Geometry Fidelity for Spherical Images

Anders Christensen [† 1 2] **Nooshin Mojab** [3] **Khushman Patel** [3] **Karan Ahuja** [3 4] **Zeynep Akata** [2 5 6]
Ole Winther [1 7 8] **Mar Gonzalez-Franco** [3] **Andrea Colaco** [3]

Abstract

Spherical or omni-directional images offer an immersive visual format appealing to a wide range of computer vision applications. However, geometric properties of spherical images pose a major challenge for models and metrics designed for ordinary 2D images. We show that direct application of Fréchet Inception Distance (FID) is insufficient for quantifying geometric fidelity in spherical images. To remedy this, we introduce Omnidirectional FID (OmniFID), an extension of FID, which additionally captures field-of-view requirements of the spherical format.

1. Introduction

Spherical images, offering a full 360-degree horizontal and 180-degree vertical field of view hold immense potential for a broad range of computer vision applications such as virtual reality, game design and immersive panoramic image viewing. However, spherical images have geometric properties not exhibited by regular 2D images. Most existing datasets are not representative of this format of images, consequently most existing models, such as generative models, are also not directly applicable or optimized for spherical images. To reduce this challenge, we can project between a spherical 3D image and 2D representations of it. However, such projections present a series of trade-offs between conformity to the spherical image, and representing area of the sphere equally on the 2D plane. A wide range of map projections have been developed to project spherical images to the plane, among which equirectangular and cubemap projections are the most commonly used in practice (Zucker & Higashi, 2018). These projections have been designed to improve on specific properties, such as reducing distortions in the resulting viewpoint images (Chen et al., 2018; Chiariotti, 2021; Hussain & Kwon, 2021). For successful application, models applied on such representations must be aware of the inherent distortions in order to adhere to the geometric constraints of the 3D sphere.

One of the key metrics to measure image fidelity for generative 2D image models is FID (Heusel et al., 2017). Having been reported in a range of works on generation of spherical images (Chen et al., 2022; Lu et al., 2023; Wang et al., 2023; Akimoto et al., 2022), FID has also been established as the de-facto fidelity metric for spherical images. However, in this paper we demonstrate that FID fails to capture distortions related to the unique geometric requirements of spherical images when applied on equirectangular projections. This is a severe limitation of the metric for spherical image applications. We show this by proposing a noise transformation of equirectangular images, effectively reducing the field of view with little distortion in its 2D equirectangular representation. Fundamentally, the FID metric relies on features extracted from the Inception V3 convolutional neural network trained on ordinary 2D linear perspective images (Szegedy et al., 2016). Hence, to increase compatibility of the underlying Inception network with spherical image data, we present an extension of FID, namely Omnidirectional FID. OmniFID utilizes cubemap representations as an alternative to the hitherto primarily used equirectangular representations. Unlike equirectangular images, cubemap views are the result of rectilinear projections, providing better conformity to the shapes in the actual spherical rendering. Further, since the resulting views are square, the aspect ratio is maintained while resizing to 299 × 299 pixels in the FID calculations. Through our experiments we showcase that OmniFID is able to capture reductions in field-of-view, a crucial aspect for a quality metric for spherical image generation, while maintaining other positive properties of FID, such as sensitivity to noise.

2. Related work

Fréchet Inception Distance (FID) is a widely established metric often used to measure image fidelity for evaluating

[†] Work done at Google [1] Technical University of Denmark [2] Helmholtz Munich [3] Google [4] Northwestern University [5] Technical University of Munich [6] Munich Center of Machine Learning [7] University of Copenhagen [8] Copenhagen University Hospital. Correspondence to: Anders Christensen <andchri@dtu.dk>, Nooshin Mojab <nooshinmojab@google.com>.

Proceedings of the Geometry-grounded Representation Learning and Generative Modeling at 41st International Conference on Machine Learning, Vienna, Austria. PMLR Vol Number, 2024.

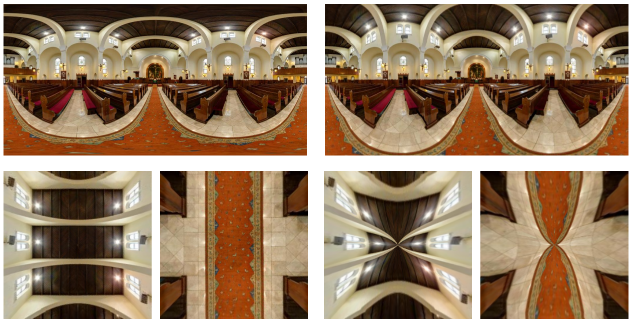

Figure 1. Visually, it is difficult to recognize field-of-view issues in the equirectangular format, but the problem is evident when rendered as a sphere and looking up/down. Top left: original spherical image with 180° vertical FOV represented as an equirectangular image. Top right: Resulting noisy equirectangular image, with a reduced vertical FOV of 140°. Bottom row: comparison of resulting views when looking upwards and downwards, respectively, in the two spherical images.

image generative models, in part due to some agreement with human perception and sensitivity to various noise types (Heusel et al., 2017). Under the assumption that this extends to equirectangular projections of spherical images, a majority of works in generative spherical imagery employ FID on this 2D representation as the main performance metric to measure the quality of generated images (Chen et al., 2022; Lu et al., 2023; Wang et al., 2023; Akimoto et al., 2022). However, unlike regular images, 2D representations of spherical images must satisfy unique geometric constraints. We showcase the shortcomings of FID in evaluating geometric fidelity of spherical images, and we present an extension of the metric enabling a more efficient evaluation designed for spherical images by leveraging projections of spherical images. As such, our paper is an addition to prior works like (Naeem et al., 2020; Borji, 2022) that detect and tackle issues with FID. Additionally, we note that using different projections for increasing compatibility of 2D image pretrained models with spherical images has previously been explored in other works like (Eder et al., 2020) for semantic segmentation.

