HarmonyView: Harmonizing Consistency and Diversity in One-Image-to-3D

Appendix

A. More Related Work

The challenge of *one-image 3D generation* has recently attracted significant attention, with various approaches and methods proposed to address this complex problem [2]. In this section, we provide a brief review of the literature.

Classical 3D generative methods. Early works can be broadly categorized into two main groups: primitive-based approaches and depth estimation approaches. Primitive-based approaches [13, 34, 36], focus on the fitting of primitive 3D shapes to 2D images, seeking to align synthetic models with observed image features. They often employ iterative optimization to refine the pose and shape of the model until a satisfactory fit is achieved. On the other hand, depth estimation approaches [40, 43] typically follow a two-step process: They first use a monocular depth estimator (*e.g.*, MiDaS [28]) to predict the 3D geometry, which is then used to render artistic effects through multi-plane images [15, 31] or point clouds [24]. To address imperfections, a pre-trained inpainting model [41] is often applied to fill in missing holes. However, these early approaches may struggle with generalization to real-world data or new object categories.

3D native models. A line of research [5, 7, 8, 11, 13, 32, 38] follows an encoder-decoder framework for modeling the image-to-3D data distribution, which involves the use of global shape latent codes to directly encode the shape information from 3D assets (*e.g.*, ShapeNet [1], Pix3D [33]). In contrast, other works utilize local features and representation-specific 3D generative models that leverage priors constructed from 3D primitives in various formats: point clouds [10, 13, 23, 39, 42], voxels [3, 6, 7, 35], meshes [4, 12, 17, 21, 36], or parametric surfaces [14, 30]. While these *3D native* models show impressive performance, they often require extensive 3D data and are constrained to specific object classes within that data. They also suffer from quality degradation when handling real-world images due to domain disparities. Recently, Point-E [25] and Shap-E [16] propose learning text-to-3D diffusion models on large-scale 3D assets to mitigate some of these limitations.

