Consistency-diversity-realism Pareto fronts of conditional image generative models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Building *world models* that accurately and comprehensively represent the real world is a holy grail for image generative models as it would enable their use as world simulators. For conditional image generative models to be successful world models, they should not only excel at image quality and prompt-image consistency but also ensure high representation diversity. However, current research in generative models mostly focuses on creative applications that are predominantly concerned with human preferences of image quality and aesthetics. We note that generative models have inference time mechanisms – or *knobs* – that allow the control of generation consistency, quality, and diversity. In this paper, we use state-of-the-art text-to-image and their knobs to draw consistency-diversity-realism Pareto fronts that provide a holistic view on consistency-diversity-realism multi-objective. Our experiments suggest that realism and consistency can both be improved simultaneously; however there exists a clear tradeoff between realism/-consistency and diversity. By looking at Pareto optimal points, we note that earlier models are better at representation diversity and worse in consistency-realism, and more recent models excel in consistency-realism while decreasing significantly the representation diversity. Overall, our analysis clearly shows that there is *no best model* and the choice of model should be determined by the downstream application. With this analysis, we invite the research community to consider Pareto fronts as an analytical tool to measure progress towards world models.

1 Introduction

Progress in foundational vision-based machine learning models has heavily relied on large-scale Internet-crawled datasets of real images (Schuhmann et al., 2022). Yet, with the acceleration of research on generative models and the unprecedented photorealistic quality achieved by recent text-to-image generative models (Podell et al., 2023; Esser et al., 2024; Ramesh et al., 2022; Saharia et al., 2022), researchers have started exploring their potential as *world models* that generate images to train downstream representation learning models (Astolfi et al., 2023; Hemmat et al., 2023; Tian et al., 2024).

World models aim to represent the real world as accurately and comprehensively as possible. Therefore, visual world models should not only be able to yield *high quality* image generations, but also generate content that is representative of the *diversity* of the world, while ensuring *prompt consistency*. However, state-of-the-art conditional image generative models have mostly been optimized for human preference, and thus, a single high-quality and consistent sample fulfills the current optimization criteria. This vastly disregards representation diversity (Hall et al., 2024; Sehwag et al., 2022; Zameshina et al., 2023; Corso et al., 2024; Hemmat et al., 2023; Sadat et al., 2024), and questions the potential of state-of-the-art conditional image generative models to operate as

Submitted to Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

effective world models. Optimizing for human preferences only partially fulfills the multi-objective optimization required to leverage conditional generative models as world models.

At the same time, state-of-the-art conditional image generative models have built-in inference time mechanisms, hereinafter referred to as *knobs*, to control the realism (also referred to as quality or fidelity), consistency, and diversity dimensions of the generation process. For example, it has been shown that the guidance scale in classifier free guidance of diffusion models (Ho & Salimans, 2021), trades image fidelity for diversity (Saharia et al., 2022; Corso et al., 2024). Similarly, post-hoc filtering (Karthik et al., 2023) is used to improve consistency. Although recent works have carried out extensive evaluations of image generative models (Ku et al., 2024; Lee et al., 2024), these evaluations have been primarily designed from the perspective of creative applications. To the best of our knowledge, a comprehensive and systematic analysis of the effect of the knobs controlling the different performance dimensions of conditional image generative models has not yet been carried out.

In this paper, we benchmark conditional image generative models in terms of the world models' multi-objective. In particular, we perform an optimization over both knobs and state-of-the-art models with the goal of capturing the consistency-diversity, realism-diversity, and consistency-realism Pareto fronts that are currently reachable. In our analysis, we include text-to-image (T2I) models, considering several versions of latent diffusion models (LDM), namely $\text{LDM}_{1.5}$ and $\text{LDM}_{2.1}$ (Rombach et al., 2022), as well as LDM_{XL} (Podell et al., 2023). We perform the core of our analysis using the ubiquitous MSCOCO (Lin et al., 2014) validation dataset. To quantify the multi-objective, we use inter-sample similarity and recall (Kynkäänniemi et al., 2019) to measure representation diversity; image reconstruction quality and precision (Kynkäänniemi et al., 2019) to quantify realism; and the Davidsonian scene graph score (Cho et al., 2024)) to assess prompt-generation consistency.

By drawing the Pareto fronts, we observe that progress in conditional image generative models has been driven by improvements in image realism and/or prompt-generation consistency, and that these improvements result in models sacrificing representation diversity. On MSCOCO, our analysis suggests that more recent models should be used when the downstream task requires samples with high realism – $\text{LDM}_{\text{XL-Turbo}}$– and consistency – LDM_{XL}–. However, older models – $\text{LDM}_{1.5}$ and $\text{LDM}_{2.1}$– are preferable for tasks that require good representation diversity. We believe that the proposed evaluation framework and the findings that arise from it will enable faster progress towards enabling the use of conditional image generative models as world models, and we hope it will encourage the research community to work on models that present softer consistency-diversity-realism tradeoffs.

2 Methodology of the analysis

In this section, we summarize the metrics we use to evaluate conditional image generative models, and describe existing knobs that control the consistency-diversity-realism multi-objective. For a more detailed description of the metrics and knobs adopted, we refer to Appendix D.

Evaluating conditional image generation We evaluate conditional image generation in terms of prompt-sample consistency, sample diversity and realism (also referred to as quality or fidelity in the literature). We consider two complementary ways of quantifying the performance of conditional image generative models: conditional and marginal. On the one hand, *conditional metrics* are prompt-specific scores computed on the set of image generations resulting from a prompt. An overall score may be obtained by averaging out all prompt-specific scores. On the other hand, *marginal metrics* are overall scores computed on the generations resulting from *all* the prompts directly. In practice, marginal metrics compare a set of generated images to a reference dataset while ignoring the prompts used to obtain the sets. In the reminder of this subsection, we define consistency – that is always conditional –, conditional and marginal diversity, as well as conditional and marginal realism.

