

VaLID: Variable-Length Input Diffusion for Novel View Synthesis

Shijie Li^{1,2 *} Farhad G. Zanjani² Haitam Ben Yahia²
Yuki M. Asano^{2,3} Juergen Gall¹ Amirhossein Habibian²

¹ University of Bonn

² Qualcomm AI Research[†]

³ University of Amsterdam

{lishijie,gall}@iai.uni-bonn.de y.m.asano@uva.nl {fzanjani,hyahia,ahabibia}@qti.qualcomm.com

Abstract

Novel View Synthesis (NVS), which tries to produce a realistic image at the target view given source view images and their corresponding poses, is a fundamental problem in 3D Vision. As this task is heavily under-constrained, some recent work, like Zero123 [18], tries to solve this problem with generative modeling, specifically using pre-trained diffusion models. Although this strategy generalizes well to new scenes, compared to neural radiance field-based methods, it offers low levels of flexibility. For example, it can only accept a single-view image as input, despite realistic applications often offering multiple input images. This is because the source-view images and corresponding poses are processed separately and injected into the model at different stages. Thus it is not trivial to generalize the model into multi-view source images, once they are available. To solve this issue, we try to process each pose image pair separately and then fuse them as a unified visual representation which will be injected into the model to guide image synthesis at the target-views. However, inconsistency and computation costs increase as the number of input source-view images increases. To solve these issues, the Multi-view Cross Former module is proposed which maps variable-length input data to fix-size output data. A two-stage training strategy is introduced to further improve the efficiency during training time. Qualitative and quantitative evaluation over multiple datasets demonstrates the effectiveness of the proposed method against previous approaches. The code will be released according to the acceptance.

1. Introduction

Zero-shot novel view synthesis (NVS) has attracted more and more attention recently as the development of AR/VR applications [14] and content creation [25, 47] rises. NVS

*Work completed during internship at Qualcomm Technologies, Inc.

[†]Qualcomm AI Research is initiative of Qualcomm Technologies, Inc.

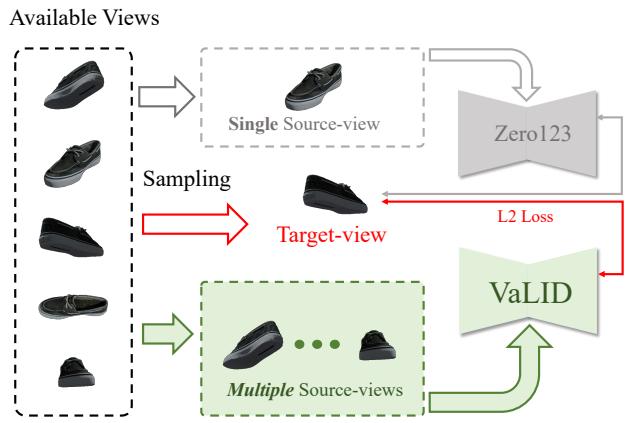


Figure 1. Compared to the previous method (Zero123[18]), which can only receive a single image as input even when multiple images are available, the proposed architecture (VaLID) can accept variable input views in both training and inference. Thus it can learn a robust representation during training and can utilize more information in the inference time.

tries to synthesize high-quality and visually consistent views given source view images and relative poses. Yet, because of the general under-constrainedness of reconstruction and due to occlusion, it is a highly challenging task that usually involves solving this large ambiguity.

Current solutions can be roughly partitioned into inference-time optimization-based methods and optimization-free methods. The former [22, 24] can produce high-quality outcomes by conducting time-consuming iterative optimization for each object and thus cannot generalize to other objects directly. As a comparison, the latter [18–21, 25, 29, 41, 41, 44, 46] models the whole task by a neural network. Although the generation is a single forward in the inference time which is highly efficient, the quality of generated images is not desirable. Compared to the above methods, Zero123[18] shows a promising alternative solution. By utilizing the powerful diffusion model, it can produce high-quality images with good generalization ability and efficiency. However, it has been constrained to a single source-view image as input whereas, in the case of

the availability of multi-view source images, it cannot be utilized in the process of NVS. The flexibility of a model to a variable number of source-view images is more desirable in real applications where more source images reduce ambiguity through multi-view fusion. Figure 3 (a) shows an overview of the Zero123[18] architecture. The model adopts an appearance-pose disentanglement conditioning strategy. The appearance information is injected at the input of U-Net whereas the corresponding relative pose information is injected in the attention modules in the U-Net. Due to separate conditioning of the diffusion process on appearance and pose embeddings, it is non-trivial to generalize the model to multi-view source images, once they are available.

To resolve this restriction and enable the model with the ability to receive variable size source-view images as input, an appearance-pose-entanglement conditioning strategy is proposed which is shown in Figure 3 (b). Based on this conditioning strategy, we propose a novel architecture named **Variable-Length Input Diffusion** (VaLID). Compared to prior works including Zero123[18], VaLID can receive variable size source-view images as input to perform NVS, during both training and inference time. This flexibility allows to handle single-view and multi-view input at the same time and to produce even higher fidelity results when more views are available. The proposed VaLID first transfers the source-view image(s) into rich visual embeddings through a Vision Transformer (ViT) encoder, pre-trained with the Masked Auto-Encoder (MAE) task [9]. The ViT encoder produces multiple spatial output tokens. Similar to Zero123[18], the camera pose(s) are transferred into the pose embedding(s) through an MLP network. The image tokens are concatenated with their corresponding camera pose features to form the input for conditioning the diffusion network.