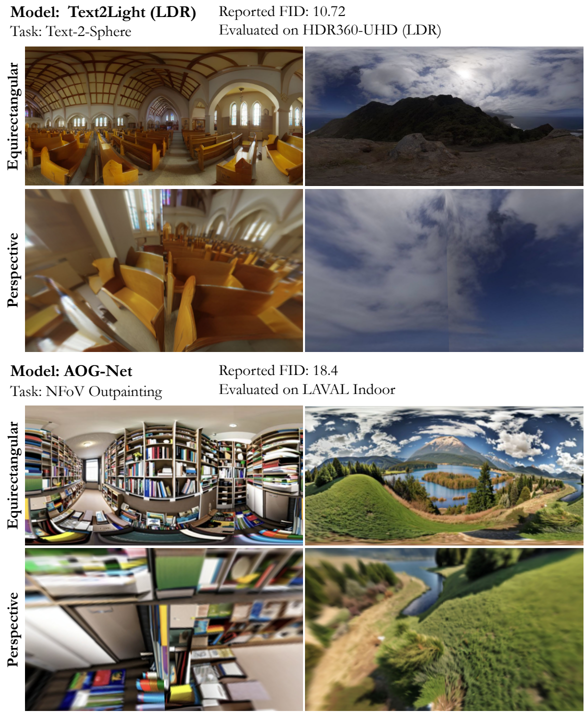

Figure 2. Although recent spherical image generation models (Text-2-Sphere and Image-2-Sphere) have begun achieving low FID scores, models are still struggling to produce images with full 180° vertical field-of-view and no seams. Above, we show equirectangular images from the models Text2Light (Chen et al., 2022) and AOG-Net (Lu et al., 2023) (top row in each block), along with their reported FID score. These images are from their respective papers. Below each image we display a perspective view when looking backwards, showing the resulting stitching across image borders (and at the poles). We find that FID does not sufficiently capture geometry fidelity issues in the generated images, such as benches converging to a point at the poles, or inconsistencies across image borders.

3. Omnidirectional FID Evaluation Metric

The traditional FID metric (Heusel et al., 2017) is computed between two sets of images, typically to assess the quality of a generative model by comparing the distance between the training set distribution and a corresponding generated distribution. All images from both image distributions are passed through the Inception V3 (Szegedy et al., 2016) convolutional model to obtain 2048-dimensional feature vectors. For each dataset, the obtained feature vectors are assumed to follow a multivariate Gaussian distribution with mean μ and covariance matrix Σ. The distance between the two distributions is then calculated using the Wasserstein-2 distance in $\mathbb{R}^{2048}$ (Heusel et al., 2017), i.e. as

$$
\begin{aligned}
FID(X_1, X_2) &:= d_{W-2}\left(\mathcal{N}\left(\mu_1, \Sigma_1\right), \mathcal{N}\left(\mu_2, \Sigma_2\right)\right) \\
&= \|\mu_1 - \mu_2\| + \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1\Sigma_2\right)^{\frac{1}{2}}\right)
\end{aligned}
$$

Although the underlying Gaussian assumptions have been shown to be faulty (Luzi et al., 2023), FID has been established as a popular metric due to sensitivity to noise and some correlation with human perception (Heusel et al., 2017).

Notably, however, spherical images present additional geometric structure compared to regular 2D images. It is not clear a priori whether the features produced by the Inception backbone, and hence the FID metric by extension, will

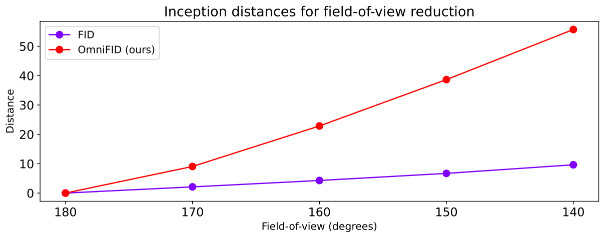

Figure 3. FID results compared to our modification, OmniFID, for detecting issues with field-of-view reductions on the 360-Indoor spherical image dataset (Chou et al., 2019). FID increases negligibly, despite reducing the vertical field-of-view from 180 degrees to 140 degrees, while our proposed OmniFID captures the difference.