Consistency-diversity-realism knobs. We steer the generations of conditional generative using two well-known knobs: guidance scale and post-hoc filtering. The guidance scale is a parameter that controls the strength of the conditioning in the denoising process of diffusion models; higher values steer the generation to be more aligned with the textual prompt. Post-hoc filtering first generates multiple samples given the same prompt, but different prior, and then select the ones with the highest consistency to the prompt based on automatic metrics like CLIPScore.

Pareto fronts. We perform an optimization over state-of-the-art models and their knobs with the goal of capturing the consistency-diversity, realism-diversity, and consistency-realism Pareto fronts that are currently reachable, and building understanding on the consistency-diversity-realism multi-objective. More precisely, we quantify consistency, diversity and realism for each pair of (model, knob-value) using the metrics presented in Section **??**. We then leverage all the resulting measurements to obtain the Pareto fronts that capture the optimal consistency-diversity-realism values achieved by current state-of-the-art conditional image generative models. For visualization ease, we transform the multi-objective into three bi-objectives: consistency-diversity, realism-diversity and consistency-realism.

3 Experiments

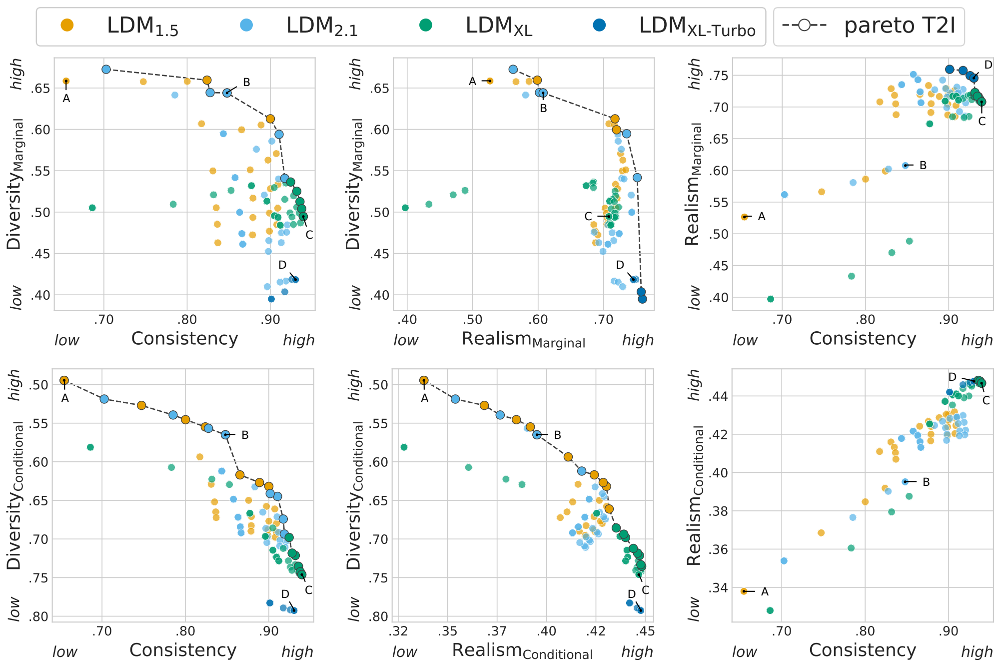

Figure 1: Consistency-diversity (left), realism-diversity (middle) and consistency-realism (right) Pareto fronts for T2I generative models. (top) marginal, (bottom) conditional metrics. Each dot is a configuration of model's knobs. Labeled dots (A-D) are visualized in Fig. 2.

3.1 Setting

Models. We consider different versions of latent diffusion models: $LDM_{1.5}$, $LDM_{2.1}$ (Rombach et al., 2022), LDM_{XL} (Podell et al., 2023)[1], and $LDM_{XL\text{-}Turbo}$ (Sauer et al., 2023). We report the knobs ablated in Appendix. Moreover, in Appendix we extend Pareto fronts to measure the geographical diversity of the same models.

Datasets. We benchmark the models on MSCOCO (Lin et al., 2014; Caesar et al., 2018). In particular, we use the validation set from the 2014 split (Lin et al., 2014), which contains 41K images, to compute the marginal metrics, and the 2017 split (Caesar et al., 2018), which contains 5K images, to compute the conditional metrics. This choice is mostly to limit computational costs, as conditional metrics require multiple samples per conditioning.

3.2 Consistency-diversity-realism multi-objective for text-to-image models

In Fig. 1, we depict consistency-diversity, realism-diversity and consistency-realism Pareto fronts for open source T2I generative models. In particular, Fig. 1 (top) depicts marginal realism and diversity metrics while Fig. 1 (bottom) shows their conditional counterparts. Note that consistency is computed in the same way (DSG) in both figures. We now discuss each of the pair-wise metrics Pareto fronts.

Consistency-diversity. The Pareto fronts in Fig. 1 (left, top and bottom), are composed of three models: $LDM_{1.5}$, $LDM_{2.1}$ and LDM_{XL}. We observe that improvement in diversity, both marginal (Recall) and conditional (DreamSim score), comes at the expense of consistency (DSG). On the one hand, $LDM_{2.1}$ and $LDM_{1.5}$ achieve the best marginal and conditional diversities, respectively. On the other hand, and perhaps unsurprisingly, LDM_{XL} reaches the best consistency ($\geq 95\%$ of DSG accuracy), while $LDM_{1.5}$ and $LDM_{2.1}$ dominate the middle region of the frontier. Moreover, by comparing these

