POLITECNICO DI TORINO
Repository ISTITUZIONALE

DiMViS: Diffusion-based Multi-View Synthesis

Original
DiMViS: Diffusion-based Multi-View Synthesis / DI GIACOMO, Giuseppe; Franzese, Giulio; Cerquitelli, Tania; Chiasserini, Carla Fabiana; Michiardi, Pietro. - STAMPA. - (2024). (Intervento presentato al convegno ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling 2nd SPIGM @ ICML tenutosi a Vienna (Austria) nel 21-27 July 2024).

Availability:
This version is available at: 11583/2989596 since: 2024-12-17T10:17:20Z

Publisher:
ICML

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

22 December 2024

DIMVIS: Diffusion-based Multi-View Synthesis

Giuseppe Di Giacomo [1] **Giulio Franzese** [2] **Tania Cerquitelli** [3] **Carla Fabiana Chiasserini** [1 4] **Pietro Michiardi** [2]

Abstract

Multi-view observations offer a broader perception of the real world, compared to observations acquired from a single viewpoint. While existing multi-view 2D diffusion models for novel view synthesis typically rely on a single conditioning reference image, a limited number of methods accommodate a multiple number thereof, by explicitly conditioning the generation process through tailored attention mechanisms. In contrast, we introduce DIMVIS, a novel method enabling the conditional generation in multi-view settings by means of a joint diffusion model. DIMVIS capitalizes on a pre-trained diffusion model, while combining an innovative masked diffusion process to implicitly learn the underlying conditional data distribution, which endows our method with the ability to produce multiple images given a flexible number of reference views. Our experimental evaluation demonstrates DIMVIS's superior performance compared to current state-of-the-art methods, while achieving reference-to-target and target-to-target visual consistency.

1. Introduction

Generative models play a pivotal role in learning data distributions, which endows them with the capability of generating synthetic samples that closely resemble real-world data, such as, for example, images or text. Over the years, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational AutoEncoders (VAEs) (Kingma & Welling, 2013) have dominated the field; however, recently, diffusion models (Ho et al., 2020; Song et al., 2021) have emerged as a powerful alternative. Specifically, Latent Diffusion Models (LDMs), such as Stable Diffusion (Rombach et al., 2022), have become the state-of-the-art for image generation: large-scale training on billions of 2D images (Schuhmann et al., 2022) equips such models with rich semantic priors and, hence, a strong generalization capability.

Motivated by the impressive performance of diffusion models, many works trained 3D variants thereof; despite their remarkable results, 3D datasets are still inadequate compared to the large available 2D datasets. To fill this gap, Zero123 (Liu et al., 2023) capitalizes on the rich 2D priors provided by Stable Diffusion, leveraging multi-view image data to learn 3D priors: once trained, given a single-view image, Zero123 is able to generate images of the underlying object from different target viewpoints. Subsequent works build on Zero123 to improve the generated views consistency (Liu et al., 2024; Weng et al., 2023).

Such methods perform the task usually referred to as Novel View Synthesis (NVS) and try to emulate the experience-based human capability of inferring 3D shapes and occluded views from a *single* observed reference view. However, in real-world scenarios, observation may be acquired from different viewpoints, offering a more comprehensive understanding compared to single-view observations, which may fall short in capturing the complexity and diversity of an object or a scene. Driven by this motivation, (Li et al., 2023) extend Zero123, to allow multiple conditioning views for generating a novel image from a target viewpoint.

We provide a detailed review of the related work in Appendix A, where we underline that most existing approaches lack flexibility in accommodating a varying number of both input and output views, namely reference and generated images. This drawback hinders such methods' deployment in practical applications requiring higher flexibility. Consider a set of recording cameras used, for instance, to ensure people's safety or to improve a classification task: in these real-use case scenarios, cameras may malfunction or be occluded. In such circumstances, we are interested in recovering the missing views, considering the observed ones. Consequently, the number of input views and images that must be generated cannot be fixed *a priori*.

To this end, (Höllein et al., 2024; Tang et al., 2024b) propose a diffusion model based on a multi-branch U-Net, with branches sharing the same architecture and weights; a simi-

[1]Department of Electronics and Telecommunications, Politecnico di Torino, Italy [2]Department of Data Science, EURECOM, France [3]Department of Control and Computer Engineering, Politecnico di Torino, Italy [4]Chalmers University of Technology, Sweden. Correspondence to: Giuseppe Di Giacomo <giuseppe.digiacomo@polito.it>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

lar approach is also envisioned in (Kong et al., 2024). Importantly, these methods employ elaborate attention mechanisms to ensure consistency among conditioning and target views and across the target views themselves. Differently from this approach, we propose a novel method called **Diffusion-based Multi-View Synthesis (DIMVIS)** that we specifically design for *conditional generation by implementing a mechanism that allows latent variables to evolve according to different time values in the forward process, and that produces a correlation between latent variables in the backward process through a joint diffusion model*.