However, when multiple source-view images are available, there will be an inconsistency among tokens from different images. What is worse, the required computation resource will increase a lot due to the increased number of image tokens from multiple images. To make tokens from different views work harmoniously while efficient, we propose a new “Multi-view Cross Former” module. In this module, all the tokens will be fused and transferred to a fixed number of tokens independent of the number of input source-view images. Finally, these transferred tokens will be injected into the attention modules in the diffusion model to guide the target-view image synthesis.

Although the proposed architecture has the ability to receive multiple source-view images as input, this also requires more training resources. To alleviate this issue, an efficient two-stage training strategy is introduced where the first stage mainly focuses on learning NVS from a single source-view image whereas the second stage mainly solves

inconsistency when multiple source-view images are available by only finetuning relevant modules. A token sampling strategy is applied at stage 2 to further constrain training cost.

In summary, our contributions can be summarised as:

- We introduce a diffusion-based NVS model to address variable-size multi-view image fusion, both in training and inference times. The proposed appearance-pose-entanglement conditioning strategy can outperform previous methods quantitatively and qualitatively even when only a single source-view image is used.
- We introduce Multi-view Cross Former to transfer the variable-size input tokens into a fixed-size representation by learning a set of learnable tokens. This improves the consistency and efficiency of conditioning to generate novel views while being agnostic to the number of input images.
- With the proposed two-stage training strategy, the training efficiency is improved. What’s more, the performance of single-view image-conditioned NVS is also improved by involving multiple views in the training.

2. Related work

Novel View Synthesis is a fast-moving research area with many concurrent works. We broadly categorize the existing and concurrent works into single and multi-view models and highlight the key differences to our work.

2.1. Single image novel-view synthesis

Several NVS methods based on single-view input image have been presented on NeRF-based [22] approaches to act on a single view [21, 32, 42, 43, 48] to generate novel views. While these methods produce impressive results, NeRFs famously suffer from requiring test time optimization and the single-view variants typically do not work as well as their multi-view counterparts [48]. To counter the lack of information from a single view, some include additional geometric priors as supervision, input or inductive bias, such as depth maps [8, 13, 36, 42, 43] point clouds [23] or epipolar attention [35]. Our work, in contrast, does not use any geometric representation except for the relative pose. Although these geometric representations work well to aid in reconstructing the observed angle, it does not circumvent that some angles are simply occluded or unobserved. Therefore a large body of work aims at filling these gaps with generative models. Specifically, diffusion models [11, 31] have recently become the primary method to fill in gaps in the observed views as in RealFusion [21], DreamFusion [25] and Zero123 [18]. In particular, Zero-1-to-3 [18] has shown that this is possible with a single view and a relative camera pose. They use CLIP [26] to extract appearance features, combined with a relative pose to generate novel views. Our work is closest to their setup, but we reduce high variance in

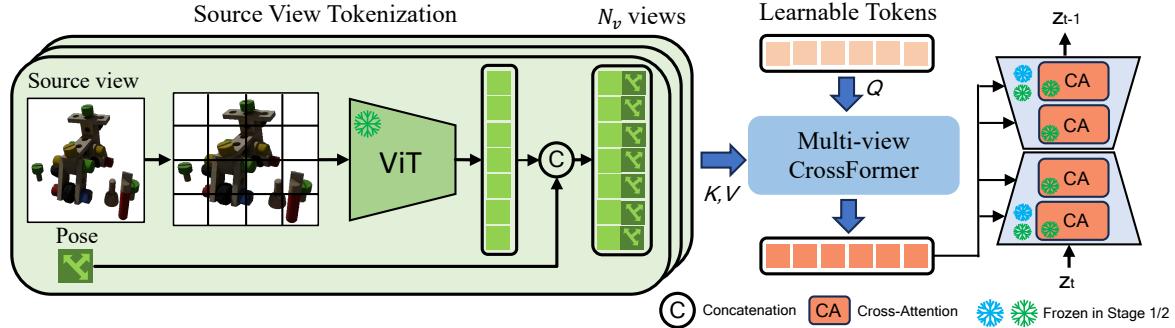


Figure 2. Overview of the proposed architecture (VaLID).

unobserved regions by extending it to the multi-view setting for training and allowing for single or multi-view inference. This is the *key difference* between single-view methods and ours: we investigate and quantify the effect of increasing the number of views, either during training or inference, without any geometric priors. Our work is orthogonal to many of the aforementioned methods and could be further improved as a combination.

2.2. Multi-image novel view synthesis

In this section, we look in particular at multi-view works in sparse view regimes. Similar to the single-view setup, there are several NeRF-based methods [10, 12, 16, 32, 38, 45], which try to mitigate the number of views needed for novel view synthesis. Several papers focus on geometry-free NVS such as ENR [6] that lifts the extracted features of a pair input images into 3D space and minimizes the cross-rendering loss. The introduced 3D lifting is computationally expensive and degrades the quality of rendered images. LFNs [30] samples a single point per ray cast and SRT [28] uses transformers to align multiple frames implicitly. Our work likewise is a geometry-free approach, but we additionally use generative modeling to synthesize novel views. EG3D [2] and 3DiM [40], similarly use generative modeling to align 3D features, using a GAN [7] or diffusion model respectively. 3DiM [40] is closest to our work with its diffusion model, but they only use multiple views during training time and generate poses in an auto-regressive manner at inference that suffers from accumulation of errors in generating and feeding in the additional views and being much slower than our work that consumes all the inputs in a single generation process.

