StereoSpace: Depth-Free Synthesis of Stereo Geometry via End-to-End Diffusion in a Canonical Space

Supplementary Material

Dataset	Baseline	Setting	#Samples	Year
NeRF-Stereo [24]	🔀	🌲🏠	27K	2023
SceneSplat [13]	🔀	🏠	5K	2025
TartanAir [28]	25 cm	🌲🏠	306K	2020
Dynamic Replica [9]	🔀	🏠	145K	2023
IRS [27]	10 cm	🏠	103K	2021
Falling Things [25]	6 cm	🌲🏠	61K	2018
LayeredFlow [31]	🔀	🏠	31K	2024
VKITTI2 [4]	53.3 cm	🌲	21K	2020
InfinigenSV [6]	🔀	🌲	17K	2024
SimStereo [7]	16 cm	🏠	14K	2022
UnrealStereo4K [23]	🔀	🌲🏠	8K	2021
Spring [14]	🔀	🌲	5K	2023
PLT-D3 [22]	12 cm	🌲	3K	2024
Sintel [3]	10 cm	🌲	1K	2012

Table 1: **Training data.** Stereo datasets used for mixed training of *StereoSpace*. For each dataset, we report the baseline (fixed value when available, 🔀 indicates a variable baseline), scene type (indoor 🏠 or outdoor 🌲), the number of samples, and the release year. Datasets above the dotted line ..... are multi-baseline.

Section 1 reports additional details concerning the training of StereoSpace, Section 2 describes the main conditioning mechanisms used to allow for camera control, Section 3 describes the evaluation split composition in detail and finally Section 5 shows further qualitative results.

1. Training Details

Dataset Composition. Our training data is drawn from 14 publicly available data sources. Here, we provide additional statistics in Tab. 1 for the stereo datasets. We report the nominal baseline, whether a dataset primarily contains indoor or outdoor scenes, the number of training stereo pairs and the release year, respectively. Together, these sources cover a diverse range of camera baselines, scene scales, and photorealism levels.

Disparity Supervision. Most datasets provide ground-truth left-to-right and right-to-left disparity maps. When some direction is missing, we infer the respective disparity using a strong off-the-shelf stereo matcher (FoundationStereo [30]), and treat the resulting pair as pseudo-ground-truth. Having both directions is required to compute the left–right consistency mask used in the warping loss (Sec. 3.3 in the main paper), which restricts supervision to pixels that are co-visible in the source and target views. The synthetic LayeredFlow dataset is a notable exception: its multi-layer depth representation violates the single-surface assumption underlying our view warping formulation. Therefore, for LayeredFlow samples, we disable the warping loss and re-weight the training batches accordingly.

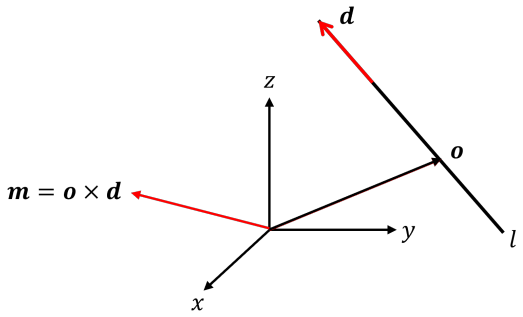

Figure 1. **Plücker coordinates** of line ℓ are given by the 6D homogeneous vector $(\mathbf{d}, \mathbf{m})$.

Rendered Multi-Baseline Tuples. For SceneSplat-7K [13], we derive multi-baseline, rectified stacks from the pre-optimized Gaussian splats. We restrict to the Hypersim [18], Replica [20], and ScanNet++ [33] subsets, filtering scenes using the dataset-provided PSNR, SSIM, LPIPS, and depth ℓ_1 metrics to discard photometrically or geometrically unstable reconstructions. Within each retained scene, we form candidate stacks with moderate baselines and subsamples up to 20 diverse bundles via k-means over stack centroids per splat. Each selected stack is rendered to RGB and depth with a small global focal-length scaling jitter, and stacks with insufficient geometric support are removed based on simple depth- and opacity-based heuristics.