[1]For LDM_{XL} we use the base model v1.0 without the refiner

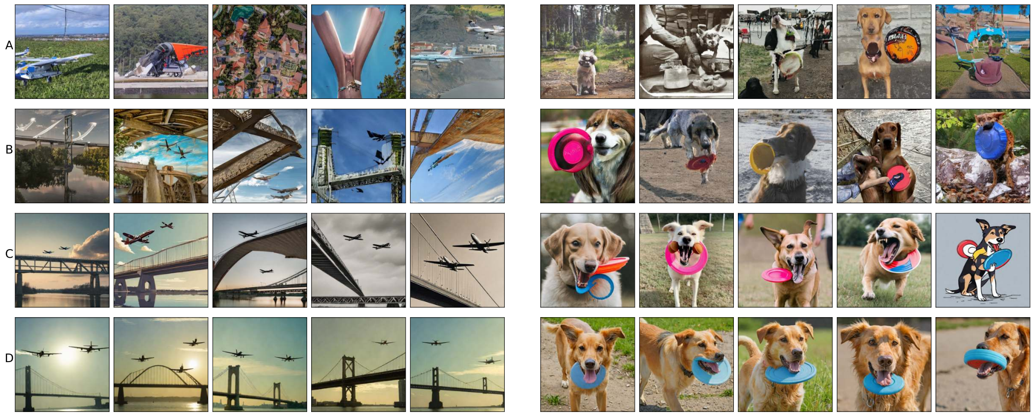

Figure 2: T2I qualitative results on MSCOCO. A-D refer to the models marked in Fig. 1. (left) `Two planes flying in the sky over a bridge.` (right) `There is a dog holding a Frisbee in its mouth.`

two models, we notice that Pareto optimal hyperparameter configurations of $\text{LDM}_{2.1}$ obtain slightly higher consistency scores. In Fig. 2, we validate these observations showcasing samples from $\text{LDM}_{1.5}$ (A) at high-diversity/low-consistency, $\text{LDM}_{2.1}$ (B) from the middle of the frontier, and LDM_{XL} (C) at high-consistency/low-diversity. Both in the case of the "two planes" and of the "dog", the variance of colors and backgrounds are reduced when visual quality is increased. Other samples are in **??**.

Realism-diversity. The marginal realism-diversity (Precision-Recall) Pareto front in Fig. 1 (middle, top), is composed of three models: $\text{LDM}_{1.5}$, $\text{LDM}_{2.1}$ and $\text{LDM}_{\text{XL-Turbo}}$. In this case, we also observe a tradeoff: higher marginal diversity coincides with lower realism for $\text{LDM}_{1.5}$ and $\text{LDM}_{2.1}$. $\text{LDM}_{\text{XL-Turbo}}$ obtains the samples of highest realism. However, we observe that the realism gain compared to $\text{LDM}_{2.1}$ is rather small and leads to a steep decrease in sample diversity. We attribute this drop to the adversarial objective used to distill $\text{LDM}_{\text{XL-Turbo}}$ from LDM_{XL}, as also noted by Sauer et al. (2023). Interestingly, LDM_{XL} does not appear on the Pareto front, and it is even quite far away from it. This is probably due to LDM_{XL} (without refiner) generating smooth images lacking of high frequency details (*e.g.*, see the dog in Fig. 2 and (Podell et al., 2023)), and the marginal metrics, which are computed with InceptionV3 features, are sensitive to those frequencies (Geirhos et al., 2018). Instead, by looking at the conditional metrics in Fig. 1 (middle), which are based on DreamSim that extract more sematical features (Fu et al., 2023), we observe that LDM_{XL} belongs to the Pareto front together with $\text{LDM}_{1.5}$, $\text{LDM}_{2.1}$. In particular, LDM_{XL} achieves the best conditional realism, obtained at the expense of conditional diversity. Here, we remark that $\text{LDM}_{\text{XL-Turbo}}$ only gets comparable (slightly lower) realism but considerably lower diversity. This difference is evident by looking at C (LDM_{XL}) vs. D ($\text{LDM}_{\text{XL-Turbo}}$) in Fig. 2. When comparing Pareto optimal points of $\text{LDM}_{1.5}$ and $\text{LDM}_{2.1}$, we note that $\text{LDM}_{1.5}$ reaches slightly better conditional realism than $\text{LDM}_{2.1}$.

Consistency-realism. In Fig. 1 (right, top and bottom) we observe that realism and consistency show relatively strong positive correlation as improvement in one metric oftentimes leads to an improvement in the other metric, with the correlation being stronger for the conditional metrics than for the marginal ones. We observe that the Pareto front is dominated by LDM_{XL} and $\text{LDM}_{\text{XL-Turbo}}$ model, highlighting how the advancement of T2I generative models have favored consistency-realism over the diversity objective. Indeed, we can also notice that in the distribution of non-Pareto-optimal points, $\text{LDM}_{2.1}$ seems better than $\text{LDM}_{1.5}$, matching the historical development of these models.

Conclusions

- Progress in T2I models has been driven by improvements in realism and/or consistency. State-of-the art T2I models improve consistency and/or realism by sacrificing representation diversity. Yet, improvements in realism are correlated with improvements in consistency.
- More recent models should be used when the downstream task requires samples with high realism – $\text{LDM}_{\text{XL-Turbo}}$– and consistency – LDM_{XL}–. However, older models – $\text{LDM}_{1.5}$ and $\text{LDM}_{2.1}$– are preferable for tasks that require good representation diversity.
- Both marginal and conditional metrics display correlated Pareto fronts.
- There is no best model and the choice of model should be determined by the downstream application.

References

Pietro Astolfi, Arantxa Casanova, Jakob Verbeek, Pascal Vincent, Adriana Romero-Soriano, and Michal Drozdzal. Instance-conditioned gan data augmentation for representation learning. *arXiv preprint arXiv:2303.09677*, 2023. 1

Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20041–20053, 2023. 9

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018. 3

Marlene Careil, Matthew J. Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ktdETU9JBg. 13

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 14

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043–3054, 2023. 9

Jaemin Cho, Yushi Hu, Jason Michael Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ITq4ZRUT4a. 2, 13

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KqbCvIFBY7. 1, 2

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 15

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 14

Zibin Dong, Yifu Yuan, Jianye HAO, Fei Ni, Yao Mu, YAN ZHENG, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bxfKIYfHyx. 9

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 1

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. 14

Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DEiNSfh1k7. 4, 13, 14

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 4

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 9

Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero-Soriano. DIG in: Evaluating disparities in image generations with indicators for geographic diversity. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=FDt2UGM1Nz. Featured Certification. 1, 9, 10, 15

Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023. 1

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 13

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 14

Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1552–1565, 2020. 9

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI. 2, 14

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 14

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023. 13

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 9

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023. 2

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023. 9

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OuV9ZrkQlc. 2, 9

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 2, 14

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 9

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. 2024. 9

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014. 2, 3

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 13

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 15

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020. 14

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14277–14286, 2023. 9

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 9

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 3, 4

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 9

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36, 2024. 9

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*, 2022. 1

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 2, 3

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X. 1

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 14

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 3, 4

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2022. 1

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 15

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015. 14

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 1

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 15

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 9

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=QLcBzRI3V3`. 9

Mariia Zameshina, Olivier Teytaud, and Laurent Najman. Diverse diffusion: Enhancing image diversity in text-to-image generation. *arXiv preprint arXiv:2310.12583*, 2023. 1

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 13

Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32, 2019. 9

Xiangru Zhu, Penglei Sun, Chengyu Wang, Jingping Liu, Zhixu Li, Yanghua Xiao, and Jun Huang. A contrastive compositional benchmark for text-to-image synthesis: A study with unified text-to-image fidelity metrics. *arXiv preprint arXiv:2312.02338*, 2023. 9

A Related work

The evaluation of recent state-of-the-art image generative models is often carried with human studies focusing on human preference (Ku et al., 2024; Dong et al., 2024; Kirstain et al., 2023; Otani et al., 2023; Zhou et al., 2019), where human annotators are asked to choose among images generated with different models. They are usually asked to select either the image they like the most or the image that is more aligned with the prompt used to generate it. However, due to the high cost of human annotations, works like Xu et al. (2024) use the collected human preferences to train a model to predicts them, in order to compute these metrics at lower cost. While the outcome of all these studies is useful to detect the most appealing generations, it provides only limited signal when the objective is to evaluate image generative models as world models, where several aspects need to be evaluated simultaneously. To this end, other works have focused on extending the evaluation to different aspects of the generation like fine-grained prompt-image alignment (*e.g.*, object counting and color consistency) (Ghosh et al., 2024; Hinz et al., 2020), compositionality (Li et al., 2024; Huang et al., 2023; Zhu et al., 2023; Park et al., 2021) and reasoning (Cho et al., 2023). Finally, HEIM (Lee et al., 2024) and HRS (Bakr et al., 2023) recently proposed to holistically evaluate T2I models, addressing up to 13 aspects including robustness, generalization, bias, fairness, and others, in addition to prompt-image alignment and image quality. However, some crucial aspects such as sample diversity are not investigated in these works, and more importantly, the several aspects analyzed are not combined together to understand the trade-offs and the multi-objective optimization of world models. In this regard, Yang et al. (2024); Rame et al. (2024) have investigated the multi-objective optimization in the context of finetuning foundation models including multimodal models and T2I models. In particular, these studies use Pareto fronts of multiple objectives as rewards to be directly optimized via reinforcement learning. However, none of these works considers the consistency-diversity-realism multi-objectives for conditional generative models as we do.

B Limitations

Our analysis only considers open models as evaluating closed models is very expensive or sometimes not possible. It would be interesting placing the dots of closed state-of-the-art models within the multi-objective pareto front. Moreover, it would be interesting to extend the analysis to ablate further knobs. For example, we have not included the knob of structured conditioning, like layouts, sketches or other form of control typically used to increase consistency. Another aspect that our analysis does not ablate is the effect of different data distribution on the consistency-diversity-realism pareto fronts –this aspect is currenty very hard to study due to the closed data filtering recipes of most models. Furthermore, for certain evaluated knobs like the retrieval augmented generation, the analysis could be deepen by considering for example the effect of different retrieval databases or stronger/more recent models than RDM—unfortunately, there is a scarcity of open models using RAG. Finally, our work suggests future research to understand whether the observed tradeoffs are fundamental, or could be overcome by future generations of better generative models.

C Additional results

C.1 Pareto fronts for geographic disparities in T2I models

We extend the use of consistency-diversity-realism Pareto fronts to characterize potential geographic disparities of state-of-the-art conditional image generative models. In particular, we follow Hall et al. (2024) and investigate geographic disparities of T2I models using the GeoDE dataset (Ramaswamy et al., 2024).

Consistency-diversity. Fig. 3 (left) depicts the region-wise consistency-diversity Pareto fronts. We observe that Europe, the Americas, and Southeast Asia exhibit the best Pareto fronts, with consistently higher diversity and consistency than Africa and West Asia. As previously noted, improving diversity (computed as marginal or conditional) comes at the expense of consistency. When considering marginal metrics (top), we observe that Europe and the Americas present the best Pareto fronts. Remarkably, $\mathrm{LDM}_{1.5}$ appears in all region-wise Pareto fronts, whereas $\mathrm{LDM}_{2.1}$ appears remarkably less frequently, and does not appear at all in the Pareto front of Europe. This is in line with prior works that demonstrate that recent advancements on standard benchmarks may