MVDiffusion [33] uses text prompts, depth maps, pixel correspondences, multi-view, and diffusion models on scene reconstruction. Magic3D [15] trains a coarse NeRF with a diffusion model on top for high-resolution NVS based on a text prompt. SyncDreamer [19] does NVS from a single view, by lifting several noisy target images into a frustum and conditioning on this in a pre-trained Zero123. Several works try to enforce consistency across Zero-1-to-3 in dif-

ferent ways [17, 41, 44, 46] *i.e.* by projecting features to 3D [44]. The closest to us is Consistent123 [41], which also uses cross-attention among different views, but the main difference is that we look at single-view output rather than simultaneous multi-view output. MVDREAM [29] uses text prompts and diffusion models. Wonder3D [20] estimates normal maps and textures and combines them using optimization of a neural implicit signed distance field (SDF) to amalgamate all 2D generated data.

Unlike the previously mentioned methods, the proposed VaLID does not require optimization during inference and can accommodate a varying number of source views, making it suitable for practical applications.

3. VaLID Novel View Synthesis

We aim for generating a realistic image y at the target-view, given any number of source-view images $x = \{x_i\}_{i=1}^{N_v}$, and corresponding camera poses relative to the target-view $\pi = \{\pi_i\}_{i=1}^{N_v}$, where $N_v \geq 1$ is a variable number of input source-view images. VaLID involves two steps: *i)* Extracting features from the source-view images as described in Sec. 3.1. *ii)* Aggregating features from source-view images to generate the target-view as discussed in Sec. 3.2. A high-level overview of our pipeline is illustrated in Figure 2.

3.1. Source view tokenization

Limitations of existing conditionings The prior work Zero123[18] is made of two conditioning mechanisms as illustrated in Fig 3 (a): *i) U-Net conditioning*, where a source-view image x_i will go through a frozen Auto-Encoder. The output latent map f_i^{AE} is then concatenated with a noisy latent feature map z_t from the previous diffusion timestep $t + 1$ as input to the U-Net. *ii) Attention conditioning*, where the source-view image x_i will go through a frozen CLIP image encoder. The output CLIP embedding f_i^{CLIP} is concatenated with the relative pose π_i and then fed into the attention modules in the U-Net. Zero123[18] can thus

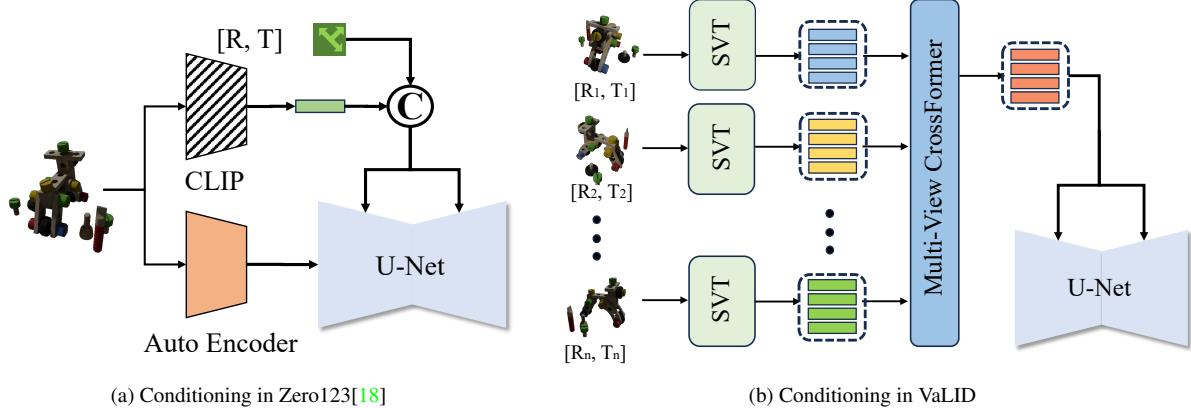


Figure 3. View conditioning in prior work (a) vs. in VaLID (b). By learning a joint appearance pose representation, VaLID seamlessly handles variable length input, which is not possible in a disentangled representation of appearance and pose as in Zero123[18].

be formulated as:

$$z_{t-1} = \epsilon_\theta (f_i^{AE} \oplus z_t, f_i^{\text{CLIP}} \oplus \pi_i, t), \quad (1)$$

where ϵ_θ denotes the U-Net, t the timestep in the diffusion process and symbol \oplus denotes the concatenation operator. We found that the Attention conditioning mainly ignores the CLIP features and only propagates pose information as shown in Figure 4. More examples are available in the supplemental materials. We hypothesize that the CLIP encoding, a high-level single token image embedding, is too coarse to retain image details. Thus Eq. (1) reverts to $z_{t-1} \approx \epsilon_\theta (f_i^{AE} \oplus z_t, \pi_i, t)$, where we see that the extracted features of input source-view image f_i^{AE} and relative pose π_i are injected into the U-Net at different stages. This works well when only a single source-view image is available as in Zero123[18]. However, when multiple source-view images are available, this disentangled appearance-pose conditioning strategy would require the model to align the poses with the source-view features at a later stage in the network.

Learning joint appearance and pose conditioning. We propose to entangle appearance and pose features together. Since only U-Net conditioning encodes appearance information, one straightforward solution is that we feed both source-view images and corresponding poses to the U-Net input. The problem here is that the number of input views would be inherently tied to the first convolutional layer in the U-Net, making it inflexible. Furthermore, when multiple source-view images are available, feeding them directly into the U-Net will produce inconsistent results, which are shown in Figure 5. Thus we utilize the flexibility of the attention mechanism and opt for feeding both source-view images and corresponding poses to the attention modules in U-Net. In this way, we need to replace the CLIP image encoder by another architecture to better encode source-view images. We achieve this with the **SVT** Module which can extract fine spatial information from source-view images.