Training Schedule. StereoSpace, including all its tested variations, is trained for 3 epochs, corresponding to approximately 48.6 K optimizer steps. This schedule matches the GenStereo [16] setup and ensures a comparable training budget between variants in the ablation study (Sec. 4.4). We also experimented with the extension of training to 5 epochs, but observed saturated performance and no consistent improvement relative to added computation, so all reported results adhere to the 3-epoch schedule. In contrast to prior work, StereoSpace is trained in both stereo directions (left→right and right→left). Empirically, this bidirectional supervision did not degrade performance, while enabling inference-time stereo generation in either direction, whereas competing approaches typically support only the left→right direction.

Dataset	Metric	GenStereo	Lyra	StereoSpace (ours)	StereoDiffusion	ZeroStereo
Middlebury	PSNR ↑	**18.54**	18.06	17.04	17.06	17.48
	SSIM ↑	**0.5989**	0.5793	0.5410	0.5246	0.5372
	LPIPS ↓	**0.184**	0.247	0.234	0.304	0.264
DrivingStereo	PSNR ↑	**24.95**	24.25	23.44	22.89	23.64
	SSIM ↑	**0.7939**	0.7792	0.7424	0.7365	0.7511
	LPIPS ↓	**0.136**	0.172	0.164	0.193	0.170
Booster	PSNR ↑	21.74	**23.21**	21.91	20.42	18.69
	SSIM ↑	0.7348	**0.7813**	0.7351	0.6986	0.5733
	LPIPS ↓	0.237	**0.196**	0.202	0.322	0.443
LayeredFlow	PSNR ↑	17.44	**18.67**	17.67	17.08	15.83
	SSIM ↑	0.5912	**0.6327**	0.5940	0.5957	0.4421
	LPIPS ↓	0.370	**0.321**	0.335	0.433	0.521

Table 2: **Evaluation with conventional metrics – PSNR, SSIM, LPIPS.** Experiments on Middlebury [19], DrivingStereo [32], Booster [17] and LayeredFlow [31].

2. Viewpoint Conditioning

In this section, we recall the principles behind the conditioning mechanisms used by StereoSpace.

Plücker rays. A 3D line ℓ can be represented by a point $\mathbf{o} \in \mathbb{R}^3$ and a (unit) direction $\mathbf{d} \in \mathbb{R}^3$ (Fig. 1). Its Plücker (Grassmann) coordinates are the homogeneous 6D vector

$$\ell \equiv (\mathbf{d},\, \mathbf{m}), \qquad \mathbf{m} = \mathbf{o} \times \mathbf{d}. \tag{1}$$

By construction $\mathbf{d}\cdot\mathbf{m} = 0$ (the Plücker constraint), and for any other point $\mathbf{o}' = \mathbf{o} + \lambda\mathbf{d}$ on the same line we have $(\mathbf{o}' \times \mathbf{d}) = \mathbf{m}$. Hence, Plücker coordinates are invariant to sliding $\mathbf{o}$ along the line, which makes them a natural parametrization of camera rays [5].

For a pinhole camera with center $\mathbf{c}$ (in world coordinates), the ray through the pixel (i, j) has Plücker coordinates $\ell_{ij} = (\mathbf{d}_{ij}, \mathbf{m}_{ij})$ with

$$\mathbf{m}_{ij} = \mathbf{c} \times \mathbf{d}_{ij}. \tag{2}$$

We form dense Plücker embeddings $\mathbf{F}_{\text{plucker}} \in \mathbb{R}^{6\times H\times W}$ by concatenating $(\mathbf{d}_{ij}, \mathbf{m}_{ij})$ for each pixel of an image of size (H, W). Because $(\mathbf{d}, \mathbf{m})$ are homogeneous, $s(\mathbf{d}, \mathbf{m})$ with $s \neq 0$ represents the same (unoriented) line. For rays, we fix this gauge by normalizing $\|\mathbf{d}\| = 1$ and choosing the sign so that $\mathbf{d}$ points from the camera into the scene.