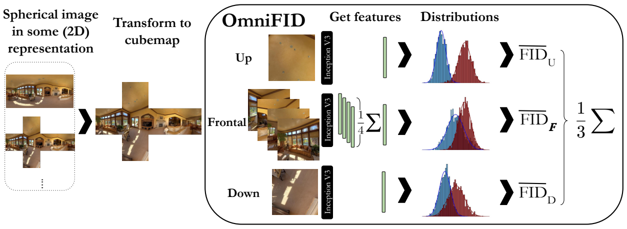
Spherical image in some (2D) representation
Transform to cubemap
OmniFID
Get features
Distributions
Up
Frontal
Down
Inception V3
$\frac{1}{4}\Sigma$
$\overline{\text{FID}}_U$
$\overline{\text{FID}}_F$
$\overline{\text{FID}}_D$
$\frac{1}{3}\sum$

Figure 4. Visualisation of our proposed Omnidirectional FID. Utilizing cubemaps and using view-point dependent image features allows OmniFID to detect issues with the spherical geometry, such as insufficient field-of-view.

capture divergences from these geometric constraints, even when the ground truth reference set contains proper projections of spherical images. Indeed, local image properties may well look reasonable, but global information in 2D representations is required to assess whether the geometric constraints are fulfilled (e.g. seamless stitching at poles and across image borders). Given the proven record of FID, rather than tampering with the metric itself, we adapt the data to the metric, improving behaviour on spherical images.

Increasing compatibility of FID to spherical images In order to improve conformity of the spherical images to the Inception backbone of the FID metric, we utilize the commonly used tangential spherical cubemap projection as the grounds for evaluation. Since transformations between projections will incur some image quality degradation (Hanhart et al., 2018), we believe it crucial to evaluate image fidelity on representations that are optimized for rendering for a representative evaluation. Further, since hardware and shaders have been optimized both for equirectangular and cubemap projections, evaluation on cubemap projections is not only valid, but perhaps even desirable. We note that although we focus on evaluation on the tangential spherical cubemap in this work, fair comparisons are also possible on other cubemap representations and re-projections, as long as the representations are unified. This could be relevant if generative models are to be evaluated for a specific shader or other transforms of the representations.

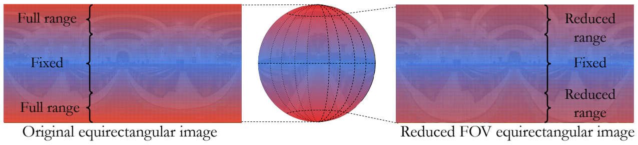
Full range
Fixed
Full range
Original equirectangular image
Reduced range
Fixed
Reduced range
Reduced FOV equirectangular image

Figure 5. Visualisation of the noise transformation used for reducing field-of-view in spherical images. The proposed transformation reduces vertical field-of-view while maintaining proportions in the central horizontal parts of the equirectangular image.

For a spherical image dataset X, we denote the set of 2D images resulting from cubemap projections by

$$\mathcal{C}^X := \{\mathcal{C}_F^X, \mathcal{C}_R^X, \mathcal{C}_B^X, \mathcal{C}_L^X, \mathcal{C}_U^X, \mathcal{C}_D^X\}, \qquad (1)$$

with the resulting view-specific image sets being denoted as $\mathcal{C}_{view}^X$, where F, R, B, L, U, and D represent the front, right, back, left, up, and down view of the cubemap projections, respectively. Visually, cubemaps are often represented as a dice with its faces folded out (see e.g. Figure 4). With the notation above we focus on the set structure of the individual cubemap views.

A priori we expect that the Inception feature distributions across cubemap views will differ. Concretely, we hypothesize that the feature vectors of the frontal views (front/right/left/back) are identically distributed, since the orientation of these views are arbitrary, but that the up and down view feature distributions will be dissimilar. Properties of the tangential spherical cubemap projection additionally support this, since structural distortions are larger at the polar faces (upwards and downward) compared to frontal faces, as a results of stitching at the poles. To get empirical evidence for this at the feature level, we compare the feature means of the different views on the 360-Indoor dataset (Chou et al., 2019). Between any two frontal views, the L^2 distances between mean features are 0.34 ± 0.13, while it is 24.27 ± 0.69 and 33.13 ± 0.59 between frontal views and up/down views, respectively. The L^2 distance is 17.05 between the average features of up and down views.

OmniFID On this basis, we group the cubemap views $\mathcal{C}^X$ into three disjoint subsets according to semantic similarities between frontal views $\mathcal{F} := \{F, R, B, L\}$, upward views U and downward views D. The Frontal group consists of four times as many perspective images as Up and Down. Since FID is biased by sample size (Chong & Forsyth, 2020), we average the Inception features within each view group of the perspective images for every cubemap, denoted by $\overline{FID}$. We then compute the FID metrics over the resulting subsets individually and average over the resulting distances, giving our proposed extension OmniFID:

$$OmniFID(X_1, X_2) := \frac{1}{3} \sum_{V \in \{U,D,\mathcal{F}\}} \overline{FID}(\mathcal{C}_V^{X_1}, \mathcal{C}_V^{X_2}) \tag{2}$$

OmniFID can be extended to perspective images from finer partitionings of the sphere. In particular, icosahedron tangent images (Eder et al., 2020) provide a fitting way to reduce distortion in the 2D perspective images at the cost of less semantic content in individual views, additional computation, and overlapping content between images. Due to properties of the icosahedron, tangent images can be grouped based on the latitude of their centers, as done with regular cubemaps above. As an example, a base level 0 icosahedron subdivision of the sphere gives 20 perspective images, which come in four groups of five images with centers at the same latitude. OmniFID^{20}, with the superscript denoting the number of perspective images, can then be computed on these images by first computing the Inception features on each image, averaging the features over each of the four groups, and computing the corresponding latitude-wise $\overline{FID}$ scores. OmniFID^{20} can then be obtained by averaging the $\overline{FID}$ scores as before. In our experiments we focus on using cubemaps, and demonstrate that this partitioning is adequate for detecting structural issues related to field-of-view in spherical images.