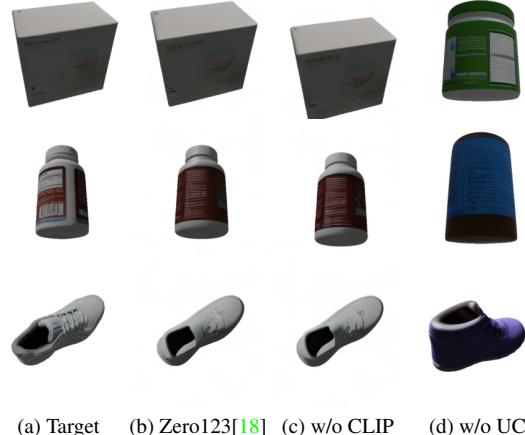


Figure 4. Conditioning limitations in Zero123 [18]: Conditioning is dominated by the U-Net input (UC) whereas CLIP embedding is ignored (CLIP).

Our final architecture can be formulated as:

$$z_{t-1} = \epsilon_\theta (z_t, \text{SVT}(x_1, \pi_1, \dots, x_{N_v}, \pi_{N_v}), t). \quad (2)$$

Then the network is trained by optimizing a simplified variational lower-bound

$$\mathcal{L} = \mathbb{E}[||\epsilon_t - z_{t-1}||^2]. \quad (3)$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$.

Source View Tokenization Module In the Source View Tokenization (**SVT**) Module, each input source-view pose image (x_i, π_i) is processed separately and converted into pose-image tokens encoding both appearance and geometry information. We extract features for each input source-view image x_i with a ViT encoder (Φ), pre-trained with the Masked Auto-Encoder (MAE) task [9]. The input image will be converted into small image patches $\chi = \{x_j\}_{j=1}^{N_p}$

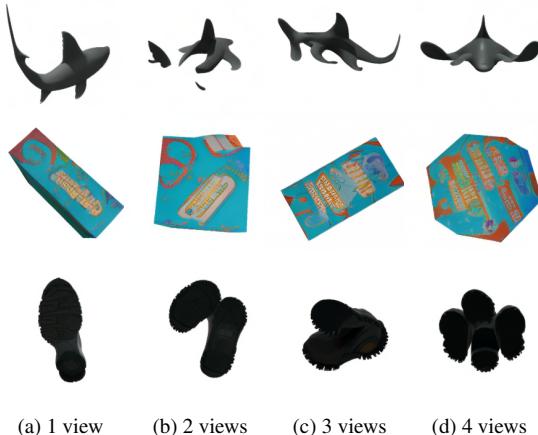


Figure 5. We can observe without the stage 2 training, there will exist heavily inconsistency in the outcomes when we feed multiple source-view images to the proposed architecture.

where N_p is the number of image patches. Positional embeddings are added to the images as in [4] for spatial awareness. These image patches are then fed into the ViT encoder and converted into image tokens:

$$\mathbf{M} = \Phi(\mathcal{X}), \quad \mathbf{M} = \{m_j\}_{j=1}^{N_p} \quad (4)$$

At this stage, each image token m_j only contains appearance and 2D spatial information whereas camera pose information is missing. To inject such information, the 3D poses are appended to the end of each token:

$$\tilde{\mathbf{M}} = \{\tilde{m}_j\}_{j=1}^{N_p}, \text{ where } \tilde{m}_j = m_j \oplus \pi^s. \quad (5)$$

where π^s is the corresponding pose to each image token m_j . For now, each pose-image token \tilde{m}_j is aware of both appearance information and spatial information (2D spatial and 3D pose). Compared to the CLIP image encoder, the Source View Tokenization can extract rich spatial information and thus better describe image details. The proposed architecture is shown in Figure 2. Later, the extracted tokens from multiple source views can be fused together in our Multi-view Token Cross Former module described in the next Section.

3.2. Multi View Fusion

3.2.1 Multi-view Cross Former

In the Source View Tokenization Module, we extract pose-image tokens $\tilde{\mathbf{M}}_i$ for each input source-view image. As mentioned earlier, inconsistent results are produced when providing multi-views directly to the U-Net (Figure 5). Alternatively feeding all source view image tokens will increase the computation proportionally to the number of images, and causes out-of-memory issues. Inspired by text-to-image diffusion models [27], where a fixed number of

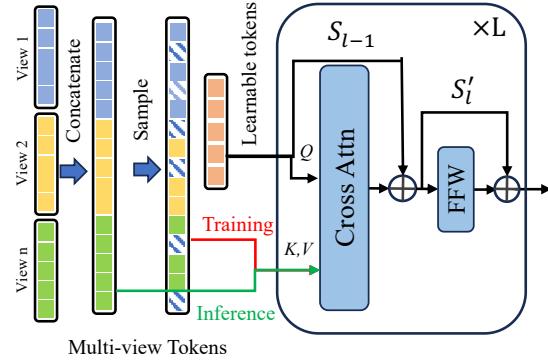


Figure 6. Multi-view Cross Former. During **training** time (stage 2), we randomly sample image tokens from all available source-view images. With this strategy, we can make the Multi-view Cross Former receives multi-view information while restrict training cost. In the **inference** time, we either feed all the tokens to include all information or sample during training.