B. Additional Experimental Setup

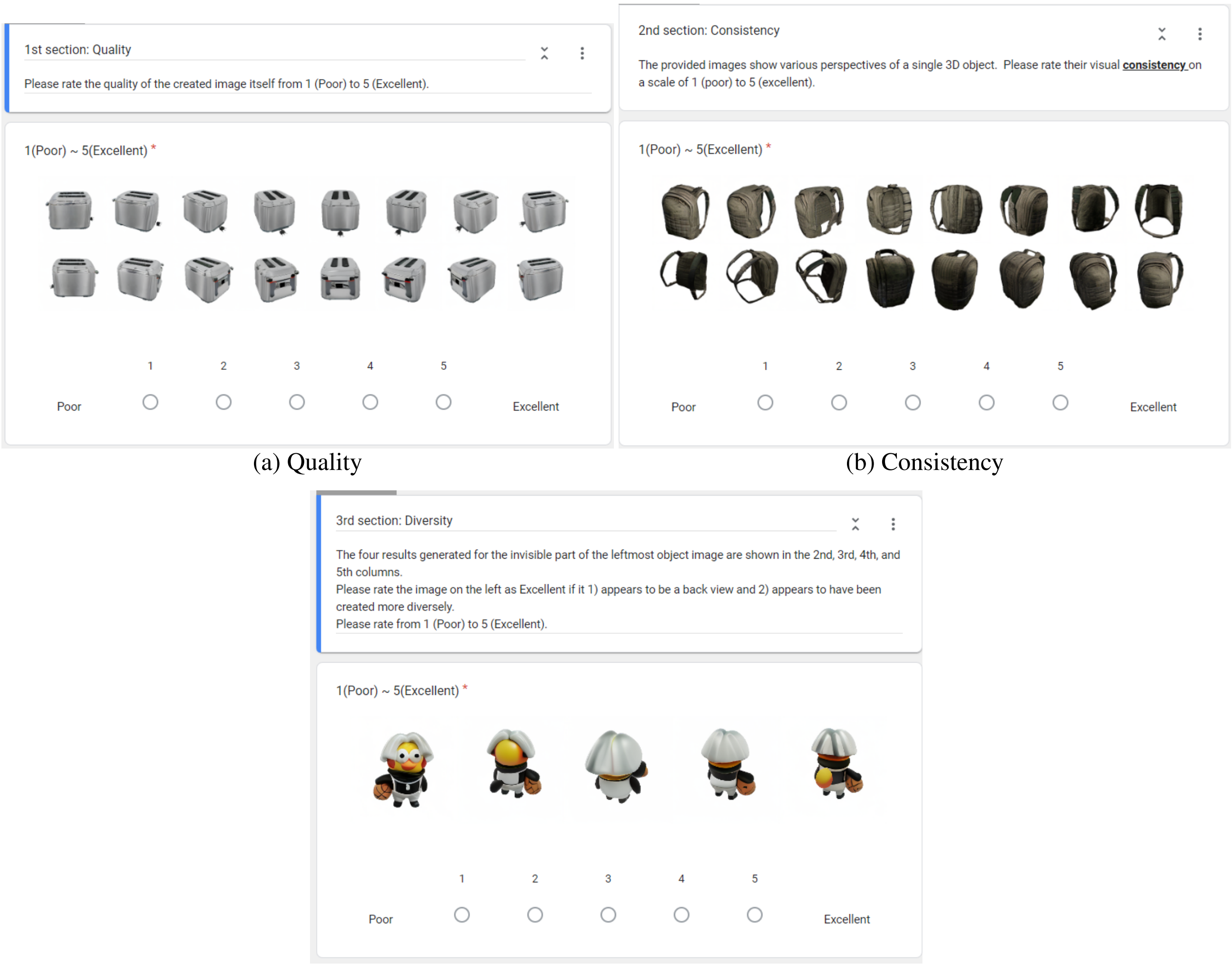
Diversity evaluation. Due to the inherent stochastic nature of diffusion models, the outputs they generate can be different w.r.t. the random seed used for their generation. Therefore, the computed metrics can differ depending on the seed we use. To evaluate the diversity of generated samples from each model, we randomly sample 4 instances using different random seeds from the same input image. We then use the CD score to quantify the diversity. By calculating the CD score across these sampled instances (each derived from a different random seed but originating from the sample input images), we obtain an average CD score. This average CD score represents the overall dissimilarity or diversity observed among the generated samples. The reported values in the main paper are the average CD score calculated across these sampled instances.

Technical details. HarmonyView is built upon the pre-trained models of SyncDreamer [20], which generates a set of $N = 16$ multi-view images, each with an elevation of 30° and azimuths evenly distributed in the range of $[0°, 360°]$. We assume that the azimuth of both the input view and the first target view is set to 0°. The viewpoint differences $\Delta\mathbf{v}^{(n)}$ are calculated based on the differences in elevation and azimuth between the input view and target view. At test time, similar to [19, 20, 22, 27], we estimate an elevation angle and use it as an input. To reconstruct the 3D mesh, we use foreground masks for generated images using CarveKit[1], and train the NeuS [37] for $2k$ steps. For text-to-image-to-3D, 2D images from the input text are created with the assistance of DALL-E-3[2].

Baselines. In our work, we employ several state-of-the-art methods as baseline models: Zero123 [19], RealFusion [22], Magic123 [27], One-2-3-45 [18], Point-E [25], Shap-E [16], and SyncDreamer [20]. Zero123 [19] is able to generate novel-view images of an object from various viewpoints given a single-view image. Moreover, its integration with the SDS loss [26] bolsters its capability for 3D reconstruction from single-view images. RealFusion [22] leverages Stable Diffusion [29] and the SDS loss for achieving high-quality single-view reconstruction. Magic123 [27] builds upon the strengths of Zero123 [19] and RealFusion [22], resulting in a method that further improves the overall quality of 3D reconstruction. One-2-3-45 [18] takes a direct approach by regressing Signed Distance Functions (SDFs) from the output images of Zero123 [19]. Point-E [25] and Shap-E [16] represent 3D generative models trained on an extensive 3D dataset. Both models exhibit the capability to convert a single-view image into either a point cloud or a shape encoded in an MLP. SyncDreamer [20] produces multi-view coherent images from a single-view image by synchronizing intermediate states of generated images using a 3D-aware feature attention mechanism.