At inference time, given a fixed-size set of images, DIMVIS accepts any number of reference views to conditionally generate the missing ones. DIMVIS uses the pre-trained weights of Stable Diffusion Image Variations (Pinkney), while adopting both for training and inference the masked diffusion method introduced in (Bounoua et al., 2024), which addresses the problem of modeling multiple input modalities (image, audio, and text data) representing the same concept. Remarkably, DIMVIS does not need any addition or change with respect to the standard attention mechanism employed in the pre-trained model, as it relies on a joint diffusion model that allows latent image variables to mutually influence one another. Therefore, DIMVIS is able to effectively aggregate multiple reference views: when incorporating additional images, thanks to the larger conditioning information, DIMVIS performance is enhanced while the inherent stochasticity of the image generation process is reduced. We experimentally find that DIMVIS outperforms current state-of-the-art NVS 2D diffusion models while achieving visual consistency of the generated images.

To summarize, DIMVIS addresses the following key requirements:

- **Generalization** capability inherited by the strong 2D generative prior provided by Stable Diffusion;
- **Flexibility**, as it both accepts and generates multiple views by leveraging a joint diffusion model that seamlessly combines masked diffusion with the pre-trained model;
- **Efficient multi-view aggregation**, with improvement up to 15% in perceptual similarity when multiple views are available.

2. Preliminaries

For the sake of clarity, we summarize in this section the joint latent diffusion model introduced in (Bounoua et al., 2024), which we adapt to the NVS domain.

Given the set of input views $X{=}\{X^1, \ldots, X^V\}$, first, a (deterministic) encoder e_ϕ produces latent variable Z^v for each X^v. We denote with $q_\phi(z)$ the produced distribution of the concatenated latent variable $Z{=}[Z^1, \ldots, Z^V]$. A score-based diffusion model is then employed to learn $q_\phi(z)$, which endows the model with the capability to generate a new sample $\hat{Z}{=}[\hat{Z}^1, \ldots, \hat{Z}^V]$.

The score-based diffusion model relies on two stages, namely, the forward and the backward diffusion processes. The forward process is a stochastic noising process injecting noise into the input data, i.e., the latent representations, and is defined by the following Stochastic Differential Equation (SDE):

$$\mathrm{d}R_t = \alpha(t)R_t\mathrm{d}t + g(t)\mathrm{d}W_t, \quad R_0 = Z \sim q(r,0), \quad (1)$$

where $\alpha(t)R_t$ and $g(t)$ are the drift and diffusion terms, respectively. W_t is a Wiener process, while $q(r,t)$ denotes the time-varying probability density of the stochastic process at time $t{\in}[0,T]$, with finite T and initial conditions $q(r,0){=}q_\phi(r)$.

To generate a new sample, we need to reverse the noising process by simulating the reverse-time SDE:

$$\begin{aligned}\mathrm{d}R_t{=}\left(-\alpha(T{-}t)R_t{+}g^2(T{-}t)\nabla\log(q(R_t,T{-}t))\right)\mathrm{d}t \\ {+}g(T{-}t)\mathrm{d}W_t, \quad R_0{\sim}q(r,T). \quad (2)\end{aligned}$$

To solve Eq. 2, a parametric score network $s_\chi(r,t)$ is used to approximate the true score function; furthermore, $q(r,T)$ is approximated with the noise distribution $\epsilon{\sim}\mathcal{N}(0,I)$. Finally, a decoder d_ψ is used to map back the latent variables into the input space.

Conditional generation. Our model accommodates conditional generation: specifically, the model leverages masked forward and backward diffusion processes to produce samples from the conditional distribution $q_\phi(z^M \mid z^C)$, being C and M the sets of conditioning and missing views to be generated, and, hence, z^C and z^M the respective latent variables. Formally, we define the masked forward SDE as:

$$\begin{aligned}\mathrm{d}R_t = \mathcal{M}(M)\odot\left[\alpha(t)R_t\mathrm{d}t + g(t)\mathrm{d}W_t\right], \\ q(r,0) = q_\phi(r^M \mid z^C)\delta(r^C - z^C), \quad (3)\end{aligned}$$

where $R_0{=}\mathcal{C}(R_0^M, R_0^C)$, with $R_0^M{\sim}q_\phi(r^M \mid z^C)$, $R_0^C{=}z^C$, and $\mathcal{C}(\cdot)$ being the concatenation operator. Importantly, the mask $\mathcal{M}(M)$ is used to freeze or diffuse the latent variable z^C and z^M, respectively.