tokens have been shown to perform favorably, we introduce a fixed number (equal to 64) of learnable seed tokens S_0 , initialized as $S_0 \sim \mathcal{N}(0, 1)$ at the beginning of training. These are used as queries in the Multi-view Cross Former block, visualized in Figure 6. Since these initial tokens S_0 are shared across training examples, they learn to be query token biases, which aim to extract relevant information from pose-image tokens $\tilde{\mathbf{M}}_i$. The Multi-view Cross Former block, in particular, is first computed with an attention operation followed by a residual:

$$\mathbf{S}'_l = \text{Attn}_l(Q, K, V) + \mathbf{S}_{l-1}. \quad (6)$$

Then the output tokens will pass through a feed-forward layer FFW_l which consists of layer normalization, a linear layer, GeLU activation, another linear layer sequentially, followed by a final residual connection:

$$\mathbf{S}_l = \text{FFW}_l(\mathbf{S}'_l) + \mathbf{S}'_l. \quad (7)$$

The query, key, and value in the attention block are calculated with $Q = \mathbf{S}_{l-1}$ and $K = V = \tilde{\mathbf{M}}_1 \oplus \dots \oplus \tilde{\mathbf{M}}_{N_v}$. The output tokens \mathbf{S}_l will be used as new target-view seeds for the following layer and the whole block will be repeated L times. Finally, the target-view seeds \mathbf{S}_L will be fed into attention modules at each timestep of the diffusion process to produce realistic images at the target view. With this design, the final target-view seeds \mathbf{S}_L learn to warp the source view tokens into the target-view tokens that are being used for conditioning the diffusion process. This also enables our model to always feed a fixed number of query tokens into the U-Net attention module which improves efficiency a lot.

3.2.2 Efficient Training and Inference

As feeding multiple source-view images during training is still expensive, we split the whole training procedure into two steps. In the first stage, the proposed architecture is trained to produce target-view images given a single source-view image. After this stage, our model already has the ability to produce realistic target-view images. As multi-view information aggregation only takes effect in the Multi-view Cross Former, we freeze all modules apart from the Multi-view Cross Former in the second stage of the training. With this strategy in the second stage of training, all the training resources will be used to make multi-view information work harmoniously to produce consistent outcomes. Additionally, we propose a sampling strategy during the second stage training where we feed a variable number of views to the model and randomly sample pose-image tokens from them. By reducing the number of pose-image tokens fed into the Multi-view Cross Former, the training cost is further reduced.

Stage 1: Single-view Optimization We build the proposed model on top of a pretrained stable-diffusion model. Thus we can assume that the U-Net has already the ability to produce realistic images. With this assumption, we only optimize the attention modules in the U-Net. Thus in this stage, the Source View Tokenization Module, Multi-view Cross Former, and attention modules in the U-Net are optimized with a single source-view image as input.

Stage 2: Variable-view Optimization After stage 1 training, we only finetune the Multi-view Cross Former where multi-view information fusion takes place, highlighted in Figure 2. In each training iteration, variable input views are fed into the model while only partial pose-image tokens are fed into the Multi-view Cross Former for optimization. With this design, we not only improve the training efficiency due to reducing the number of used pose-image tokens but also improve the robustness of the proposed architecture when only partial information is available. At inference time, we can choose to feed all pose-image tokens to the Multi-view Cross Former to provide more information to the model. We can alternatively sample fewer tokens for efficiency purposes though we find this also hurts performance (shown in Sec. 4.3).

4. Experiments

4.1. Experimental Setting

Dataset We follow Zero123’s [18] experimental setting and fine-tune our diffusion model on the Objaverse dataset [3] which contains more than 800K high-quality 3D object CAD models. For a fair comparison, we use the processed rendering data provided by Zero123[18] which are 12 random views for each object. For evaluation, we mainly focus on the performance of out-of-distribution data. Thus we

conducted the evaluation on the Google Scanned Objects dataset [5] (GSO), which contains high-quality scanned household items, and RTMV [34] dataset, which provides more complex scenes each with around 20 random objects. In both training and evaluation, all the images are resized to 256×256 resolution with a white background. To be compliant with the visual transformer architecture in the Source View Tokenization Module, we conduct the center crop thus the final input size will be 224×224 .

Evaluation Setting On the Google Scanned Objects dataset, 24 views are randomly sampled in 3D as the target view. To demonstrate the performance of our method with variable-length input source-view images, multiple source-view images are rendered following some predefined rules. A reference view (view 0) is randomly picked up first with polar angle 60° for each object. Then 3 other views (view 1,2,3) will be clockwisely selected by only varying azimuth (azimuth angle interval as 90°). In this setting, the available input information is controlled increasing gradually proportional to the number of input source-view images. Because the RTMV dataset already provides rendered images, we conducted the evaluation following the Zero123 [18] experimental setting for fair comparison. To quantitatively evaluate the quality of generated images, we use Peak signal-to-noise ratio (PSNR), Structural Similarity Index (SSIM) [39], and Learned Perceptual Image Patch Similarity (LPIPS) [49] to measure similarity between rendered images and ground truth images.

Baselines We mainly compare our method to the previous state-of-the-art method, Zero123[18] which can produce realistic images at the target view iteratively with a diffusion model. Besides, an image-conditioned diffusion model Image Variations (IV)[1] is also selected as the baseline method which can produce semantic consistent images based on input images. This is achieved by finetuning a Stable Diffusion model to be conditioned on images instead of text prompts. Although Neural Radiance Field [22] (NeRF) is widely adopted for novel view synthesis, it only works well when a large number of images are available. Hence, to further evaluate the effectiveness of our proposed method, we compare it to DietNeRF [12], a technique that regularizes NeRF using a CLIP image-to-image consistency loss. DietNeRF has demonstrated its capability to produce high-quality results even when only a limited number of images are available. Additionally, zero-shot 3D content creation, which optimizes a Nerf with a pretrained diffusion model becomes popular. We use SJC[37] which combines an image-conditioned diffusion model and NeRF as another baseline method to represent these methods.