[1] https://github.com/OPHoperHPO/image-background-remove-tool

[2] https://cdn.openai.com/papers/dall-e-3.pdf



1st section: Quality

Please rate the quality of the created image itself from 1 (Poor) to 5 (Excellent).

1(Poor) ~ 5(Excellent) *

	1	2	3	4	5	
Poor	○	○	○	○	○	Excellent

(a) Quality

2nd section: Consistency

The provided images show various perspectives of a single 3D object. Please rate their visual **consistency** on a scale of 1 (poor) to 5 (excellent).

1(Poor) ~ 5(Excellent) *

	1	2	3	4	5	
Poor	○	○	○	○	○	Excellent

(b) Consistency

3rd section: Diversity

The four results generated for the invisible part of the leftmost object image are shown in the 2nd, 3rd, 4th, and 5th columns.
Please rate the image on the left as Excellent if it 1) appears to be a back view and 2) appears to have been created more diversely.
Please rate from 1 (Poor) to 5 (Excellent).

1(Poor) ~ 5(Excellent) *

	1	2	3	4	5	
Poor	○	○	○	○	○	Excellent

(c) Diversity

Figure 1. **User evaluation examples.** We perform a user study to evaluate the effectiveness of our approach, HarmonyView, in comparison to SyncDreamer [20] and Zero123 [19]. Participants were asked to rate the three approaches using a 5-point Likert-scale (1-5), assessing (a) Quality, (b) Consistency, and (c) Diversity.

C. Correlation between CD score and Human Evaluation

To assess the efficacy of HarmonyView against SyncDreamer [20] and Zero123 [19], we conducted a user study where participants rated the three approaches using a 5-point Likert-scale (1-5), evaluating (a) Quality, (b) Consistency, and (c) Diversity. Our user study, showcased in Fig. 1, reveals a consistent alignment between the CD Score (CD Score = D/S_{Var}) and human evaluation metrics. Throughout the study, we observed that the CD Score reliably reflects the correlation between two key factors: S_{Var}, measuring the diversity in generated images' alignment with a given text prompt, and D, evaluating creative variation against a reference view using CLIP image encoders.

1. **Semantic Variance ($\mathbf{S}_{Var}$) and Consistency**: Lower Semantic Variance consistently corresponds to higher consistency in human evaluation. In simpler terms, when the generated images are more aligned in their interpretation of the text prompt, human evaluators tend to agree more on the perceived consistency. This correlation implies that there's a negative relationship between Semantic Variance and Consistency — lower variance often leads to higher agreement among evaluators.

2. **Diversity Score (D) and Quality Perception**: Higher Diversity Scores tend to lead to lower quality perceptions in human evaluation. This suggests a somewhat negative correlation between Diversity Score and Quality Perception. Put differently, when the diversity among the generated images is higher — meaning they deviate more from the reference image — human evaluators tend to perceive lower quality. Conversely, higher similarity between the generated images and the reference image correlates with higher perceived quality. In essence, when the visual similarity between the