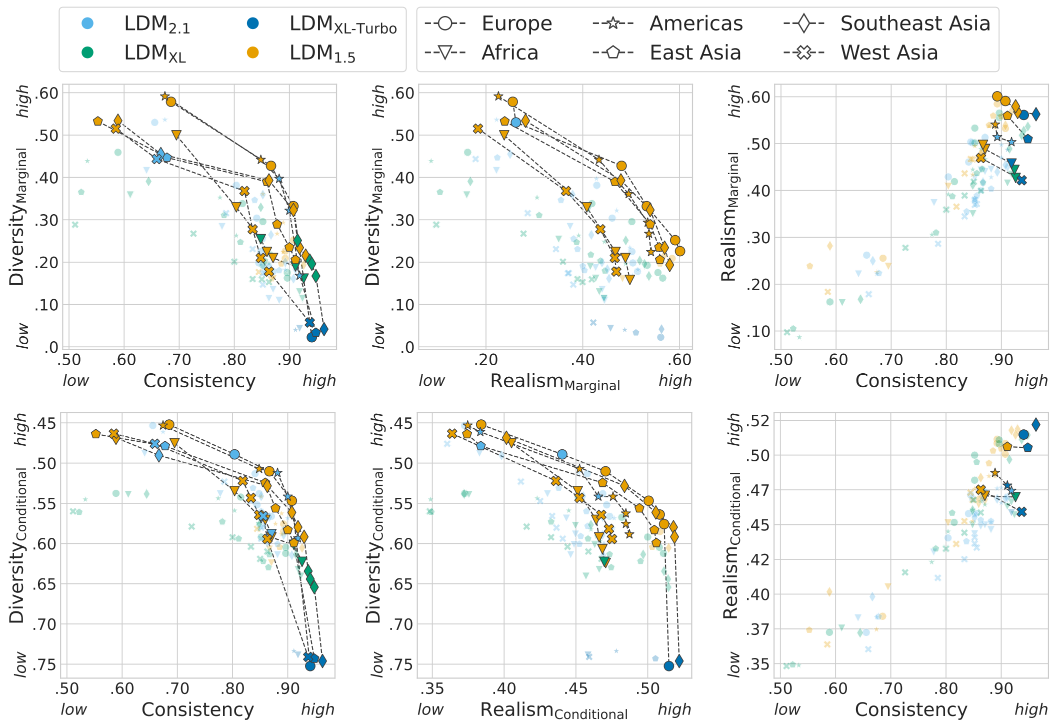

Figure 3: Consistency-diversity (left), realism-diversity (middle) and consistency-realism (right) Pareto fronts for T2I models on the GeoDE dataset. Consistency measures only the presence of the object in the image. Each models' configuration differ solely for guidance scale value.

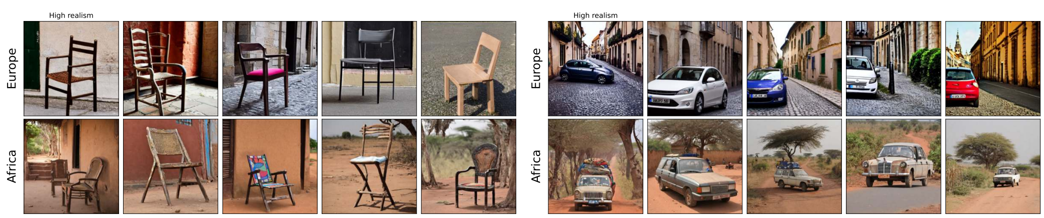

Figure 4: GeoDE qualitative. Left: `A chair in {region}`. Right: `A car in {region}`

have come at the cost of reduced real world geographic representations (Hall et al., 2024). However, we positively discover that disparity *reduction* occurs via LDM$_{\text{XL}}$ which appears in the Pareto fronts of Africa, West Asia and South East Asia, bringing the results of Africa closer to those of Europe or the Americas. Yet, LDM$_{\text{XL-Turbo}}$ only appears in the Pareto fronts of some regions, and presents the highest consistency. We observe that the improvements achieved by LDM$_{\text{XL}}$ for Africa are notably reduced when distilling the model into LDM$_{\text{XL-Turbo}}$. When considering conditional metrics (bottom), we see that all T2I models appear in the Pareto fronts. Once again, LDM$_{1.5}$ shows the highest diversity and LDM$_{\text{XL-Turbo}}$ the highest consistency. As in the previous case, LDM$_{\text{XL}}$ only appears in the Pareto fronts of West Asia, Africa, and South East Asia, and bridges the consistency and diversity performance gap between Africa and both Europe and the Americas. Yet, the improvements observed in LDM$_{\text{XL}}$ for Africa disappear when considering LDM$_{\text{XL-Turbo}}$.

Realism-diversity. Fig. 3 (middle) depicts the region-wise realism-diversity Pareto fronts. In the top panel (precision vs. recall), we observe that, similarly to MSCOCO2014 (Fig. 1), realism and diversity performance of T2I models present a clear tradeoff. Focusing on the regions, we see that the Pareto fronts of West Asia and Africa are visibly worse than the others. In terms of models, LDM$_{1.5}$ is the model that generally dominates the Pareto fronts of all regions. Moving to conditional metrics (bottom), we notice similar trends. However, LDM$_{\text{XL}}$ appears in the highest realism part of the Pareto front of Africa, and LDM$_{\text{XL-Turbo}}$ appears in the highest realism part of the Pareto fronts of Europe and Southeast Asia. By looking at the inter-region disparities along different areas of the Pareto fronts, we notice a gradual increase of the inter-region variance when moving from high diversity (low realism) to high realism (low diversity). This result suggest that maximizing realism might exacerbate stereotypes – as suggested by the lower diversity – and increase geographical disparities – as suggested by the increased variance across region-wise Pareto fronts. We provide a visual validation of this phenomenon in Fig. 4 (See **????** in **??** for more examples).

Consistency-realism. Fig. 3 (right) depicts the region-wise consistency-realism Pareto fronts. As shown in the figure, consistency and realism correlate as previously noticed on MSCOCO2014. The region-wise stratification shows that West Asia and Africa are again the regions with the worst Pareto fronts. The regions that exhibit the best Pareto fronts are East Asia, Southeast Asia, and Europe. Focusing on the top plot (marginal metrics), the Pareto fronts of all regions except the Americas contain $\text{LDM}_{1.5}$ and $\text{LDM}_{\text{XL-Turbo}}$. Note that $\text{LDM}_{1.5}$ consistently stands out in terms of realism, whereas $\text{LDM}_{\text{XL-Turbo}}$ shines in consistency. $\text{LDM}_{2.1}$ and LDM_{XL} are only present in the Pareto of the Americas and Africa, respectively. In the bottom plot (conditional metrics), the situation is very similar, but we notice that for Europe and Southeast Asia the Pareto is only composed by $\text{LDM}_{\text{XL-Turbo}}$.

Key insights

- Improving generation diversity comes at the expense of consistency for all regions considered. Realism and diversity also present a clear tradeoff for all regions, whereas realism and consistency appear correlated.
- Interestingly, the oldest model, $\text{LDM}_{1.5}$ dominates the most recent ones, and consistently appears in the Pareto fronts of all regions, when considering any pair-wise objective. However, LDM_{XL} reduces the disparities between Africa and Europe or the Americas in terms of diversity and consistency, as we move towards the high consistency part of the Pareto fronts.
- Advances in T2I models reduce region-wise disparities in terms of consistency and increase the disparities in terms of realism, while sacrificing diversity across all regions.