4. Experiments

Field-of-view reduction noise In order to evaluate the ability of FID to capture issues related to the geometric requirements of spherical images, we construct a noise transformation for reducing the vertical field-of-view in equirectangular projections of spherical images. The noise transformation is visualized in Figure 5, and the effects, which are a common artifact in generated equirectangular images, can be seen in Figure 1. Concretely, the field-of-view is reduced by an angle v by first cropping the top and bottom horizontal parts of the equirectangular image corresponding to $\frac{1}{2}v$ each. The central 90° horizontal part of the image is kept fixed, while the remaining parts of the image, each covering $90° - \frac{1}{2}v$, are resized using bi-linear interpolation to re-obtain the original image resolution. This is equivalent to ignoring the upper and lower $\frac{1}{2}v$ of the field-of-view of the spherical image when performing the equirectangular projection.

For our evaluations of FID and OmniFID, we use the 360-Indoor dataset (Chou et al., 2019), a collection of 3335 equirectangular images of indoor scenes with 360° horizontal and 180° vertical field-of-view. The images have resolution 1920×960, and we resize them to 1024×512. Compared to other datasets of spherical images, 360-Indoor is optimal for this purpose since it has both full field-of-view and has enough samples for the mean and covariance estimates in the FID and OmniFID metrics to be valid, although the number of samples is still low. In the experiments, we use an uncorrupted copy of the 360-Indoor dataset, and a copy which we gradually corrupt - here with reduction of vertical field-of-view.

In Figure 3, we see that decreasing the field-of-view from 180 to 140 degrees results in an FID of just 10. This is a particularly low value considering the bias of FID dependent on sample size, since the 360-Indoor dataset contains just 3335 spherical images (Chong & Forsyth, 2020). Further, when comparing with other types of noise (as shown in Figure 6), it is also evident that FID captures this geometric issue insufficiently. On the other hand, OmniFID crucially captures the difference in geometric fidelity between the real and corrupted dataset. This confirms that while FID fails to capture important aspects of the quality of spherical images, the adjustments made in OmniFID allows the metric to better quantify fidelity related to vertical field-of-view.

Noise sensitivity The FID metric became an established metric in part due to its sensitivity to various forms of noise (Heusel et al., 2017). In Appendix A we validate that OmniFID has not lost these properties of the FID metric through our extension. We compare the two metrics on various types and degrees of image corruptions. As above, we use two copies of the 360-Indoor dataset, gradually corrupting one with salt & pepper noise, Gaussian noise, and Gaussian blurring, respectively. We then compute FID and OmniFID between the two dataset. For each type of corruption, we increase the noise across four increasing levels of noise strengths. We note that the noise is applied on the equirectangular image, i.e. before transforming the images to cubemaps for OmniFID. The results are visualized in Figure 6, along with example equirectangular images showing the level of noise. We observe that for the different noise types, OmniFID follows the trend of FID closely, demonstrating that our extension retains these desired properties of FID. We further note that similarities between the FID and OmniFID scores across these types of noise confirm that the difference in scores on the field-of-view reduction task are not a matter of scaling.

Qualitative evaluation of OmniFID In Appendix B we compare OmniFID and FID scores on generated examples of varying quality from a set of different checkpoints based on a version of Dreambooth (Imagen) model (Saharia et al., 2022; Ruiz et al., 2023), trained on internal data sources and finetuned on our dataset. We showcase that OmniFID decreases as adherence of generated images to the spherical structure improves, while FID is unaffected - in fact, the lowest of the FID scores are achieved on a set of images with clear geometric issues.

5. Conclusion

In this work we showcased that the standard image fidelity metric FID, commonly used in evaluation of generative models, fails to capture crucial properties of spherical images associated with their unique geometrical constraints. To remedy the limitations of existing 2D image-based metrics, we presented an extension of FID, called OmniFID. Our experiments demonstrate the effectiveness of our proposed metric to measure geometry fidelity for spherical images through utilizing cubemap representations.

Acknowledgements

AC thanks the ELLIS PhD program and the Danish Pioneer Centre for AI, DNRF grant number P1, for support. ZA is supported by the ERC (853489-DEXIM). OW is funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606).

References

Akimoto, N., Matsuo, Y., and Aoki, Y. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11441–11450, 2022.

Borji, A. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

Chen, Z., Li, Y., and Zhang, Y. Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation. *Signal Processing*, 146:66–78, 2018.