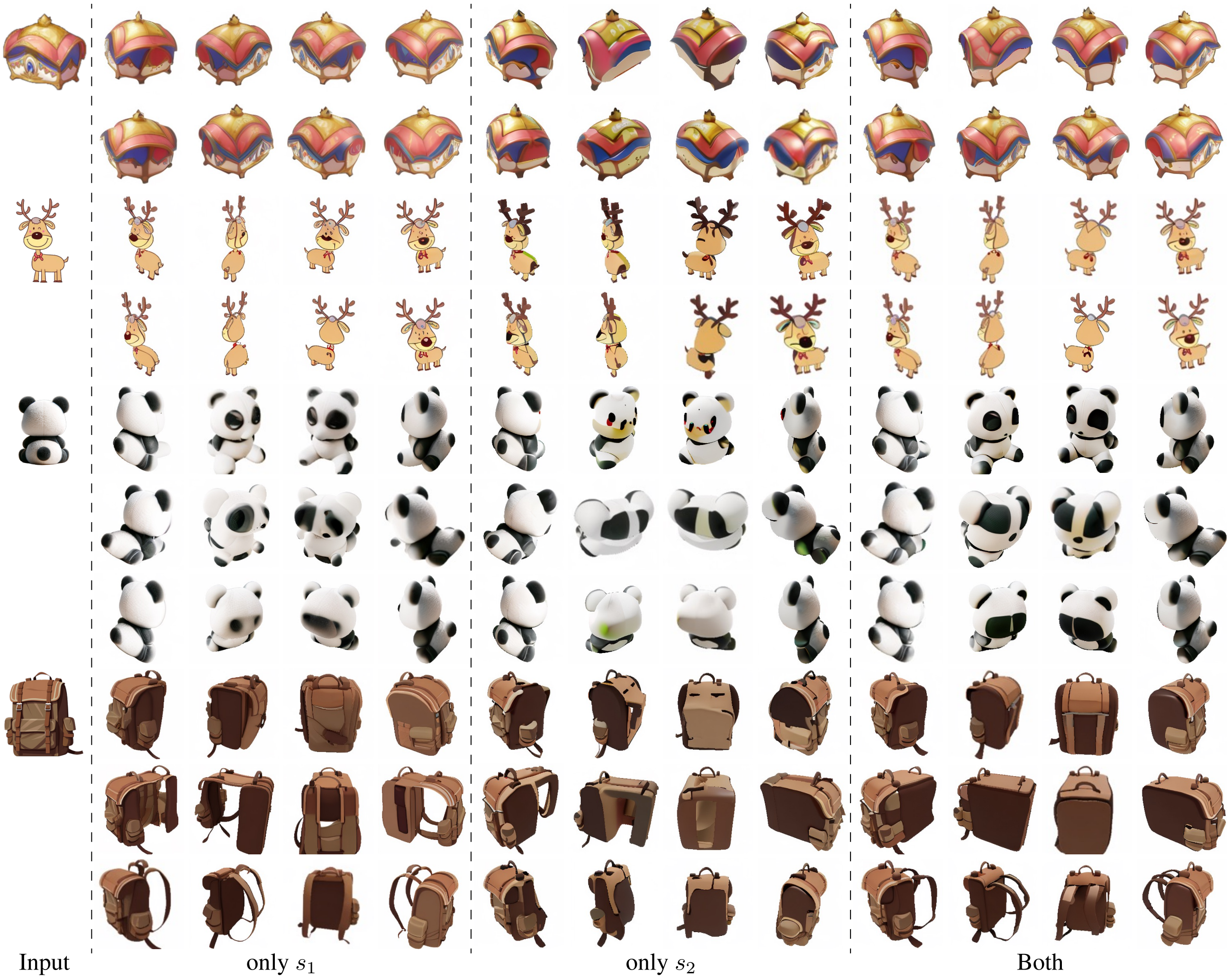


Figure 2. **Qualitative ablation study on novel-view synthesis.** Our HarmonyView guides the multi-view diffusion process with two parameters, s_1 and s_2 (see **??**). The nuanced interplay between s_1 and s_2 impacts consistency and diversity throughout the generation process. By skillfully balancing these guiding principles, we can achieve a win-win scenario: generate diverse images that maintain coherence across multiple views and stay faithful to the input view.

input and target views is higher, the quality tends to be perceived as better by human evaluators.

These findings collectively underscore the critical balance needed between semantic diversity and adherence to the reference image in the pursuit of generating high-quality images aligned with text prompts. Achieving this delicate equilibrium is pivotal to ensure that generated images are diverse enough to capture different interpretations while also being faithful enough to the reference to maintain perceived quality. HarmonyView demonstrated the highest CD score compared to SyncDreamer [20] and Zero123 [19], indicating that our generated images strike a winning balance between consistency and diversity, excelling in both aspects of fidelity to the reference image and semantic variation.