4.2. Comparison to the state-of-the-art methods

We first compare our method to the baseline methods quantitatively on Google Scanned Objects and RTMV datasets.

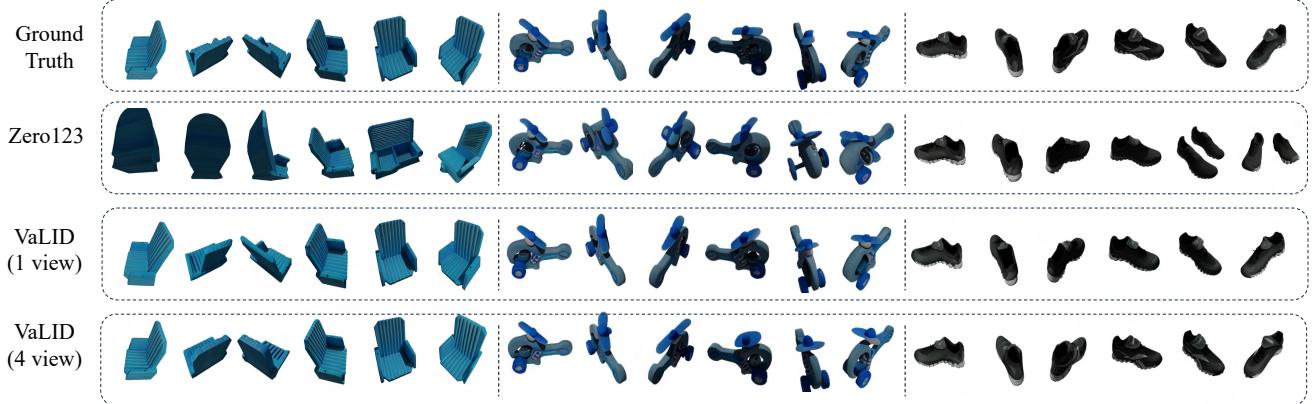


Figure 7. Qualitative results on GSO dataset. We can observe our method can produce high-quality outcomes at the target view compared to the previous state-of-the-art methods. Furthermore, the generated images from our method also maintain consistency across different views. This is demonstrated by generated images along a predefined camera trajectory. Especially when multiple source-view images are available, the quality of generated images improves a lot. More examples can be found in the supplemental materials.

	DietNeRF [12]	IV [1]	SJC-I [37]	Zero123 [18]	VaLiD			
View num	1	1	1	1	1	2	3	4
PSNR \uparrow	8.933	5.914	6.573	19.000	20.034	20.405	21.053	21.305
SSIM \uparrow	0.645	0.540	0.552	0.865	0.881	0.884	0.891	0.895
LPIPS \downarrow	0.412	0.545	0.484	0.115	0.091	0.085	0.073	0.069

Table 1. NVS results on Google Scanned Objects dataset. The proposed method outperforms the previous state-of-the-art method by a significant margin on all metrics. Furthermore, as the number of input source-view images increased, the performance gradually improved.

	DietNeRF [12]	IV [1]	SJC-I [37]	Zero123 [18]	VaLiD
Input views	1	1	1	1	1
PSNR \uparrow	7.130	6.561	7.953	8.893	9.024
SSIM \uparrow	0.406	0.442	0.456	0.515	0.519
LPIPS \downarrow	0.507	0.564	0.545	0.432	0.432

Table 2. NVS results on RTMV dataset. Compared to the Google Scanned Objects dataset, this dataset contains more challenging scenes. We can observe our method still can outperform other methods.

The results are shown in Tab. 1 and Tab. 2. For the GSO dataset, the proposed method can achieve a 5% improvement on PSNR when only a single source-view image is used. The improvement on SSIM and LPIPS are also obvious. By increasing the input source-view images, the performances on all metrics increase gradually. This demonstrates the capability of the proposed method to utilize information from multiple source-view images when available. Furthermore, the proposed architecture has the flexibility to receive variable-length inputs and thus is more suitable for realistic applications or when the model receives input images progressively. As for the RTMV dataset, we follow the Zero123[18] experiment setting because this dataset already provides rendered images. For each scene, the first view is used as the source view and the following views are used as the target view. Compared to the GSO dataset, this dataset contains more objects for each scene thus more challenging. Even in this situation, our method can still outperform other baseline methods.

Finally, we show some qualitative results in Figure 7. Apart from the quality of each individually generated image, the consistency among different target views is also important which is usually missing in previous work. To demonstrate the consistency of the generated images among different target views, we select a predefined camera trajectory to demonstrate the qualitative results. From these qualitative results, we can observe the proposed method not only can produce more realistic images compared to previous methods but also can maintain consistency among different target views even no explicit geometry constraints are applied. Especially when multiple source-view images are available.

4.3. Ablation Study

In the ablation study, we show the influence of sampling strategy in both inference and training times. First, we show the effect of the sampling strategy, involved in stage 2 of training in Figure 8 (a)-(c). It can be observed that in stage 2 of training, tokens sampling from two source views actually achieve quite close performance, compared to feeding all the tokens to Multi-view Cross Former. Moreover, the experiments show that providing more source-view images improves the quality of the generated images. However, increasing up to six views doesn't show improvement in comparison to four views. This may be because a fixed number of learnable tokens is applied in Multi-view Cross Former.