The reverse-time process of Eq. 3 is defined as follows:

$$\begin{aligned}\mathrm{d}R_t{=}\mathcal{M}(M)\odot\left[(-\alpha(T{-}t)R_t{+}\right. \\ \left.g^2(T{-}t)\nabla\log\big(q(R_t,T{-}t\mid z^C)\big)\right)\mathrm{d}t{+}g(T{-}t)\mathrm{d}W_t\Big], \quad (4)\end{aligned}$$

with $R_0{=}\mathcal{C}(R_0^M, z^C)$ and $R_0^M{\sim}q(r^M, T \mid z^C)$. Also in this case, $q(r^M, T \mid z^C)$ is approximated by its corresponding

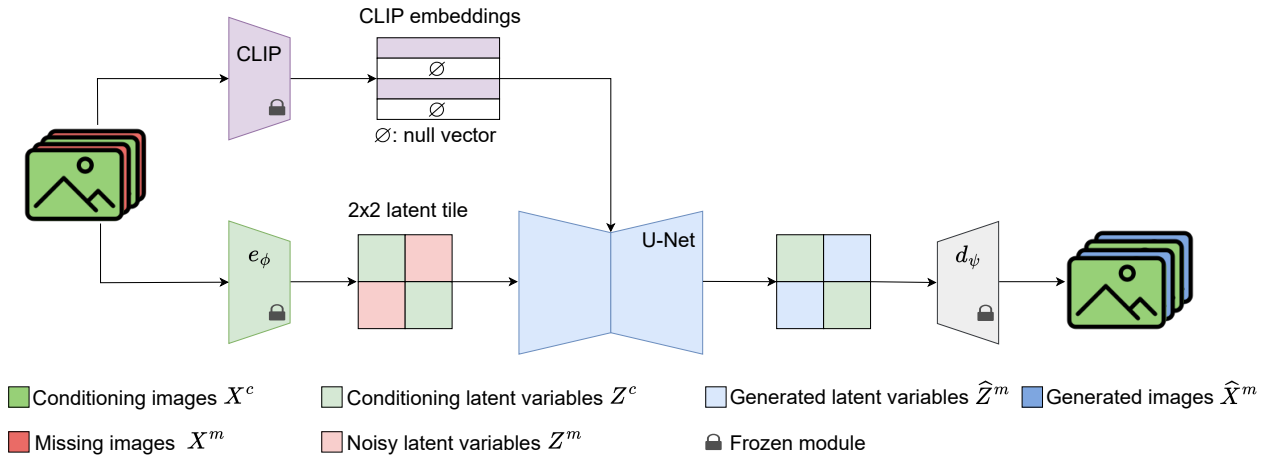

Figure 1. **DIMVIS Architecture**. Given a set of multi-view images, DIMVIS is able to generate the missing views given the conditioning observed views. In this example, we consider two conditioning images and two missing ones.

steady state distribution $\epsilon \sim \mathcal{N}(0, I)$, and the true conditional score function $\nabla \log\big(q(r, t \mid z^C)\big)$ is estimated with a conditional score network $s_\chi(r^M, t \mid z^C)$.

The joint diffusion model in (Bounoua et al., 2024) implements masked diffusion by using a multi *multi-time vector* $\tau = [t_1, \ldots, t_V]$, which concurrently indicates the diffusion time and which views are missing. Formally, the multi-time vector is defined as $\tau(M, t) = t\,\big[\mathbb{1}(1 \in M), \ldots, \mathbb{1}(V \in M)\big]$. Note that this formulation can be easily adapted to include conditioning signals, such as CLIP (Radford et al., 2021) embeddings.

Finally, the original continuous-time model definition from (Bounoua et al., 2024) can be cast in discrete-time, such as Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020). Indeed, continuous- and discrete-time diffusion converge to a mathematically equivalent expression with a sufficiently high number of integration steps. In DIMVIS we use a discrete-time diffusion formulation, which enables our approach to exploit pre-trained models such as Stable Diffusion (Rombach et al., 2022).

3. Our Method: DIMVIS

Given a set of V views, we define the generic subsets of conditioning, i.e., reference, and missing views as $X^C = \{X^c\}_{c \in C}$ and $X^M = \{X^m\}_{m \in M}$, where C and M are the sets of indices of conditioning and missing views, respectively. The goal is to generate the missing views given the reference ones: we model this problem as a conditional generation task, by using a *joint* diffusion model.

To exploit the generalization ability provided by the strong 2D prior acquired through large-scale training of pre-trained 2D diffusion models, we capitalize on the Stable Diffusion (Rombach et al., 2022) architecture; specifically, we initialize the model weights from Stable Diffusion Image Variations v2 (Pinkney), which was obtained by fine-tuning Stable Diffusion v1.4. However, we need to accommodate a variable number of reference and target views at inference; consequently, to learn the conditional data distribution, we adapt the original architecture to fine-tune it according to the Masked Multi-time diffusion approach introduced in (Bounoua et al., 2024) for the multi-modal domain.

At inference, our approach is depicted in Fig. 1, where we consider a set of $V = 4$ views in total and, for example, the set of missing views with $M = \{2, 4\}$. A deterministic encoder e_ϕ is used to encode each conditioning reference view X^c, with $C = \{1, 3\}$, to obtain their latent representations $Z^c = e_\phi(X^c)$; on the other hand, the missing views latent variables are represented using random noise, sampling from the Normal distribution $\mathcal{N}(0, I)$. Such latent representations are concatenated along the height and width axes, forming a 2×2 tile of latent variables, which is the input of the U-Net.