D. Additional Results

D.1. Novel-view Synthesis

Qualitative ablation study. Our HarmonyView decomposes multi-view diffusion guidance into two distinct guidance components (see Eq. (7)): s_1 primarily serves to ensure visual consistency between the input and target views, while s_2 focuses on amplifying diversity across novel viewpoints. The significance of this approach is showcased in Fig. 2, where we visually demonstrate how each guidance factor influences the synthesized images. When prioritizing s_1, the quality of synthesis improves significantly as it focuses on aligning the visual consistency between the input and target views. However, in specific cases, like the deer sample, it generates multiple faces of the deer, leading to what's known as the "Janus problem" — creating facial features on the rear

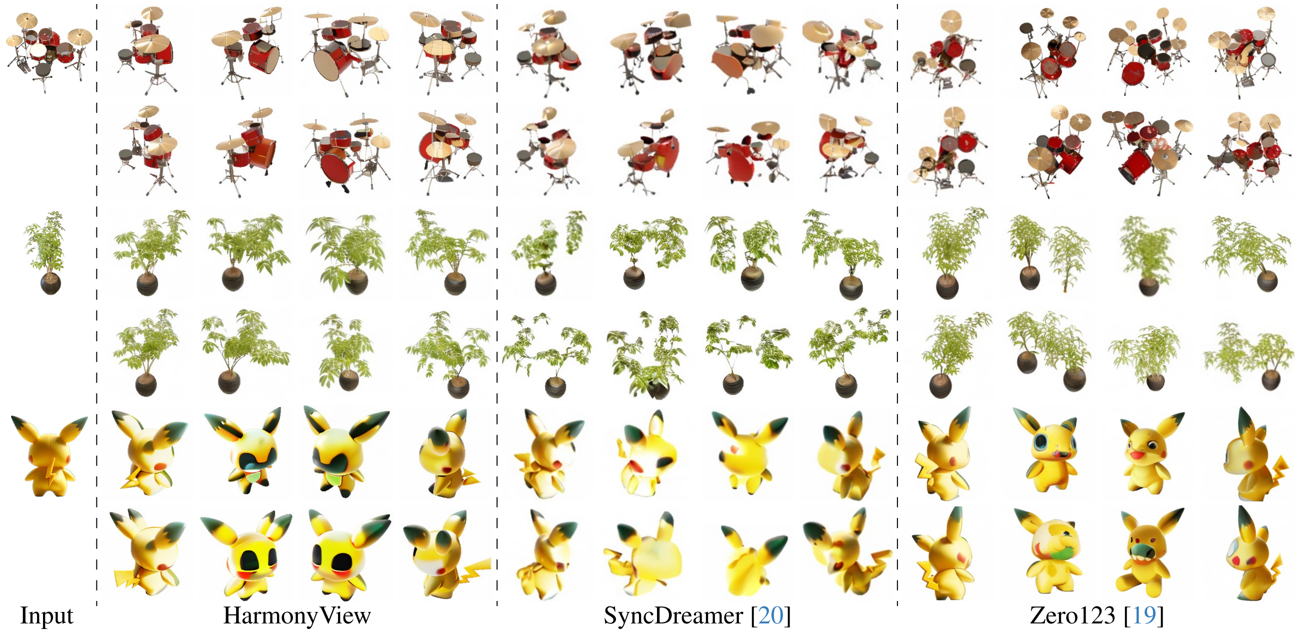


Figure 3. **Additional novel-view synthesis comparison.** HarmonyView creates diverse, coherent multi-view images for complex scenes, effortlessly generating realistic front views from rear-view input images.