C.2 The impact of knobs on consistency-diversity-realism

Finally, in this section, we study the effect of different knobs that control consistency, diversity and realism of conditional image generative models. In the interest of space, we focus on the conditional metrics, and perform the analysis on MSCOCO2014.

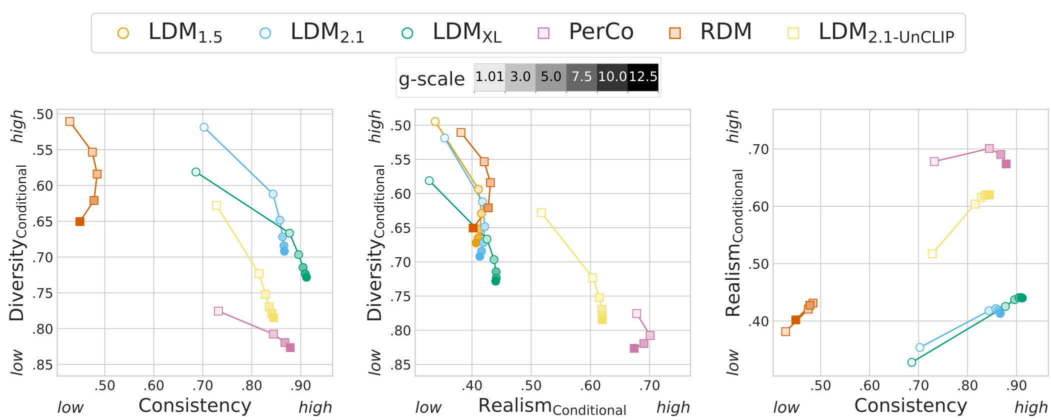

Figure 5: Ablation on guidance scale. To help readability, we report only a subset of the points presented in Fig. 1 and **??**, selecting runs with default values for other knobs.

Guidance scale. Fig. 5 depicts the effect of guidance scale on consistency-diversity (left panel), realism-diversity (middle panel), and consistency-realism (right panel) objectives. By looking at the consistency-diversity plot, we observe that increasing the guidance scale leads to improved consistency at the expense of the diversity in most cases [2], with LDM_{XL} showing the highest relative improvements. Moreover, for all models we notice that the initial increase in the guidance scale –

2

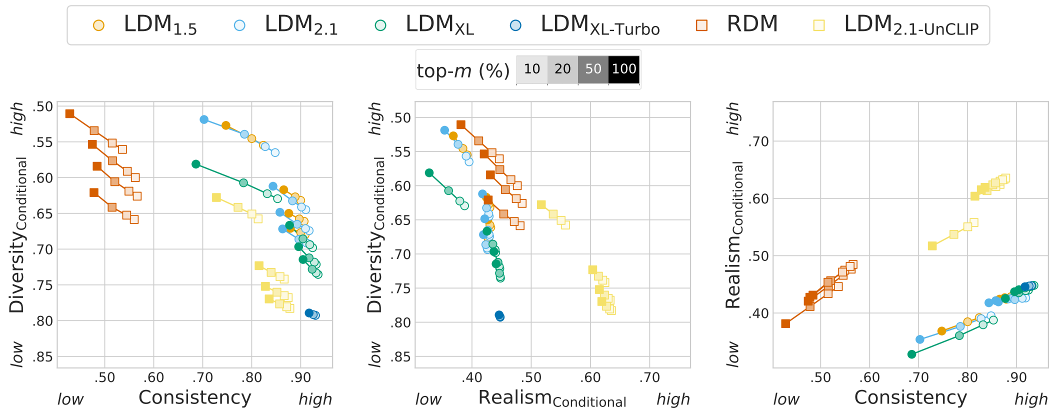

Figure 6: Ablation on top-m filtering.

from 1.01 to 3.0 – leads to the biggest consistency improvements. By looking at the realism-diversity plot, we note that the increase in the guidance scale often leads to increase in realism at the expense of diversity, with $\text{LDM}_{\text{2.1-UnCLIP}}$ and PerCo benefiting the most and the least from this knob, respectively. Moreover, we note that, in most cases, increasing the guidance scales beyond 7.5 no longer results in realism improvements. Finally, the consistency-realism plot reveals that by increasing the guidance scale the models generally improve both the consistency and realism. However, too large values of guidance may lead to decreasing the image realism; this happens for all models except of $\text{LDM}_{\text{2.1-UnCLIP}}$ and LDM_{XL}.

Post-hoc filtering. Fig. 6 depicts the effect of applying top-m filtering. In the consistency-diversity plot (left), we observe that top-m filtering (based on CLIPScore) leads to improvements in consistency for all models – the lower the value of m, the higher the consistency. Unsurprisingly, the models that initially have high consistency scores do not gain as much when leveraging top-m filtering as the models that start with low consistency scores. Moreover, we observe that the post-hoc filtering consistently leads to a diversity decrease. However, this decrease is less pronounced for the top-m filtering than for the guidance knob, as is the case for the consistency increase (*cf*. Fig. 5). The diversity-realism plot (middle) shows that post-hoc image filtering leads to an increase in the realism at the expense of diversity. By looking at the realism-consistency plot (right), we note that the post-hoc filtering is an effective way to increase both image realism and consistency, with the latter one improving faster.

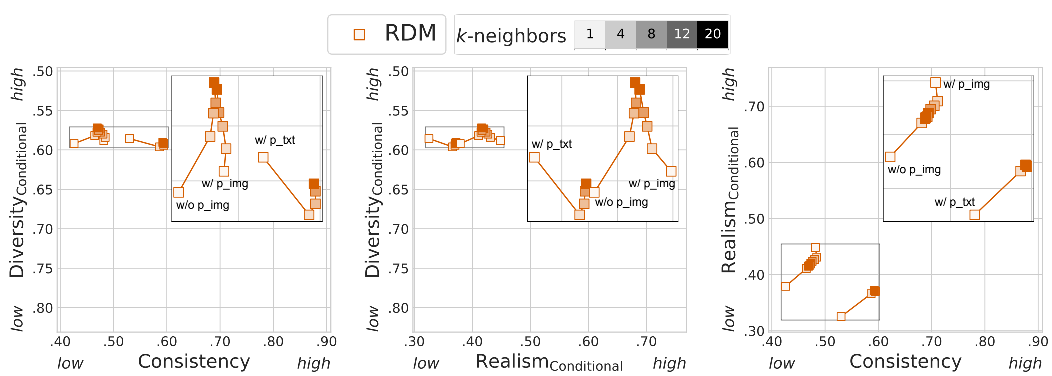

Figure 7: The effect of the neighborhood size on diversity, consistency and realism metrics. To improve readability we report a zoomed-in view in the top right of each plot.