Our architecture uses a conditioning branch containing a CLIP image encoder. The CLIP embeddings obtained from the reference views are concatenated together with the null vectors corresponding to the missing views embeddings and injected into the cross-attention modules of the U-Net. The resulting concatenated embeddings vector has two purposes: it both contains the conditioning signals relative to the reference views and, thanks to the null vectors, indicates which views are missing. In (Bounoua et al., 2024), the information concerning the missing modalities is provided by a multi-time vector; however, by using the CLIP embed-

dings, we integrate such conditioning signal without any modification to the U-Net, which instead would be necessary with the multi-time vector, which can interfere with the pre-trained model.

Finally, the U-Net generates the latent variables $\hat{Z}^m$ of the missing views X^m, and a deterministic decoder d_ψ transforms the generated latent variables back into the input space, obtaining the generated images $\hat{X}^m{=}d_\psi(\hat{Z}^m)$.

3.1. Training procedure

During training, we freeze the encoder e_ϕ, the decoder d_ψ and the CLIP image encoder, and we fine-tune the whole U-Net, to learn the conditional distribution of the missing latent variables $Z^M{=}\{Z^m\}_{m\in M}{\sim}p(z^M|z^C)$. To do so, we employ the masked diffusion approach introduced in (Bounoua et al., 2024), by using a complete training set and randomly setting some views as missing during the fine-tuning. Specifically, with probability $d{=}0.2$ we set $C{=}\emptyset$, i.e., there is no conditioning view, hence, we diffuse all latent variables and all the CLIP embeddings are null-vectors; on the other hand, with probability $1{-}d$, we perform masked diffusion: first, we uniformly sample the set of conditioning views over all the possible sets; then, the remaining views, which are assumed to be missing, are diffused and their CLIP embedding is set to the null-vector, while the latent variables of the conditioning views are frozen.

Formally, we fine-tune the denoiser U-Net ϵ_θ by minimizing the following loss:

$$\mathcal{L} = \lambda(M,C)\|\mathcal{M}(M)\odot[\epsilon-\epsilon_\theta(R_t,t,\text{CLIP}(M,X))]\|_2^2, \tag{5}$$

where R_t is obtained by first sampling from the distribution $q(r\,|\,Z,t)$ and aggregating it with the input Z using the mask $\mathcal{M}(M)$. Importantly, we use a scaling factor $\lambda(M,C){=}1{+}\frac{|C|}{|M|}$ to take into account the randomization of M and C that leads to the diffusion of different portions of the latent space.

4. Experiments

4.1. Training details

We fine-tune DIMVIS on Objaverse-1.0 (Deitke et al., 2023), a large object-centric dataset containing more than 800k 3D models, which is commonly used for multi-view diffusion models. However, the original dataset contains many samples with poor quality, that may negatively affect the model training; for this reason, we filter the dataset as in (Tang et al., 2024a), obtaining a subset of ∼82k objects. For each object, we render 16 images with azimuth evenly distributed from 0° to 360° and elevation view fixed to 30°.

To speed up fine-tuning, we first build the latent dataset by randomly sampling for every object a set of 4 images having azimuth offset by multiple of 90°; importantly, the latent variables are deterministically generated by using the mode of posterior distribution produced by the frozen encoder e_ϕ employed in the Stable Diffusion variational autoencoder.

We use the AdamW (Loshchilov & Hutter, 2018) optimizer with peak learning rate $5{\times}10^{-5}$, and cosine annealing with 100 warm-up steps. The total batch size is 112, with gradient accumulation over 12 batches, and image resolution is $256{\times}256$. We fine-tune our model for 15750 steps on 7 NVIDIA A100-80GB GPUs, which take about 5 days.

4.2. Evaluation

Baseline methods. We compare DIMVIS to two state-of-the-art NVS 2D diffusion models, namely Zero123-XL and EscherNet (Kong et al., 2024), by using their publicly available pre-trained models. Zero123-XL inherits the architecture from Zero123 (Liu et al., 2023), but it is trained with the extended version of the Objaverse dataset, i.e., Objaverse-XL (Deitke et al., 2024), which contains over 1 million 3D objects. Zero123-XL is a one-to-one model, as it accepts only one conditioning image to generate a single output representing the same object from a given different angle. Conversely, EscherNet is trained on the standard version of Objaverse-1.0 and is a many-to-many generative model, as it can accept more than one conditioning input image and generates multiple outputs from the target viewpoints. Specifically, we use the EscherNet variant employing 6 DoF camera poses. Indeed, both reference and target camera poses are required by EscherNet, and Zero123-XL; in contrast, DIMVIS is able to work without such information, as our implemented method generates a set of 4 images with azimuth evenly spaced between 0° and 360°; nevertheless, the set of potentially achievable viewpoints can be enlarged during training, such that DIMVIS could generate more views at inference.