Method	PSNR↑			SSIM↑			LPIPS↓			E_{flow}↓		
	Best	Avg.	Var.	Best	Avg.	Var.*	Best	Avg.	Var.*	Best	Avg.	Var.
Zero123 [19]	18.98	18.79	0.048	0.795	0.792	1.003	0.166	0.170	2.025	3.820	4.185	0.197
SyncDreamer [20]	20.19	19.74	0.242	0.819	0.813	4.465	0.140	0.148	7.922	2.071	2.446	0.458
HarmonyView	**20.69**	**20.24**	**0.260**	**0.825**	**0.819**	**5.295**	**0.133**	**0.140**	**8.038**	**1.945**	**2.350**	**0.510**

Table 1. **Statistical analysis of novel-view synthesis on GSO [9] dataset.** We report PSNR, SSIM, LPIPS, and E_{flow} for the best-matched instance with GT, as well as the average and variance across four instances. The variances marked as $*$ are reported with scaling by 10^{-5}.

side akin to the front, causing visual anomalies. On the other hand, emphasizing s_2 results in increased diversity across the generated samples. However, a fundamental trade-off exists between these two aspects — quality and diversity — making it challenging to optimize for both simultaneously. Yet, by employing both s_1 and s_2 in tandem, we can achieve a win-win scenario. This division allows us to precisely discern the impact of each guidance factor on the generation process. By skillfully balancing these guiding principles, our method becomes empowered to generate a rich and varied array of images, exhibiting both multi-view coherence and fidelity to the input view.

Qualitative comparison. Figure 3 provides a glimpse into the capabilities and limitations of different novel-view synthesis methods. Zero123 [19] frequently generates images that lack coherence across multiple viewpoints. These synthesized images often contain implausible variations, such as alterations in the number of cymbals or trees based on the view, or even changes in the shape of eyes. These inconsistencies underscore the struggle of Zero123 to maintain coherence and realism across different perspectives, leading to discrepancies that compromise the overall quality of multi-view synthesis. SyncDreamer [20] faces challenges in preserving the expected visual similarity across different viewpoints. The generated images often display deviations in overall size, empty or missing regions, or distorted forms, leading to an overall loss of visual completeness and integrity. Instances where facial features are erased or distorted represent the difficulties SyncDreamer encounters in maintaining the visual fidelity expected across diverse views. In stark contrast, HarmonyView stands out for its ability to generate diverse yet plausible multi-view images while preserving geometric coherence across these views. Unlike its counterparts, HarmonyView maintains a harmonious relationship between different views, ensuring the consistent appearance, shapes, and elements of objects. In addition, HarmonyView can extrapolate realistic frontal views from the rear-view input image (see third sample). This further underscores the versatility and robustness of HarmonyView. Overall, HarmonyView is able to generate a diverse set of images while maintaining a sense of realism and coherence across the multiple views.

Statistical analysis. In Table 1, we conduct a comprehensive statistical analysis on the GSO [9] dataset, eval-