Retrieval augmentation neighborhood size. The amount of neighbors used in retrieval augmentation may impact consistency, diversity, realism based on the semantic of the neighbors. In Fig. 7, we study the impact of the neighborhood size k for RDM. We notice that, in absolute terms, the impact of k is minor in all the pairs of metrics considered, suggesting that this knob is not as effective as the previous ones. In the consistency-diversity plot (left), we observe that increasing k from 4 to 20 leads to a small but consistent increase in diversity, while maintaining consistency. However, when increasing k from 1 to 4, we generally see a small improvement in consistency. This result is expected

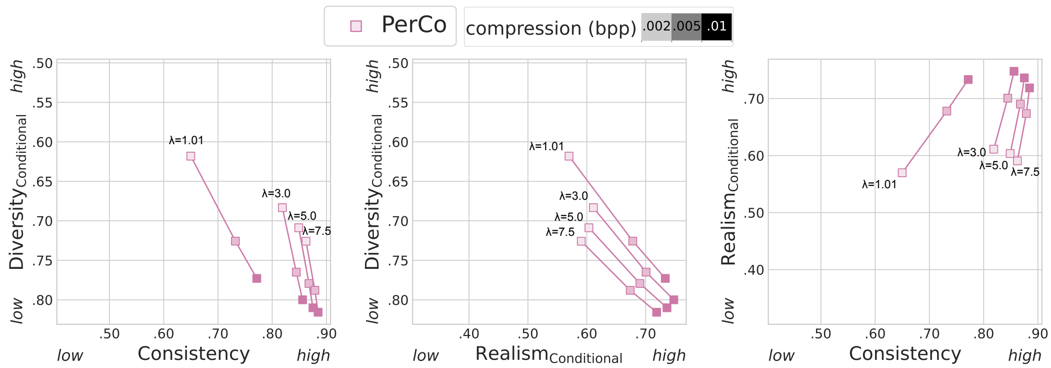

Figure 8: The effect of the compression rate on diversity, consistency and realism metrics.

as by increasing the neighborhood size we might include more diverse neighbors, and as long as those neighbors are semantically similar to the query image, they will not affect the consistency of the generation. In the realism-diversity plot (middle), we observe similar trends: increasing k from 4 to 20 results in small diversity improvements with little to no effect on realism, while increasing k from 1 to 4 results in small realism improvements. Interestingly, RDM prompted with text achieves lower realism than the others models. Moreover, increasing k when the query image is present together with the neighbors, slightly harms the realism. Finally, in the consistency-realism plot (right), we note a positive correlation between the two metrics when text query or no query is used.

Compression rate. The reconstructions produced by an image compression model are highly dependent on the selected compression rate, measured in terms of bit-per-pixel (bbp) of the compressed image, where high compression rate means low bpp. In Fig. 8 we assess PerCo with different bitrates and at different guidance scales. By looking at the left panel, we observe that decreasing the bitrate leads to notable increases in conditional diversity, which is inline with qualitative observations made by Careil et al. (2024). Moreover, these diversity increases only marginally reduce consistency, especially for guidance scales > 3, suggesting that even at high compression rates, the reconstructed images maintain their semantics. By contrast, in realism-diversity (middle), higher compression leads to a pronounced loss in realism, suggesting that the reconstructed images do not necessarily capture all the details from the original images. Finally, the results presented for consistency-realism (right) suggest, once again, that consistency and realism are correlated.

Key insights

- Guidance scale trades diversity for consistency and realism. Consistency and realism improve with higher guidance scale, but realism improvements saturate earlier than consistency improvements.
- Post-hoc filtering improves consistency and realism at the expense of diversity. Although both consistency and realism improve with this knob, consistency increases at a faster pace. Overall, post-hoc filtering appears less effective than guidance scale.
- The effect of retrieval augmentation on consistency-diversity-realism appears minor, questioning the knobs efficacy to control the multi-objective.
- Compression rate affects image realism and diversity, but has little effect on consistency, as compression models tend to maintain the image semantics.

D Analysis details

D.1 Evaluation metrics

Consistency, $\mathcal{C}$. Prompt-generation consistency is measured either with distance or similarity-based scores – *e.g.*, CLIPScore (Hessel et al., 2021), LPIPS score (Zhang et al., 2018) and DreamSim score (Fu et al., 2023) – or with visual question answering (VQA) approaches – *e.g.*, TIFA (Hu et al., 2023), VQAScore (Lin et al., 2024), and DSG (Cho et al., 2024) metrics –. In our analysis, we opt to use VQA approaches as they are reported to be more calibrated and interpretable than the distance and similarity-based scores (Cho et al., 2024). Concretely, we measure the prompt-generation consistency with DSG. DSG relies on questions $\mathbf{Q}$ generated from the prompt $\mathbf{p}$ and their corresponding answers

A. Per-prompt consistency, $\mathcal{C}^p$, is defined as:

$$\mathcal{C}^p = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}\mathbb{1}\Big(\mathrm{VQA}(\mathbf{Y}_j, \mathbf{Q}_i), \mathbf{A}_i\Big), \tag{1}$$

where N represents the number of images generated per conditioning prompt, Q_j represents the number of question per j-th image, and $\mathbb{1}$ represents the indicator function. The consistency over a set of prompts may be aggregated into a global consistency score, $\mathcal{C}$, by averaging all the conditioning-wise DSG scores, $\mathcal{C}^p$.