Metrics. For the comparison, we compute three key metrics: the peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM) (Wang et al., 2004) and the learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018). PSNR is a widely used metric based on the pixel-wise difference between two images; SSIM measures the similarity of two images by comparing luminance, contrast, and structure. SSIM aims to reflect better the human visual perception, which is also the goal of LPIPS. The latter, though, computes the difference between the features obtained from a layer of a pre-trained image Convolutional Neural Network, namely the VGG (Simonyan & Zisserman, 2014) in our implementation.

Evaluation dataset. We evaluate DIMVIS and the baseline models on the Google Scanned Objects (GSO) (Downs et al., 2022) dataset. Specifically, we use the same subset of 30 objects used in (Liu et al., 2024; Kong et al., 2024), by

Table 1. Numerical evaluation of novel view synthesis on GSO

Method	Target views	Ref. views	CFG scale	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Zero123-XL	3	1	3.0	16.13	0.772	0.181
EscherNet			3.0	14.85	0.755	0.209
DIMVIS			3.0	<u>16.32</u>	<u>0.78</u>	<u>0.171</u>
EscherNet*	3×5		3.0	**17.26**	**0.8**	**0.165**
EscherNet	2	2	3.0	19.57	0.828	0.132
DIMVIS			1.0	**21.07**	**0.854**	**0.113**
EscherNet*	2×7		3.0	<u>20.76</u>	<u>0.849</u>	<u>0.117</u>
EscherNet	1	3	3.0	20.20	0.821	0.131
DIMVIS			1.0	**23.96**	**0.886**	**0.088**
EscherNet*	1×15		3.0	<u>22.35</u>	<u>0.869</u>	<u>0.104</u>

rendering 16 images for each object following the procedure described for the training set. Notably, for the rendering, we use the same lighting conditions used in (Kong et al., 2024), although they differ from the lighting settings of our training set, which are the same as in (Liu et al., 2024). This poses an additional generalization challenge for DIMVIS.

For an extensive evaluation, we consider all possible subsets compatible with the fine-tuning dataset settings: to do so, we first build all four sets having four images with an azimuth offset by multiple of 90°, and, for each set, we finally consider all subsets with one, two and three conditioning reference views.

Results. Tab. 1 shows the considered metrics computed on the GSO dataset using the DDIM (Song et al., 2020) sampler with 50 steps, as well as the values of the classifier-free guidance (CFG) (Ho & Salimans, 2021) scale used for the generation. In particular, we compute all the metrics with CFG equal to 1 and 3, but we report only the best obtained results. Interestingly, we found that when having only one condition view, DIMVIS works better setting the CFG scale equal to 3, while with two and three conditioning images we obtain the best performance without using classifier-free guidance, i.e., setting the CFG scale equal to 1.

When considering only one reference view, DIMVIS works slightly better than Zero123-XL; furthermore, the latter is a one-to-one model, hence lacking flexibility, and the performance gap gets larger when DIMVIS is used with multiple reference views. Overall, DIMVIS also outperforms EscherNet. Interestingly, the latter does not show a clear improvement when increasing the number of reference views from 2 to 3. In this regard, EscherNet is empirically proven to work at its best with many target views, being them random or duplicates of the actual target one. For this reason, we evaluate EscherNet by repeating each target 5, 7, and 15 times when the actual number of targets is respectively 3, 2 and 1, and taking only the first output image. This workaround, referred to as "EscherNet*" in Tab. 1, leads to better performance at the expense of increased computational cost during inference, with EscherNet outperforming DIMVIS only when a single view is available. Nevertheless, even in this case, DIMVIS outperforms EscherNet when multiple views are available; we attribute the superior DIMVIS performance to its capability of better aggregating the information provided by the reference images.

Fig. 2 shows the images generated using the considered methods. As it is possible to notice, the images produced by Zero123-XL are coherent with the (single) reference view, but are not consistent with each other, as they are generated independently. EscherNet overcomes this drawback by implementing specifically designed attention mechanisms, while DIMVIS addresses this challenge by leveraging a joint diffusion model that exploits the self- and cross-attention mechanisms inherited from the pre-trained Stable Diffusion, without requiring any modification.

The images produced with DIMVIS generally exhibit higher visual quality and consistency; furthermore, DIMVIS seems to generate more diverse images throughout different runs (e.g., changing the set of conditioning images), while still maintaining reference-to-target and target-to-target consistency. Nevertheless, the shades produced by DIMVIS, even if plausible, do not always accurately reflect those present in the ground truth pictures: this is due to the different lighting conditions between our training set and the used test set, created following the same settings employed in EscherNet paper. However, this issue is alleviated when more views are observed.