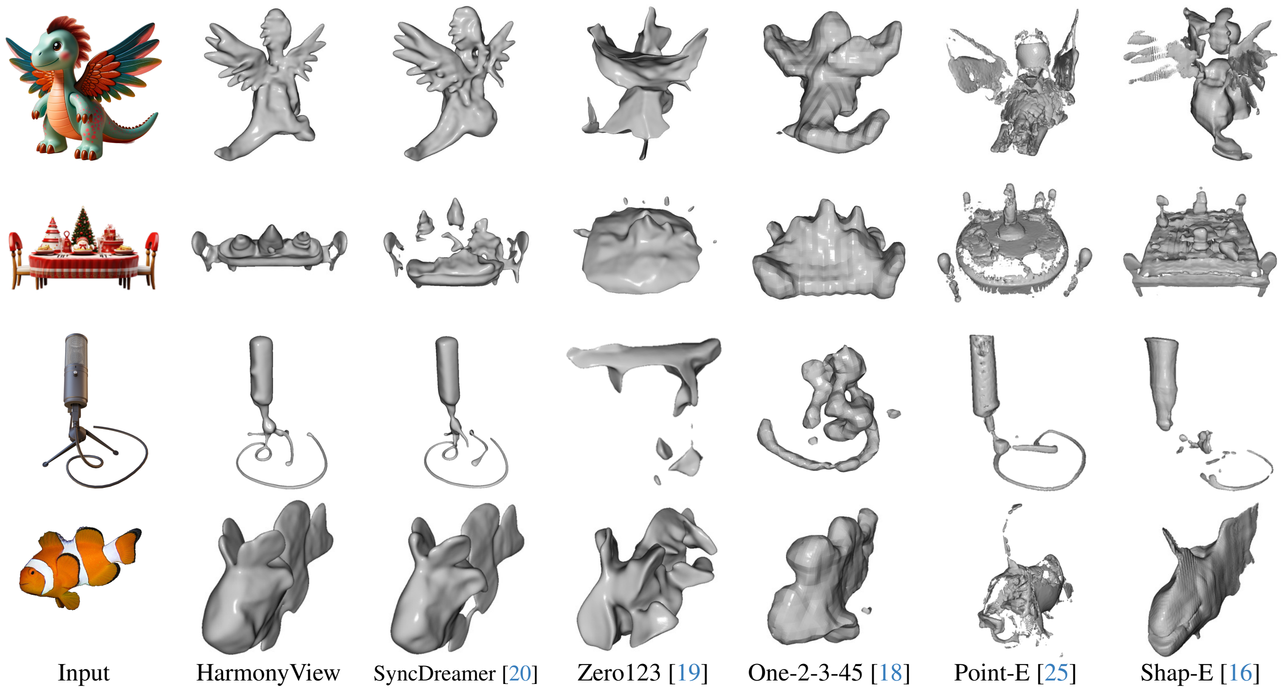


Figure 4. **Additional 3D reconstruction comparison.** HarmonyView excels at generating high-fidelity 3D meshes that achieve precise geometry with a realistic appearance while sidestepping common pitfalls for comprehensive and captivating reconstructions.

uating the performance of three methods: HarmonyView, Zeor123 [19], and SyncDreamer [20]. We report PSNR, SSIM, LPIPS, and E_{flow} for the best-matched instance with ground truth, as well as the average and variance across four instances. Upon comparison, HarmonyView demonstrates superior performance across all metrics when compared to Zeor123 and SyncDreamer. It attains the highest scores in PSNR and SSIM, indicating better image quality in terms of both fidelity and structural similarity when compared to the ground truth. Moreover, HarmonyView also exhibits the lowest LPIPS and E_{flow} scores, signifying reduced perceptual differences and flow errors when matched against the ground truth. Interestingly, HarmonyView shows higher variability (indicated by larger variance values) across instances compared to other methods. This variability might imply that while HarmonyView generally performs well, its performance might fluctuate more across different instances or scenarios compared to the Zero123 and SyncDreamer. Nevertheless, it is essential to note that this variability in performance also reflects its diversity in samples. This could imply that while HarmonyView showcases a broader range of outputs, it still maintains a high level of image quality.

D.2. 3D Reconstruction

In Fig. 4, the results exemplify HarmonyView's exceptional quality compared to other methods evaluated for 3D reconstruction. While contrasting with competing methods, it is evident that these approaches encounter various challenges in handling the reconstruction process. For instance, both Point-E [25] and Shap-E [16] struggle significantly with incomplete reconstructions, failing to capture the entirety of the intended 3D shapes. This deficiency results in reconstructions that lack certain crucial elements, undermining the fidelity of the output. In the case of One-2-3-45 [18], the method exhibits a tendency to produce ambiguous shapes, failing to accurately represent the intended shape contours. Furthermore, Zero123 [19] faces difficulties in capturing fine elements within the reconstructed shapes, which diminishes the overall fidelity and detail level of the output. SyncDremaer [20] also shows discontinuities or holes within the generated 3D meshes. These imperfections detract from the coherence and completeness of the reconstructed shape. In contrast, HarmonyView produces high-quality 3D meshes that achieve accurate geometry while maintaining a realistic appearance. Its ability to circumvent the pitfalls experienced by other methods speaks volumes about its capability to generate comprehensive, detailed, and visually compelling reconstructions.