Chen, Z., Wang, G., and Liu, Z. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

Chiariotti, F. A survey on 360-degree video: Coding, quality of experience and streaming. *Computer Communications*, 177:133–155, 2021.

Chong, M. J. and Forsyth, D. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6070–6079, 2020.

Chou, S.-H., Sun, C., Chang, W.-Y., Hsu, W. T., Sun, M., and Fu, J. 360-indoor: Towards learning real-world objects in 360 indoor equirectangular images. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 834–842, 2019.

Eder, M., Shvets, M., Lim, J., and Frahm, J.-M. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12426–12434, 2020.

Hanhart, P., He, Y., Ye, Y., Boyce, J., Deng, Z., and Xu, L. 360-degree video quality evaluation. In *2018 Picture Coding Symposium (PCS)*, pp. 328–332. IEEE, 2018.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hussain, I. and Kwon, O.-J. Evaluation of 360 image projection formats; comparing format conversion distortion using objective quality metrics. *Journal of Imaging*, 7(8): 137, 2021.

Lu, Z., Hu, K., Wang, C., Bai, L., and Wang, Z. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. *ArXiv*, abs/2309.03467, 2023.

Luzi, L., Marrero, C. O., Wynar, N., Baraniuk, R. G., and Henry, M. J. Evaluating generative networks using gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 279–288, 2023.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Wang, J., Chen, Z., Ling, J., Xie, R., and Song, L. 360-degree panorama generation from few unregistered nfov images. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6811–6821, 2023.

Zucker, M. and Higashi, Y. Cube-to-sphere projections for procedural texturing and beyond. *Journal of Computer Graphics Techniques Vol*, 7(2), 2018.

A. Sensitivity to noise

Here we present experiments providing empirical evidence that OmniFID has not lost the noise sensitivity of the FID metric through our extension using averages of Inception features from perspective images. We compare the OmniFID and FID on various types and degrees of image corruptions. We use two copies of the 360-Indoor dataset, gradually corrupting one with salt & pepper noise, Gaussian noise, and Gaussian blurring, respectively. We then compute FID and OmniFID between the two dataset. For each type of corruption, we increase the noise across four increasing levels of noise strengths. We note that the noise is applied on the equirectangular image, i.e. before transforming the images to cubemaps for OmniFID. The results are visualized in Figure 6, along with example equirectangular images showing the level of noise. We observe that for the different noise types, OmniFID follows the trend of FID closely, demonstrating that our extension retains these desired properties of FID.

B. Qualitative evaluation of OmniFID

To qualitatively evaluate our proposed Omnidirectional FID, we compute FID and OmniFID on generated images from three different checkpoints of a finetuned text-to-image generative model. The model is based on a version of Imagen (Saharia et al., 2022) trained on internal datasources, and finetuned using Dreambooth (Ruiz et al., 2023) with a batch size of 16. We finetune the model on the 360-Indoor equirectangular image dataset (Chou et al., 2019) and use captions generated by a multimodal language model. This gives us 3252 image-caption pairs after removing duplicate and empty captions.

The captions were generated by giving the multimodal model prompts with few-shot examples describing the content of corresponding equirectangular images, followed by keywords of e.g. style, lighting, and indoor/outdoor. An example of such a given few-shot example caption is: ”living room with couches, TV, coffee tables and fireplace. french style decoration, daylight, indoor”. Finally, we edit the prompt to be ”a panoramic view of a <caption>”.

Below, we show example equirectangular image generations from model checkpoints after 5000, 10000, and 20000 steps (Figure 7, Figure 8, Figure 9, respectively). The visualized generations are generated from the same prompts across the different checkpoints, where the corresponding prompts were selected randomly. Results show that the FID score is near-constant across the checkpoints (33.96, 35.42, 34.95, respectively). Further, although the example generations from the 5000 step model demonstrate that the model has issues constructing realistic geometry, the FID score is lowest for this checkpoint. On the contrary, OmniFID decreases monotonically over the checkpoints as geometry fidelity improves (63.39, 60.38, 55.07, respectively).