Conditional diversity, $\mathcal{D}_C$. We measure per-prompt conditional diversity as follows:

$$\mathcal{D}_C^p = \frac{1}{N^2 - N}\sum_{j \neq i}\mathcal{S}(f_\phi(\mathbf{Y}_j), f_\phi(\mathbf{Y}_i)), \tag{2}$$

where $\mathcal{S}$ is a similarity or distance function, and f_ϕ is an image feature extractor. In our analysis, we use cosine similarity and the DreamSim (Fu et al., 2023) feature extractor. DreamSim leverages an ensemble of modern vision encoders, including DINO (Caron et al., 2021) and two independently trained CLIP encoders, and is reported to correlate well with human perception. The conditional diversity over a set of prompts may be aggregated into a global score, $\mathcal{D}_C$, by averaging all the conditioning-wise scores, $\mathcal{D}_C^p$.

Conditional realism, $\mathcal{R}_C$. We measure per-prompt conditional realism as follows:

$$\mathcal{R}_C^p = \frac{1}{N}\sum_{j=1}^{N}\max_i(\mathcal{S}(f_\phi(\mathbf{X}_i), f_\phi(\mathbf{Y}_j))), \quad i \in \{1, \ldots, N'\}, \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{N' \times H \times W \times 3}$ represents a tensor of N' real images. Note that both $\mathbf{X}$ and $\mathbf{Y}$ represent generations and real images of the same prompt $\mathbf{p}$, respectively. Similarly to conditional diversity, we implement $\mathcal{S}$ as cosine similarity and use DreamSim as feature extractor. The conditional realism over a set of prompts may be aggregated into a global score, $\mathcal{R}_C$, by averaging all the conditioning-wise scores $\mathcal{R}_C^p$.

Marginal diversity, $\mathcal{D}_M$. Commonly used metrics of marginal diversity, such as *recall* (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) or *coverage* (Naeem et al., 2020), compare real and generated image distributions by leveraging a reference dataset of real images to ground the notion of diversity. Marginal diversity may also be measured with metrics which do not rely on a reference dataset, such as the Vendi Score (Friedman & Dieng, 2023). In our analysis, we use recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) to compute marginal diversity given its ubiquitous use in the literature. Recall measures marginal diversity as the probability that a random real image falls within the support of the generated image distribution.

Marginal realism, $\mathcal{R}_M$. The most commonly used metric to estimate image realism is the Fréchet Inception Distance (FID) (Heusel et al., 2017). FID relies on a pre-trained image encoder (usually, the Inception-v3 model trained on ImageNet-1k (Szegedy et al., 2015)) that embeds both generated and real images from a reference dataset. The metric estimates the distance between distributions of features of real images and features of generated images, relying on a Gaussian distribution assumption. The FID summarizes image realism and diversity into a single scalar. In our analysis, to disentangle both axes of evaluation, we use precision (Kynkäänniemi et al., 2019; Naeem et al., 2020) as marginal realism metric. Precision measures marginal realism as the probability that a random generated image falls within the support of the real image distribution.

D.2 Consistency-diversity-realism knobs

Guidance scale. To control the strength of the conditioning, a guidance scale (g-scale) hyper-parameter can be used to bias the sampling of diffusion models like DDPM (Ho et al., 2020), see *e.g.*, classifier (Dhariwal & Nichol, 2021) or classifier-free guidance (CFG) (Ho & Salimans, 2021). More precisely, rewriting **??** for diffusion models trained with CFG, we obtain:

$$\mathbf{Y} = \lambda g_\theta(\mathbf{Z}, \mathbf{p}) + (1 - \lambda) g_\theta(\mathbf{Z}, \emptyset), \tag{4}$$

where λ is the guidance scale, $\emptyset$ is an empty conditioning prompt, and the first and second terms indicate conditional and unconditional samplings, respectively. Importantly, λ can be arbitrarily increased (> 1) in order to steer the model to generate samples more aligned with the conditioning $\mathbf{p}$.

Post-hoc filtering. To improve the generated images, *e.g.* in terms of realism or consistency, or to avoid certain undesirable generations, a set of images generated for the same prompt may be filtered to retain the top-m images based on a predefined criterion, which can be either based on human preferences or automatic metrics. Considering the latter case, a common choice of metric is the CLIPScore, resulting in:

$$\mathbf{Y} = \mathrm{top}\bigg(m,\ \mathcal{S}(\mathbf{p}, f_\phi(\mathbf{Y}_j))\bigg), \tag{5}$$

where decreasing m ensures higher consistency.

D.3 Implementation details.

We adopt the `Diffusers` library for the LDM models (von Platen et al., 2022) and the official models' repos for RDM and PerCo. We set the number of inference steps to 50 (20 for PerCo as suggested in their paper) using deterministic sampling strategies, DPM++ (Lu et al., 2022) for `Diffusers` models and DDIM (Song et al., 2020) for others. For the conditional metrics on MSCOCO, we sample 10 images per prompt, using the 5,000 image-caption pairs of the 2017 validation split, while for the marginal metrics we sample 1 image per conditioning, using 30,000 randomly selected data points from the validation set of 2014. Note that, as MSCOCO contains multiples captions for each image, we fix the first caption as prompt for generations. For GeoDE, we sample 180 images for each of the `{object} in {region}` prompts for both conditional and marginal metrics. We disaggregate metrics by groups, per Hall et al. (2024), to measure disparities between geographic regions. For metrics based on DreamSim we use the ensemble backbone as recommended from the official repository. For marginal metrics we use the implementation of `prdc`. For DSG, we leverage `GPT-3.5-turbo` to generate questions from the prompts, and `InstructBLIP` (Dai et al., 2024) to make the predictions. When performing top-m filtering based on CLIPScore, we use `CLIP-ViT-H-14-s32B-b79K` from Hugging Face. Finally, we ablate different values for each knob as reported in Tab. 1.

Table 1: Knob values ablated per model.

Knob	values
g-scale	$[1.01, 3.0, 5.0, 7.5, 10.0, 12.5]$;
top-m filtering	$[10, 20, 50, 100]\%$