E. Discussion

E.1. Limitations & Future Work

While HarmonyView demonstrates promising results in enhancing both visual consistency and novel-view diversity in single-image 3D content generation, several limitations warrant further investigation. Firstly, our multi-view diffusion formulation somewhat mitigates inherent trade-offs between consistency and diversity to achieve a certain level of Pareto optimality. However, the complete separation of these aspects to eliminate the trade-off entirely remains a challenging pursuit. Secondly, HarmonyView's current focus primarily revolves around object-centric scenes. This poses limitations when dealing with complex scenarios involving multiple interacting objects, varying scales, and intricate geometries. Expanding the technique to encompass such diverse and intricate scenes demands innovative approaches that account for object interactions, spatial relationships, and contextual understanding within the scene. Moreover, our current setting typically involves single objects without backgrounds, simplifying the requirements for realism and diversity. The ignorance of background significantly reduces the expectations of synthesizing diverse images. To accommodate in-the-wild multi-object scenes with complex backgrounds, HarmonyView requires the use of an external background removal tool (*e.g.*, CarveKit). Addressing these limitations effectively presents ample opportunities for innovation and refinement within the field. Exploring these avenues promises to advance the field towards more comprehensive and realistic 3D content generation from single images.

E.2. Ethical Considerations

The advancements in one-image-to-3D bring forth several ethical considerations that demand careful attention. One key concern is the potential misuse of generated 3D content. These advancements could be exploited to create deceptive or misleading visual information, leading to misinformation or even malicious activities like deepfakes, where fabricated content is passed off as genuine, potentially causing harm, misinformation, or manipulation. It is essential to establish responsible usage guidelines and ethical standards to prevent the abuse of this technology. Another critical concern is the inherent bias within the training data, which might lead to biased representations or unfair outcomes. Ensuring diverse and representative training datasets and continuously monitoring and addressing biases are essential to mitigate such risks. Moreover, the technology poses privacy implications, as it could be used to reconstruct 3D models of objects and scenes from any images. Images taken without consent or from public spaces could be used to reconstruct detailed 3D models, potentially violating personal privacy boundaries. As such, it is crucial to implement appropriate safeguards and obtain informed consent when working with images containing personal information.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[2] Zhiqin Chen. A review of deep learning-powered mesh reconstruction methods. *arXiv preprint arXiv:2303.02879*, 2023. 1

[3] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1

[4] Zhiqin Chen and Hao Zhang. Neural marching cubes. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021. 1

[5] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 45–54, 2020. 1

[6] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 1

[7] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 1

[8] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31–44, 2020. 1

[9] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 4

[10] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 1

[11] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020. 1

[12] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 1

[13] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 1

[14] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 1

[15] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T

Freeman, David Salesin, Brian Curless, et al. Slide: Single image 3d photography with soft layering and depth-aware inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12518–12527, 2021. 1

[16] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 5

[17] Yiyi Liao, Simon Donne, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2916–2925, 2018. 1

[18] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 1, 5

[19] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2, 3, 4, 5

[20] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 2, 3, 4, 5

[21] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023. 1

[22] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 1

[23] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023. 1

[24] Fangzhou Mu, Jian Wang, Yicheng Wu, and Yin Li. 3d photo stylization: Learning to generate stylized novel views from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16273–16282, 2022. 1

[25] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 1, 5

[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[27] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1

[28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[29] 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*, pages 10684–10695, 2022. 1

[30] Gopal Sharma, Difan Liu, Subhransu Maji, Evangelos Kalogerakis, Siddhartha Chaudhuri, and Radomír Měch. Parsenet: A parametric surface fitting network for 3d point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 261–276. Springer, 2020. 1

[31] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 1

[32] Ayan Sinha, Asim Unmesh, Qixing Huang, and Karthik Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6040–6049, 2017. 1

[33] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 1

[34] Shubham Tulsiani, Abhishek Kar, Joao Carreira, and Jitendra Malik. Learning category-specific deformable 3d models for object reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):719–731, 2016. 1

[35] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017. 1

[36] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 1

[37] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[38] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018. 1

[39] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 829–838, 2020. 1

[40] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1–10, 2020. 1

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 1

[42] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 1

[43] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. *Advances in neural information processing systems*, 31, 2018. 1