Consistency and diversity. Importantly, when increasing the number of reference views, more information is available and the generative process is more constrained: this means that the inherent randomness of the generation decreases and the produced novel views get closer to the ground-truth images, which leads to improved performance in terms of the analyzed metrics. However, the low metrics values obtained with fewer reference views do not necessarily indicate poor generation quality. Consider for example the sofa pictures in Fig. 2 obtained with DIMVIS. When only one or two condition views are available, the diffusion model is less constrained and can be more imaginative, generating different images that are still plausible and realistic, and also consistent with each other and with the available views. However, the standard metrics used in NVS compute the similarity, either pixel-wise or perceptual, between the generated image and ground truth, thus penalizing the diversity in the produced images that are not close to the latter, but are perfectly coherent with the information provided by

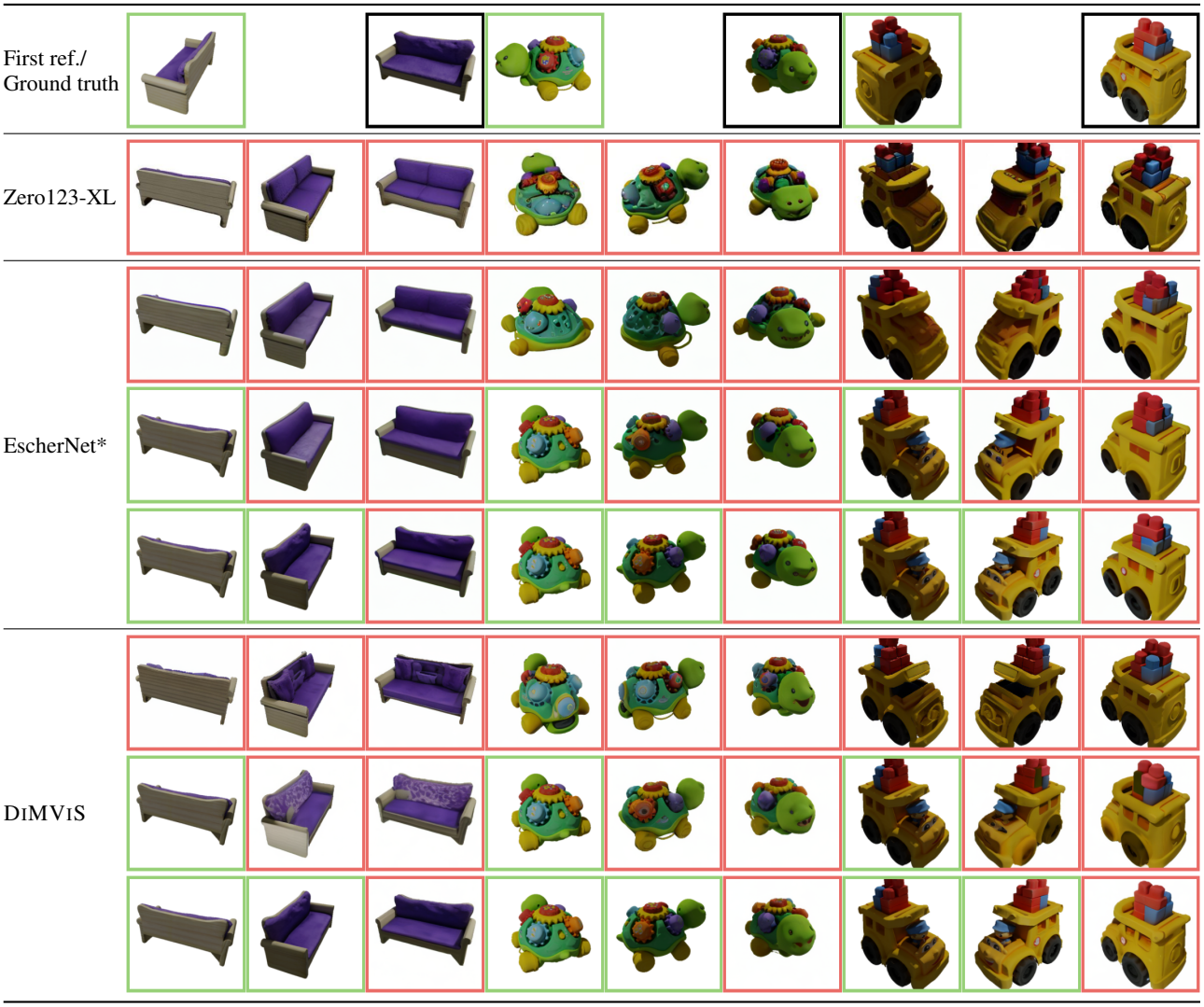

Figure 2. **Visual evaluation of novel view synthesis visualization on GSO datasets.** Ground truth, reference and generated images are respectively represented in black, green and red boxes. Note that, for each object, the first reference view is used for all generation processes, whereas we show only one ground truth image, as the other ones are already displayed as conditioning views. DIMVIS produces images that exhibit higher quality and diversity while ensuring reference-to-target and target-to-target consistency.

the reference view(s).