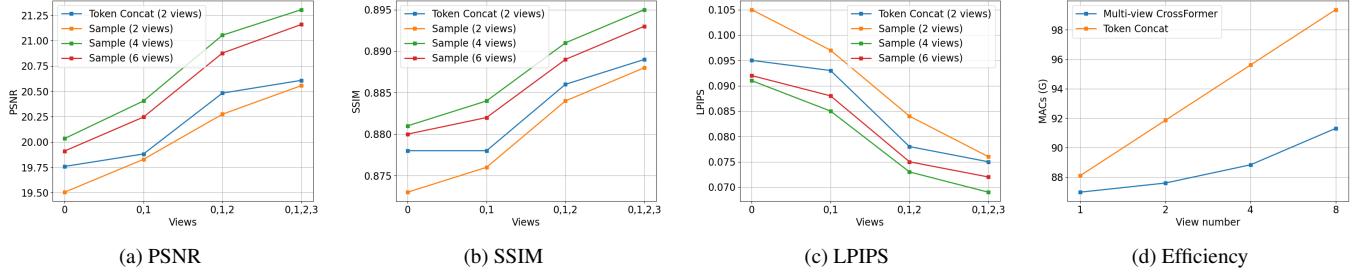


Figure 8. Token Sampling setting in the stage 2 training (a)-(c). The efficiency analysis of the proposed architecture is shown in (d).

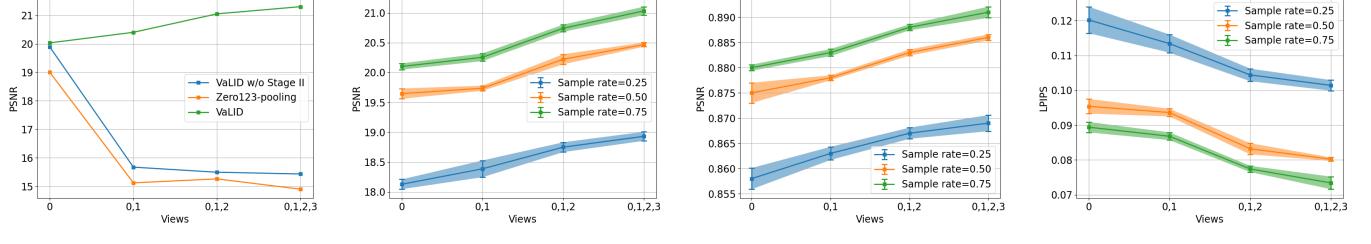


Figure 9. We show the necessity of stage 2 training in (a). The effect of token sampling at Inference time is shown in (b)-(d). **Green line** denotes sampling 75% tokens before Multi-view Cross Former. **Red line** and **Blue line** represent sample ratio equals to 50% and 25% respectively. We can observe as the available information increase, the uncertainty decrease which are measured by 5 runs.

When the number of input tokens becomes too large, it will face a bottleneck to collect all useful information.

The computational burden of preserving (without sampling) all pose-image tokens for different numbers of source-view images is shown in Figure 8 (d). It can be observed that the introduced token sampling strategy shows a significant reduction in the required Multiply-Accumulate Operations (MACs) for U-Net and Multi-view Cross Former, especially when the number of source views increased. Thanks to Multi-view Cross Former, ignoring a considerable number of pose-image tokens via sampling does not show a performance loss while it obtains high efficiency in computation.

The importance of the stage 2 training is shown in Figure 9 (a). Without the stage 2 training, when multiple source-view images are available, the performance of our method decreases a lot. This is mainly because of inconsistency in multi-view tokens which is shown in Figure 5 previously. As for Zero123[18], we apply a pooling operation to enable it with the ability to receive multiple source-view images. Unfortunately, the performance also decreases significantly when multiple source-view images are fed. Another interesting finding is that stage 2 training can also improve the performance of single source-view image-conditioned NVS. This is because receiving multi-view information during training enables the model to learn a more comprehensive representation even if only a single source-view image is available in inference time. This ablation shows the impact of stage 2 training where the Multi-view Cross Former adapts to perform multi-view token fusion when its input contains pose-image tokens belonging to distinct views.

In inference time, we try to validate the robustness of the

proposed method. This is achieved by feeding partial tokens into Multi-view Cross Former in the inference time similar to above. The results are shown in Figure 9 (b)-(d). The number is reported on 5 runs. It can be observed that when the number of available source-view images increases or a higher number of tokens is used, the performance improves gradually. Furthermore, the uncertainty is reduced gradually as the available information increases. These results also show the robustness of the proposed method as even though limited information is available, the performance is still at a high level.

5. Conclusion

This paper presents a novel diffusion-based framework for novel view image synthesis, named as Variable-Length Input Diffusion model (VaLID). Compared to previous diffusion-based methods, which can only receive a single image as input, VaLID can accept variable number of views as input in both training and inference time. This flexibility makes the proposed model more suitable for realistic applications where multiple views but with a variable number are usually available or progressively presents to model. The multi-view fusion is achieved by an appearance-pose entanglement conditioning strategy. To handle the information inconsistency among different views, a Multi-view Cross Former module has been introduced which is highly efficient, both in terms of reducing the number of tokens, which conditioned the diffusion process, to have a fix-length set and in terms of training strategy. The proposed method outperforms previous methods both quantitatively and qualitatively which demonstrates its effectiveness.