This representation encodes the camera geometry in a distributed way: instead of a single global pose vector, each pixel's ray is tagged with a 6D vector that implicitly contains the camera intrinsics and extrinsics for that viewline. Thus, the diffusion model can, in principle, attend to the 3D configuration of rays when conditioning on the view.

Plücker coordinates also admit simple expressions for line-line relations. Given two rays $\ell_k = (\mathbf{d}_k, \mathbf{m}_k)$, the reciprocal product

$$\langle \ell_1, \ell_2 \rangle \;=\; \mathbf{d}_1\cdot\mathbf{m}_2 \;+\; \mathbf{d}_2\cdot\mathbf{m}_1 \tag{3}$$

vanishes iff the two rays are coplanar (i.e., they intersect or are parallel) [5]. For non-parallel rays, their shortest distance is

$$d(\ell_1, \ell_2) \;=\; \frac{\left|\mathbf{d}_1\cdot\mathbf{m}_2 + \mathbf{d}_2\cdot\mathbf{m}_1\right|}{\|\mathbf{d}_1 \times \mathbf{d}_2\|}. \tag{4}$$

Rays that image the same 3D point are coplanar and intersect, and rays to nearby points have small reciprocal product and line-line distance. Importantly, these quantities are bilinear in $(\mathbf{d}, \mathbf{m})$, so they can be implemented by linear projections and dot products on $\mathbf{F}_{\text{plucker}}$.

Although the computation of such expressions is not enforced explicitly in the network, Plücker ray embeddings provide an inductive bias that makes cross-view geometric consistency easy to test in the input space. Empirically, such per-pixel ray conditioning has been shown to improve camera pose accuracy and 3D consistency in generative models [8, 34], and we observe similar benefits. By contrast, global pose encodings (e.g., Euler angles) offer no direct notion of pixel-to-pixel correspondences across views and often suffer from symmetries such as front-back ambiguity.

PRoPE attention. Besides Plücker-ray conditioning, we also evaluate an attention-level camera encoding based on Geometric Transform Attention (GTA) [15], CaPE [10], RoPE [21], and the recent PRoPE [12]. In this variant, each token t from camera $i(t)$ is associated with (i) a projective transform derived from the camera's projection matrix and (ii) a rotary position embedding of its 2D image coordinates (x_t, y_t). Following the GTA framework, these per-token transforms D_t are multiplied into Q, K, V so that attention logits depend on the *relative* projective transform between the cameras of the query and key tokens, while RoPE handles within-image spatial relations.

We apply PRoPE-style attention in all cross-attention layers where the denoising stream attends to reference views. The PRoPE mechanism complements Plücker embeddings: both encode the full camera frustum (intrinsics + extrinsics)

Figure 2. **Qualitative results of multiple inferences with varying baseline.** StereoSpace naturally supports rendering images captured with arbitrary baselines, including viewpoints located to the left (negative baseline) and to the right (positive baseline) of the source image.

but at different stages: Plücker rays provide per-pixel geometry in the input channels and ResNet blocks, while PRoPE aligns token features via the relative projective transform inside attention. However, as reported in our ablation study, the combination of the two does not produce any significant improvement, but Plücker embeddings alone achieve both the lowest iSQoE and MEt3R scores.

3. Evaluation Sets Composition

We evaluate on four real-world stereo benchmarks: Middlebury 2014 [19], DrivingStereo [32], Booster [17], and LayeredFlow [31], none of these datasets are used during training. For Middlebury 2014, we follow the official evaluation protocol and use the same split as GenStereo and StereoDiffusion. For DrivingStereo, we uniformly sample 50 stereo pairs from the 500 released pairs for each weather condition, yielding 200 evaluation images in total. For Booster, we report results on the union of the official train and test splits. For LayeredFlow, we evaluate on the 300 real-world stereo pairs from the official validation and evaluation splits of the benchmark. In our mixed training set, we only include the separate synthetic stereo split of LayeredFlow rendered in Blender.