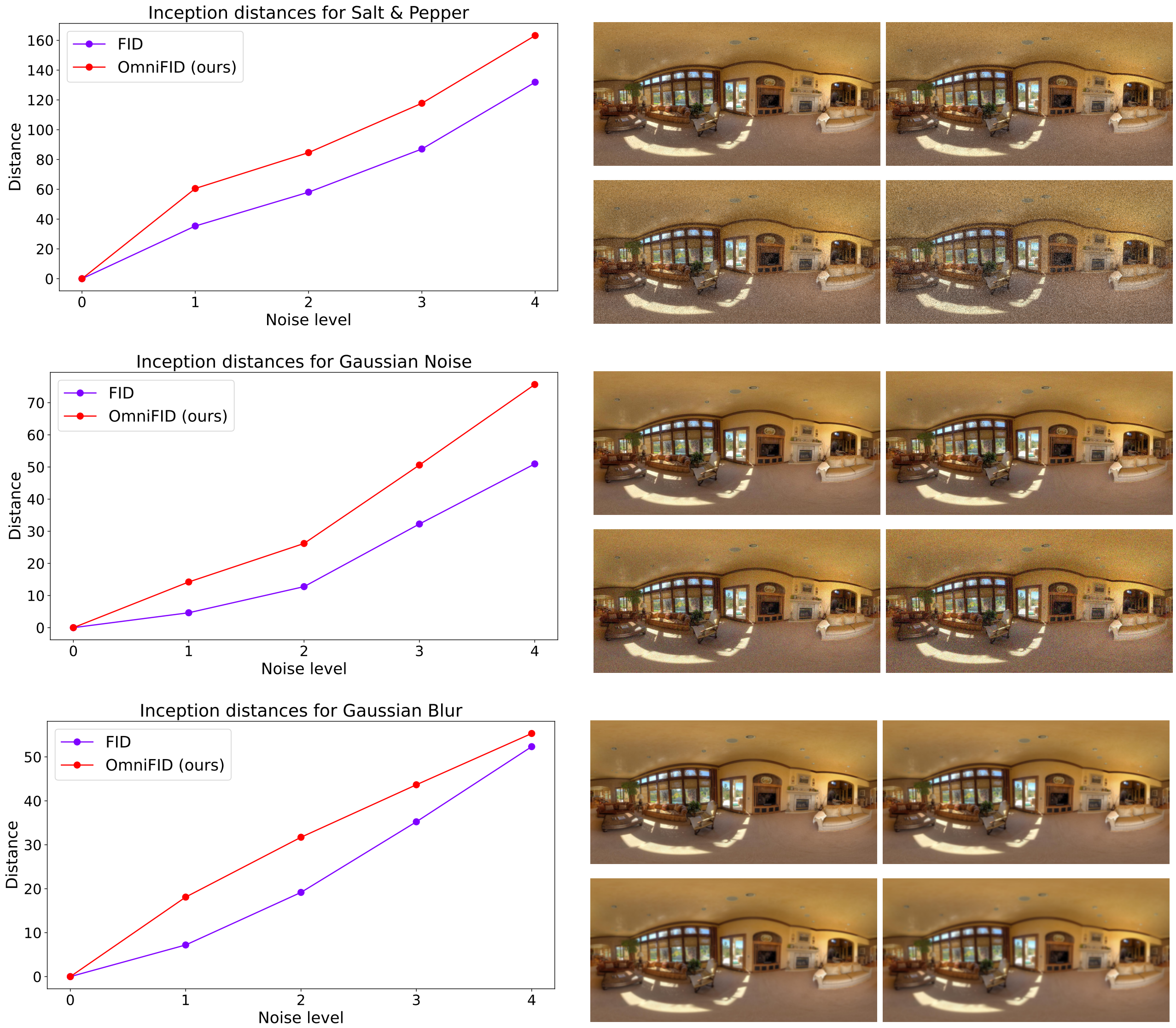

Figure 6. FID compared to our OmniFID on various noise types unrelated to the spherical geometry (salt & pepper, Gaussian noise, and Gaussian blur). Noise is applied to the equirectangular image before it is transformed to the cubemap. OmniFID retains the noise-related properties of FID.