Nevertheless, diversity is one of the three main requirements in generative modeling (Xiao et al., 2022). Then, to take diversity into account, we argue that the currently used metrics should be integrated with new evaluation methods better suitable for novel view synthesis. One metric typically used to compare generative model performance is the Fréchet Inception Distance (FID) (Heusel et al., 2017), which evaluates both quality and diversity. However, the FID score has recently been shown to present some limitations (Jayasumana et al., 2023; Kynkäänniemi et al., 2023), and hence we do not include it in our analysis. Interestingly, (Watson et al., 2023) introduce what they call a "3D consistency scoring", a novel metric that respects the above-mentioned criteria. However, it is not compliant with our sparse-view setting, as it is based on a NeRF model, whose training requires many views. Alternatively, one viable way would be to resort to human evaluation, which we defer to our future work.

5. Conclusion

In this paper, we have presented DIMVIS, a multi-view 2D diffusion model able to accommodate a varying number of reference and target images. DIMVIS capitalizes on Stable Diffusion, inheriting its generalization capability, and on masked diffusion, which permits learning the latent conditional distribution by means of a joint diffusion model. Our

experimental evaluation has demonstrated the superiority of DIMVIS, compared to two state-of-the-art baselines.

Although DIMVIS is flexible in terms of the number of conditioning and generated views, it is trained on, and hence only produces, images with azimuth offset by multiples of 90°. Future research will investigate extending the current architecture to provide further flexibility, increasing the number of reference and target views and including the related camera poses. Future work will also include training images with different elevation, diverse lighting conditions and not limited to object-centric samples in order to achieve superior generalization ability.

Acknowledgements

This work was supported by the European Commission under Grant Agreement No. 101095363 (ADROIT6G project).

References

Anciukevicius, T., Manhardt, F., Tombari, F., and Henderson, P. Denoising diffusion via image-based rendering. *arXiv preprint arXiv:2402.03445*, 2024.

Bounoua, M., Franzese, G., and Michiardi, P. Multi-modal latent diffusion. *Entropy*, 26(4):320, 2024.

Chan, E. R., Nagano, K., Chan, M. A., Bergman, A. W., Park, J. J., Levy, A., Aittala, M., De Mello, S., Karras, T., and Wetzstein, G. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4217–4229, 2023.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S. Y., et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T. B., and Vanhoucke, V. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Höllein, L., Božič, A., Müller, N., Novotny, D., Tseng, H.-Y., Richardt, C., Zollhöfer, M., and Nießner, M. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023.

Jiang, H., Jiang, Z., Grauman, K., and Zhu, Y. Few-view object reconstruction with unknown categories and camera poses. *International Conference on 3D Vision (3DV)*, 2024a.

Jiang, H., Jiang, Z., Zhao, Y., and Huang, Q. LEAP: Liberate sparse-view 3d modeling from camera poses. In *The Twelfth International Conference on Learning Representations*, 2024b.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., and Davison, A. J. Eschernet: A generative model for scalable view synthesis. *arXiv preprint arXiv:2402.03908*, 2024.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4oXTQ6m_ws8.

Li, S., Zanjani, F. G., Yahia, H. B., Asano, Y. M., Gall, J., and Habibian, A. Valid: Variable-length input diffusion for novel view synthesis. *arXiv preprint arXiv:2312.08892*, 2023.

Lin, C.-H., Ma, W.-C., Torralba, A., and Lucey, S. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5741–5751, 2021.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., and Wang, W. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Müller, T., Evans, A., Schied, C., and Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15, 2022.

Pinkney, J. Stable diffusion image variations - a hugging face space by lambdalabs.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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.

Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., and Su, H. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021.

Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., and Liu, Z. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024a.

Tang, S., Chen, J., Wang, D., Tang, C., Zhang, F., Fan, Y., Chandra, V., Furukawa, Y., and Ranjan, R. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024b.

Truong, P., Rakotosaona, M.-J., Manhardt, F., and Tombari, F. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4190–4200, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Watson, D., Chan, W., Brualla, R. M., Ho, J., Tagliasacchi, A., and Norouzi, M. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.

Weng, H., Yang, T., Wang, J., Li, Y., Zhang, T., Chen, C., and Zhang, L. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023.

Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P. P., Verbin, D., Barron, J. T., Poole, B., et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21551–21561, 2024.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion GANs. In *ICLR*, 2022.

Ye, J., Wang, P., Li, K., Shi, Y., and Wang, H. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. 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.

Zhou, Z. and Tulsiani, S. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12588–12597, 2023.

A. Related work

In this section, we divide the current novel view synthesis approaches into two main categories, namely, Neural Radiance Fields (NeRFs) and 2D diffusion models.

A.1. NeRF models

NeRFs (Mildenhall et al., 2021) are powerful methods that pioneered a substantial advancement in 3D scene reconstruction and rendering, enabling NVS. While the first models required a per-scene optimization using dense input views and accurate camera poses, recent works have tried to overcome these limitations. BARF (Lin et al., 2021) trains a NeRF from noisy or unknown camera poses, but assumes dense input views. FORGE (Jiang et al., 2024a) is a NeRF variant that jointly predicts the input images' camera poses, which are assumed to be unknown, and performs the object reconstruction. FORGE is endowed with generalization capability and carries out a quick test-time optimization for further performance improvement; nevertheless, it requires at least 2 input images to work. SPARF (Truong et al., 2023) is a NeRF variant that performs effectively given only a few sparse input images, i.e., only a few views are available, with noisy camera poses, which are jointly optimized together with the radiance field; however, it requires per-scene optimization.