4. Evaluation with Conventional Metrics

For the sake of completeness, we report classical metrics such as PSNR, SSIM, and LPIPS in Tab. 2. In most cases, StereoSpaces ranks 2nd, just behind either Lyra or GenStereo. However, as discussed in the paper, these metrics fail at faithfully assessing both geometric consistency and perceptual comfort.

5. Qualitative Results

We conclude by reporting additional qualitative results.

Native Multi-baseline Inference. Thanks to our viewpoint conditioning mechanism and training schedule, StereoSpace natively supports the generation images at arbitrary horizontal baselines on either side of the input view (Fig. 2). Warping-based frameworks can also be extended to this setting, but require either manually rescaling the monocular disparity used for warping or flipping the image to synthesize views on the opposite side of the reference image.

MEt3R Score Map Visualization. To better illustrate the margin in MEt3R [1] scores across methods, we visualize the per-pixel score maps produced by MEt3R before averaging, which makes local differences between methods directly observable. Fig. 3 shows results on five samples

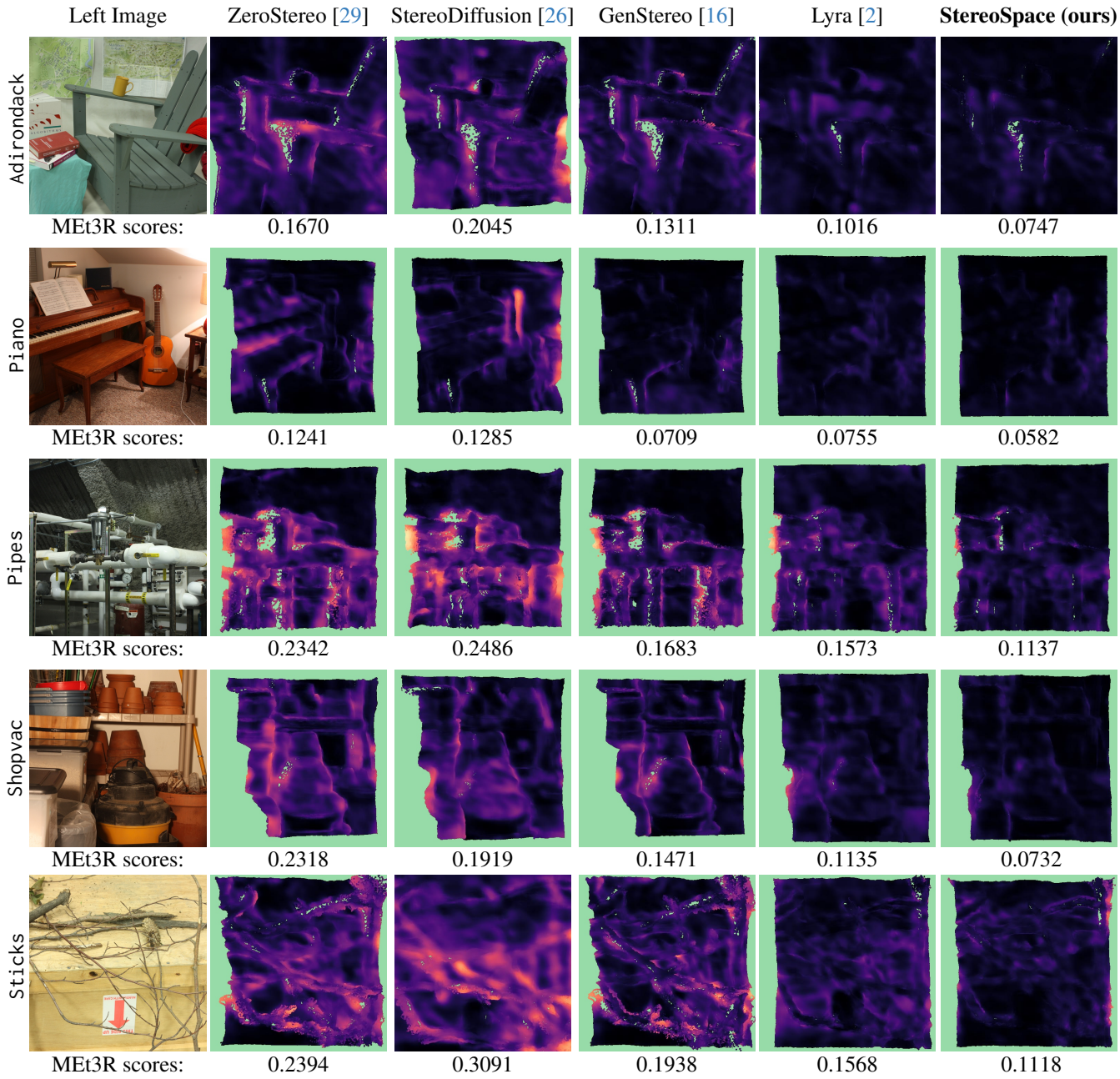

Figure 3. **Visualization of MEt3R score [1] maps on Middlebury dataset [19].** We report, from left to right, the original left image for four samples in the dataset, followed by the MEt3R score maps computed between it and the right images generated by different methods. The coloring is according to the `magma` colormap, with green regions representing occlusions (discarded by MEt3R when computing the average score). Under each score map, we report the global score computed by MEt3R (the lower, the better).

from the Middlebury dataset. Overall, warping-based methods produce higher MEt3R scores, particularly near depth discontinuities, whereas both Lyra and StereoSpace exhibit substantially lower scores. We also observe that the overall quality of the generated images affects the ability of MEt3R (through its MASt3R [11] backbone) to accurately estimate the relative transformation between the original and synthesized images. This is visible in the green regions, which represent areas of the scene that are not overlapping between the two images. Warping-based solutions show larger non-overlapping regions at both the top and bottom of the frame, although these areas should overlap in a stereo setup. By contrast, such regions are often much narrower (or absent) in the Lyra and StereoSpace score maps.