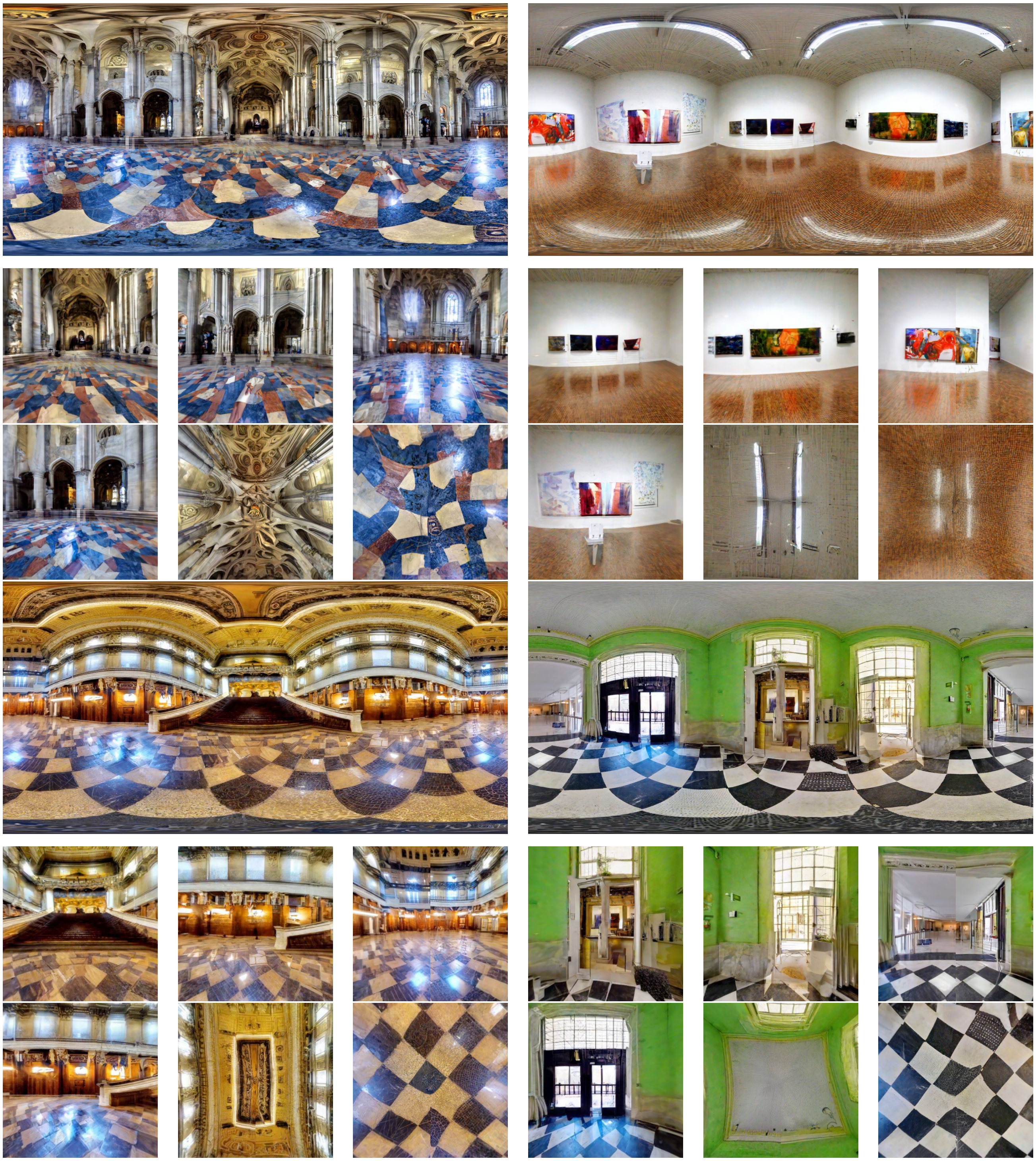

Figure 7. Four example equirectangular generations of a text-to-image model fine-tuned on 360-Indoor after 5000 steps. Under each generation we show the cubemap images to illustrate the geometry of the rendered views (top left to bottom right: front/right/back/left/up/down). FID is 33.96, OmniFID is 63.39.

.