To deal with the challenges related to NeRF models, LEAP (Jiang et al., 2024b) proposes a generalizable NeRF that can model 3D scenes or objects given a sparse set of views, relying only on their relative camera poses. However, (Tang et al., 2024b) highlight that its proposed diffusion-based method achieves much better image quality than LEAP: according to the authors, the reason relies upon the strong image priors inherited from the pre-trained latent diffusion models. Furthermore, experimental results in (Kong et al., 2024) demonstrate the superior generation quality of a reference NeRF model (Müller et al., 2022) with respect to their diffusion-based method only when considering a dense set of conditioning views; nonetheless, the NeRF model requires scene-specific training and is outperformed in sparse-view settings. In such a scenario, we argue that novel view synthesis should be modeled as a conditional generation problem: when a part of an object is not visible, it can be only obtained by using a generative model, such as a diffusion model, and not NeRF approaches, which typically lack generative capability as they do not rely on generative modeling. By doing so, it is possible to take into account the stochasticity of the generation process and, hence, to produce different plausible images. Importantly, the generation randomness should decrease when increasing the number of reference views.

In this direction, (Zhou & Tulsiani, 2023) design a category-specific model for 3D reconstruction that leverages a diffusion model to optimize a NeRF, through a diffusion distillation mechanism. A similar approach is proposed in (Wu et al., 2024), where a NeRF is coupled with a diffusion model conditioned on the CLIP embeddings of the available views and a feature map containing their geometric information. However, such proposed models are only tested when more than one input view is available, whereas in real-use case scenarios only a single image may be observed.

A.2. 2D Diffusion models

The remarkable generative capability of 2D diffusion models (Rombach et al., 2022) has motivated the development of NVS methods built upon them. 3DiM (Watson et al., 2023) uses a diffusion model operating in the image space and allows more than one input image to condition the novel view synthesis; however, it generates only one output image, whose quality, counterintuitively, becomes worse when increasing the number of conditioning inputs.

Zero123 (Liu et al., 2023) paved the way to 2D diffusion models for NVS thanks to its strong generalization capability inherited by Stable Diffusion; however, it accepts only one conditioning image and generates a single output; thus, when creating images of the same object from multiple target viewpoints, this approach suffers from inconsistency as the images are generated independently. To overcome this issue, (Shi et al., 2023; Weng et al., 2023) extend Zero123 in order to generate multiple consistent views simultaneously, while (Liu et al., 2024) propose a diffusion model, still based on Zero123, that uses a 3D-aware feature attention mechanism to ensure consistency across the produced images. To achieve this goal, (Ye et al., 2023) use epipolar attention for geometric awareness and multi-view attention, which better combines information from multi-view images. The latter method generates different images simultaneously and can potentially be fed with a sparse set of multi-view images, but all experiments are performed with only one reference image. Nevertheless, using only a single conditioning input may be a limitation in real use-case applications: to address this drawback, (Li et al., 2023) modifie the Zero123 architecture to accommodate multiple views in input, but still generates a single image from a target viewpoint. (Tang et al., 2024b) propose a pose-free method to generate at the same time multiple images given a sparse set of views. To do so, the model employs a U-Net with several branches, i.e., copies, whose number is equal

to the number of total views, both conditioning and target. Notably, a global self-attention mechanism among the U-Net features across all branches is applied to enhance the consistency of the generated views. A similar methodology based on a multi-branch U-Net is envisioned in (Höllein et al., 2024). (Kong et al., 2024) employ transformers (Vaswani et al., 2017) that, by leveraging specifically designed camera positional encodings, aim to improve reference-to-target and target-to-target views consistency.

Another area of research focuses on 3D-aware diffusion models, combining both diffusion models and NeRF-like 3D representations. (Chan et al., 2023) present a diffusion model that integrates geometric priors by conditioning it on 3D features. The model takes a variable number of input views, while still generating a single output image. Also, the model is category-specific. (Anciukevicius et al., 2024) introduce a diffusion model based on a multi-view U-Net that produces a neural scene representation, instead of images, by leveraging cross-view attention to align the induced features, similar to the previously mentioned works. Notably, the model is trained from scratch on a limited set of classes of different datasets. This hinders the comparison with our method, which exhibits generalization capability thanks to the 2D prior inherited by the pre-trained Stable Diffusion model.

Novelty. Conversely to the works based on explicit conditioning by means of tailored attention modules, DIMVIS employs a joint diffusion model endowed with conditional generation ability by leveraging a straightforward yet effective masked diffusion mechanism that encourages correlation across the latent variables, which influence each other.