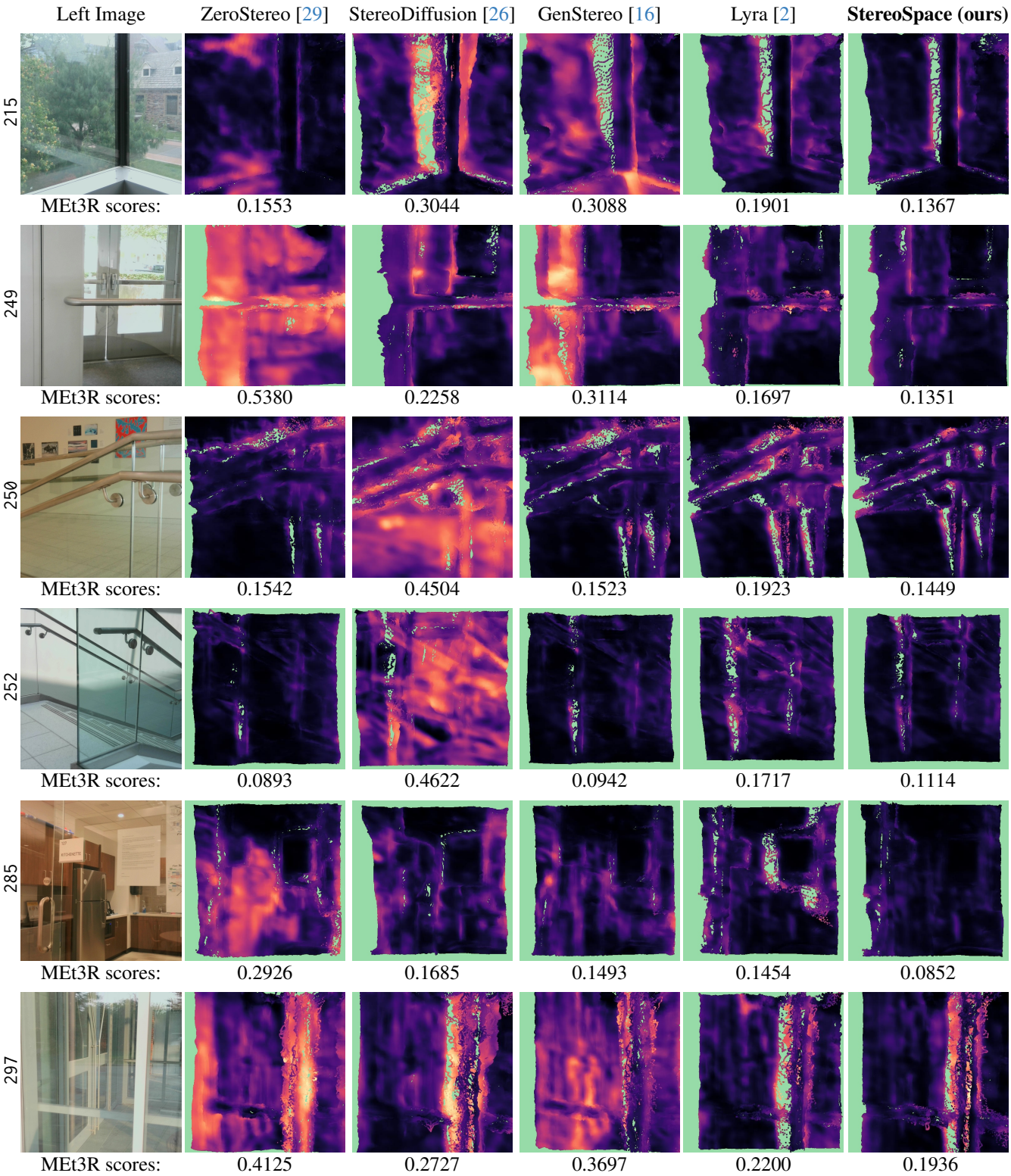

Figure 4. **Visualization of MEt3R score [1] maps on LayeredFlow dataset [31].** We report, from left to right, the original left image for four samples in the dataset, followed by the MEt3R score maps computed between it and the right images generated by different methods. The coloring is according to the `magma` colormap, with green regions representing occlusions (discarded by MEt3R when computing the average score). Under each score map, we report the global score computed by MEt3R (the lower, the better).

Figure 5. **Qualitative results on Middlebury [19] and DrivingStereo [32] datasets.**

Fig. 4 collects six samples from the LayeredFlow dataset [31]. The multi-layered geometry peculiar to these scenes leads to a significant increase in the MEt3R scores, in particular for warping-based solutions. As discussed, this effect is mostly related to the use of estimated depth, which fails to account for the multiple layers in the scene. On the contrary, StereoSpace shines in these cases, as it does not inherit the limitations inherent to warping-based methods.