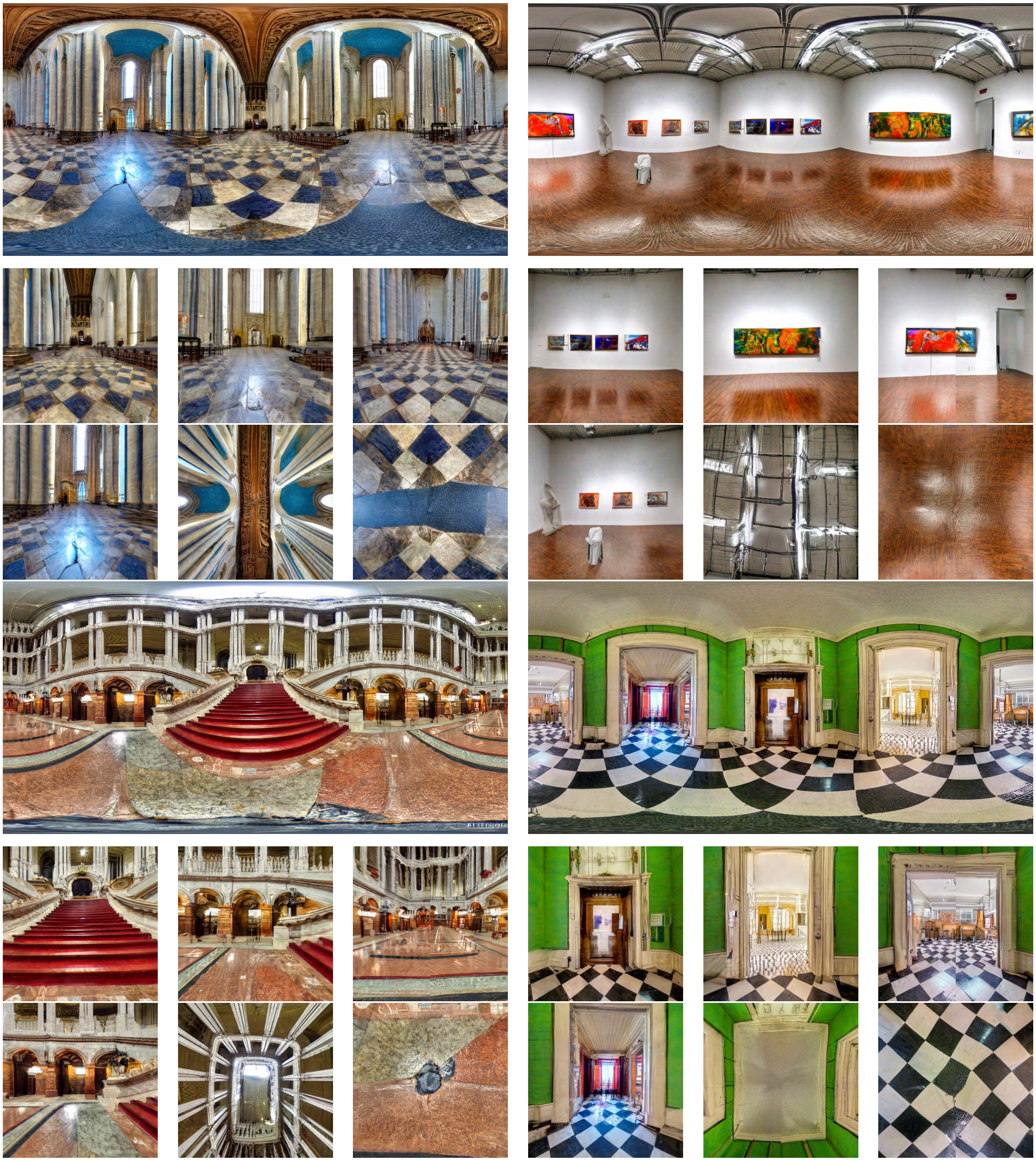

Figure 8. Four example equirectangular generations of a text-to-image model fine-tuned on 360-Indoor after 10000 steps. Under each generation we show the cubemap images to illustrate the geometry of the rendered views (top left to bottom right: front/right/back/left/up/down). FID is 35.42, OmniFID is 60.38.

.

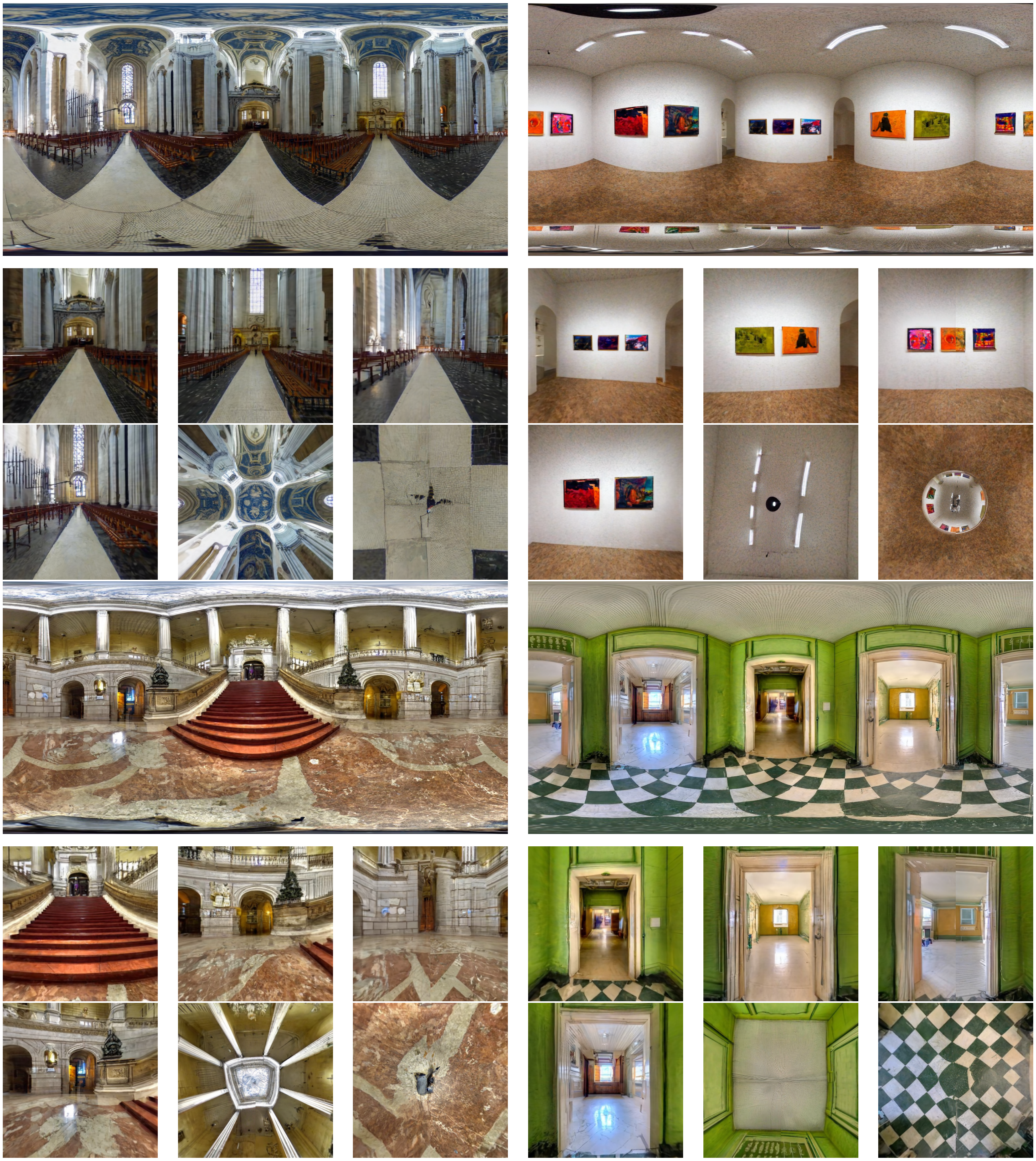

Figure 9. Four example equirectangular generations of a text-to-image model fine-tuned on 360-Indoor after 20000 steps. Under each generation we show the cubemap images to illustrate the geometry of the rendered views (top left to bottom right: front/right/back/left/up/down). FID is 34.95, OmniFID is 55.07.