitation is forced into [-1,1], the interpretation of z-scores and any gaussian likelihood assumptions becomes less clear. It would be useful to know whether the model fails to converge, converges more slowly, or converges to a different optimum when omitting the log transform or the [-1,1] scaling. Given the efficient unet style model (small to mid sized, I am assuming) and the moderate number of epochs (100), a small ablation on these preprocessing choices seems feasible. I do not view this as a blocking issue, but demonstrating the impact of these transformations would significantly help readers understand the robustness of the method. While you are at it, also share number of trainable params in your model, and any ML engineering tricks (gradient clipping, outlier trimming in train vs. test, etc.) that can help the readers.

We appreciate this suggestion and agree that demonstrating the impact of preprocessing is valuable for understanding the convergence properties of our method and providing general insights into precipitation preprocessing for machine learning practitioners.

To address this, we have conducted an additional ablation study comparing our proposed preprocessing pipeline (log-transform + standardization + rescaling to [-1, 1]) against:

I.) a variant using only standardization (z-score) and II.) log-transform + standardization. Both experiments required retraining of the model. We added a new Figure (Fig. S28) to the SI, where we compare the three different pre-processings to the ground truth ERA5.

We found that the full preprocessing pipeline (log-transform + standardization + rescaling to [-1, 1]) results in the best alignment with the ground truth (ERA5). As shown in Fig. S28, the variant using only standardization (blue line) tends to underestimate high-frequency variability in the spatial spectrum (Fig. S28A) and exhibits bias in the spatial means. Conversely, the variant using log-transform + standardization without the final rescaling (orange line) consistently overestimates mean precipitation across both latitudes and

longitudes (Fig. S28C and Fig. S28D) as well as the frequency of extreme events (Fig. S28B). Our proposed method (magenta) effectively balances these factors. The addition of the [-1, 1] rescaling mitigates the bias introduced by the log-transform while preserving the spectral fidelity, resulting in the closest match to ERA5 across the metrics.

For the “ML engineering tricks”**,** we have detailed our choices in Section 4.3, including the use of dynamic clipping, noise conditioning augmentation, and Min-SNR weighting (lines 619-620). We now also explicitly report the number of trainable parameters (730.95 million) in Section 4.3 (lines 624).