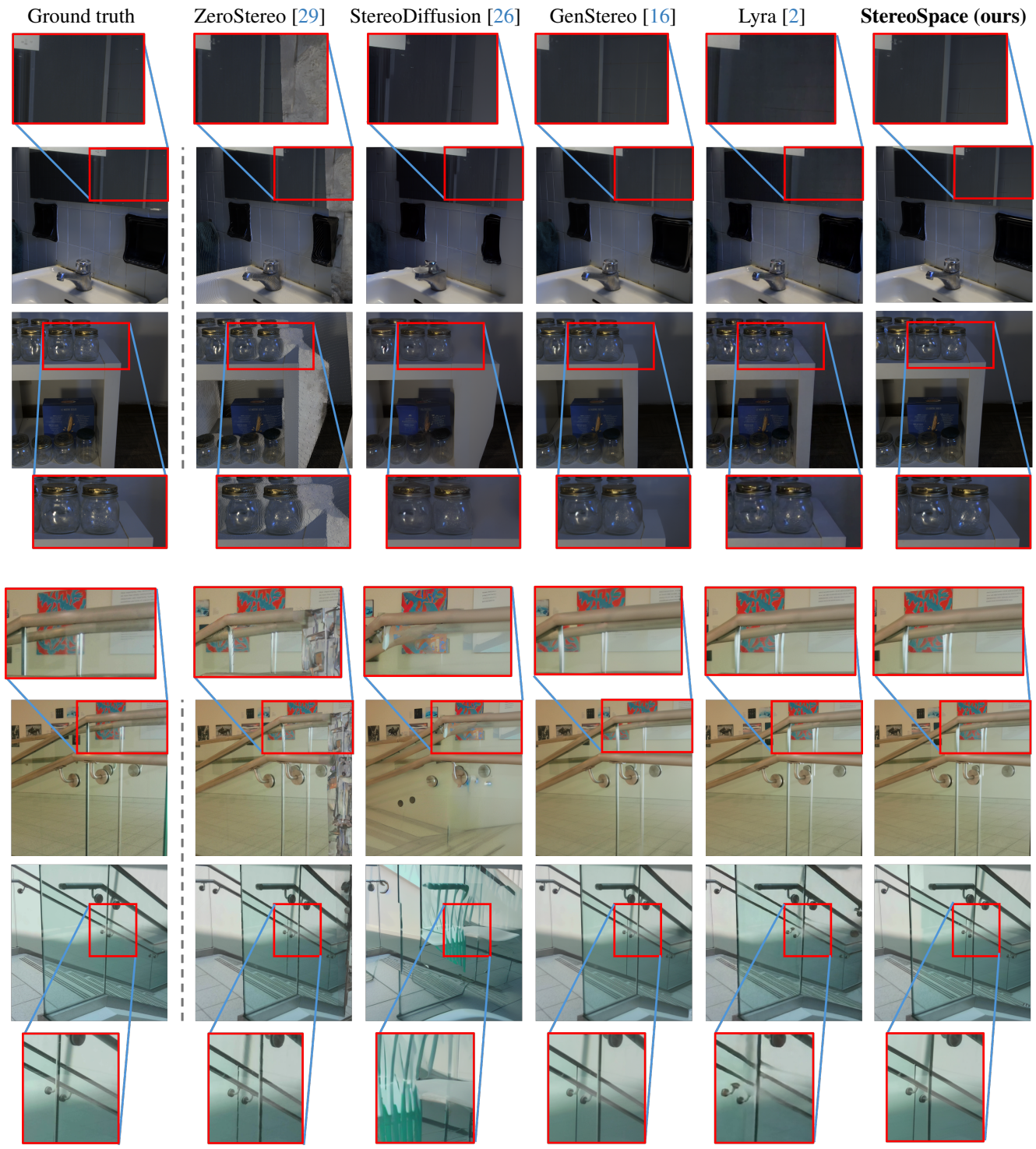

Figure 6. **Qualitative results on Booster [17] and LayeredFlow [31] datasets.**

Further Qualitative Comparisons. Finally, to highlight the superior realism of StereoSpace, we present extensive qualitative comparisons with images generated by all competing methods considered in our evaluation. Fig. 5 reports two samples from Middlebury [19] and two from the DrivingStereo [32] datasets. In the Middlebury examples, we highlight the bleeding artifacts introduced by ZeroStereo, as well as the interpolation effects between foreground and background objects produced by StereoDiffusion, clearly visible in the Shopvac scene, and both caused by warping.

We can also appreciate how Lyra itself, despite its high realism, tends to introduce oversmoothing. On DrivingStereo, we observe fewer artifacts due to simpler scene geometry and smaller disparities from larger depths. Nevertheless, StereoSpace still demonstrates superior reconstruction of thin structures.

Fig. 6 shows two scenes from Booster [17] and two from LayeredFlow [31]. At the top, we can notice how reflective and transparent objects in general are a significant challenge for methods relying on estimated depth, as highlighted by the artifacts produced in correspondence with the mirror or the deformations visible in the jars. At the bottom, depth-based approaches again struggle to deal with transparent surfaces that induce multiple depth layers, while StereoSpace maintains more faithful geometry and appearance.

References

[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring multi-view consistency in generated images. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2025. 3, 4, 5

[2] Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *preprint arXiv:2509.19296*, 2025. 4, 5, 6, 7

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012. 1

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual Kitti 2. *preprint arXiv:2001.10773*, 2020. 1

[5] Yan-Bin Jia. Plücker coordinates for lines in the space. COMS 4770/5770 Notes, Iowa State University, 2024. Lecture notes. 2

[6] Junpeng Jing, Ye Mao, Anlan Qiu, and Krystian Mikolajczyk. Match stereo videos via bidirectional alignment. *preprint arXiv:2409.20283*, 2024. 1

[7] Laurent Jospin, Allen Antony, Lian Xu, Hamid Laga, Farid Boussaid, and Mohammed Bennamoun. Active-passive simstereo-benchmarking the cross-generalization capabilities of deep learning-based stereo methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[8] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. SPAD: Spatially aware multi-view diffusers. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2024. 2

[9] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2023. 1

[10] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. EscherNet: A generative model for scalable view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2024. 2

[11] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 4

[12] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *preprint arXiv:2507.10496*, 2025. 2

[13] Yue Li, Qi Ma, Runyi Yang, Huapeng Li, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Ender Konukoglu, Theo Gevers, et al. SceneSplat: Gaussian splatting-based scene understanding with vision-language pretraining. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1

[14] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2023. 1

[15] Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. GTA: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[16] Feng Qiao, Zhexiao Xiong, Eric Xing, and Nathan Jacobs. GenStereo: Towards open-world generation of stereo images and unsupervised matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2025. 1, 4, 5, 6, 7

[17] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2022. 2, 3, 7, 8

[18] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 1

[19] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42, 2014. 2, 3, 4, 6, 7

[20] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1

[21] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2

[22] Joshua Tokarsky, Ibrahim Abdulhafiz, Satya Ayyalasomayajula, Mostafa Mohsen, Navya G Rao, and Adam Forbes. PLT-D3: A high-fidelity dynamic driving simulation dataset for stereo depth and scene flow. *preprint arXiv:2406.07667*, 2024. 1

[23] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo mixture density networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2021. 1

[24] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. NeRF-supervised deep stereo. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2023. 1

[25] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition Workshops (CVPRW)*, 2018. 1

[26] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. StereoDiffusion: Training-free stereo image generation using latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2024. 4, 5, 6, 7

[27] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. IRS: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 1

[28] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual slam. In *IEE/RSJ International Conference on Intelligent Robotics and Systems*, 2020. 1

[29] Xianqi Wang, Hao Yang, Gangwei Xu, Junda Cheng, Min Lin, Yong Deng, Jinliang Zang, Yurui Chen, and Xin Yang. ZeroStereo: Zero-shot stereo matching from single images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4, 5, 6, 7

[30] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *CVPR*, 2025. 1

[31] Hongyu Wen, Erich Liang, and Jia Deng. LayeredFlow: A real-world benchmark for non-lambertian multi-layer optical flow. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 5, 6, 7, 8

[32] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2019. 2, 3, 6, 7

[33] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1

[34] Chuanxia Zheng and Andrea Vedaldi. Free3D: Consistent novel view synthesis without 3d representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recogition (CVPR)*, 2024. 2