References

- [1] Stable diffusion image variations - a hugging face space by lambdalabs. [6](#) [7](#)
- [2] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. [3](#)
- [3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. [6](#)
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [5](#)
- [5] 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 *International Conference on Robotics and Automation*, pages 2553–2560. IEEE, 2022. [6](#)
- [6] Emilien Dupont, Bautista Miguel Angel, Alex Colburn, Aditya Sankar, Carlos Guestrin, Josh Susskind, and Qi Shan. Equivariant neural rendering. In *International Conference on Machine Learning*, 2020. [3](#)
- [7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. [3](#)
- [8] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multi-plane images. In *ACM SIGGRAPH*, 2022. [2](#)
- [9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. [2](#) [4](#)
- [10] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotný. Unsupervised learning of 3d object categories from videos in the wild. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4698–4707, 2021. [3](#)
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [2](#)
- [12] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *IEEE International Conference on Computer Vision*, pages 5885–5894, 2021. [3](#) [6](#) [7](#)
- [13] Yash Kant, Aliaksandr Siarohin, Michael Vasilkovsky, Riza Alp Guler, Jian Ren, Sergey Tulyakov, and Igor Gilitschenski. invs : Repurposing diffusion inpainters for novel view synthesis. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. [2](#)
- [14] Johannes Kopf, Kevin Matzen, Suhib Alsian, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM Transactions on Graphics (TOG)*, 39(4):76–1, 2020. [1](#)
- [15] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [3](#)
- [16] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. [3](#)
- [17] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023. [3](#)
- [18] 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. *arXiv preprint arXiv:2303.11328*, 2023. [1](#) [2](#) [3](#) [4](#) [6](#) [7](#) [8](#) [11](#) [12](#) [13](#)
- [19] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. [3](#)
- [20] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv:2310.15008*, 2023. [3](#)
- [21] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. [1](#) [2](#)
- [22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. [1](#), [2](#) [6](#)
- [23] 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. [2](#)
- [24] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. [1](#)
- [25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023. [1](#), [2](#)
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya

- Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [2](#)
- [27] 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. [5](#)
- [28] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022. [3](#)
- [29] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. [1, 3](#)
- [30] Vincent Sitzmann, Semon Rezhnikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Neural Information Processing Systems*, 2021. [3](#)
- [31] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, JMLR Proceedings, pages 2256–2265, 2015. [2](#)
- [32] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3d generative models from 2d data. *International Conference on Computer Vision*, 2023. [2, 3](#)
- [33] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Neural Information Processing Systems*, 2023. [3](#)
- [34] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Thomas Müller, Charles Loop, Nathan Morrical, Koki Nagano, et al. Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *arXiv preprint arXiv:2205.07058*, 2022. [6](#)
- [35] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2023. [2](#)
- [36] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 548–557, 2020. [2](#)
- [37] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. [6, 7](#)
- [38] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. [3](#)
- [39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. [6](#)
- [40] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *International Conference on Learning Representations*, 2023. [3](#)
- [41] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, C. L. Philip Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. [1, 3](#)
- [42] Dejia Xu, Yifan Jiang 0001, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360° views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023*, pages 4479–4489. IEEE, 2023. [2](#)
- [43] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 736–753. Springer, 2022. [2](#)
- [44] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. *arXiv*, 2023. [1, 3](#)
- [45] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. [3](#)
- [46] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023. [1, 3](#)
- [47] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6496–6503. IEEE, 2022. [1](#)
- [48] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. [2](#)
- [49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. [6](#)

VaLID: Variable-Length Input Diffusion for Novel View Synthesis

Supplementary Material

6. Conditioning Strategy

As mentioned in the main paper, Zero123[18] is made of two conditioning mechanisms; (1) *U-Net conditioning*, where a source-view image x_i will go through a frozen Auto-Encoder and the output latent map f_i^{AE} is then concatenated with a noisy latent feature map z_t from the previous diffusion timestep $t + 1$ as input to the U-Net. (2) *Attention conditioning*, where the source-view image x_i will go through a frozen CLIP image encoder. The output CLIP embedding f_i^{CLIP} is concatenated with the relative pose π_i and then fed into the attention modules in the U-Net.

Figure 10 demonstrates the impacts of these two conditioning strategies on the generated images. To drop out the CLIP embedding, we mask them out with a 0-tensor of the same size. It can be observed that in the case of following the default setting, where both U-Net conditioning and Attention conditioning exist, the model can produce plausible outcomes. As a comparison, removing U-Net conditioning (w/o concat) produces poor outcomes, e.g. the objects in the generated images are usually in the wrong pose. Moreover, the appearance of objects in generated images looks vastly different from the corresponding objects in the source-view images. We hypothesize this is because the CLIP Image encoder can only output a single token for each input image which is a high-level semantic summary. Thus, it is usually not enough to maintain image details. By removing attention conditioning, we find the outcomes are almost the same, compared to the default setting. This demonstrates U-Net conditioning dominates outcomes of Zero123 whereas CLIP embedding is almost ignored.

7. Qualitative Results

Figure 11 shows more qualitative results. It can be observed that compared to Zero123, our method can produce high-quality images. In some examples, Zero123 produces objects in the wrong pose, the wrong shapes, or multiple objects whereas only a single object exists in the source-view images. This may be because there exists high uncertainty when only a single source-view image is available. Unfortunately, Zero123 cannot handle this uncertainty well, especially when the difference between source-view and target-view is large. Intuitively, the uncertainty usually decreases as the available information (source-view images) increases. By receiving variable-length input views, the proposed method *VaLID* can utilize multi-view image information thus producing high-quality images. We can observe a clear improvement when the number of input views increases. Even if only a single source-view image is avail-

able, it can still outperform Zero123 qualitatively.

To further demonstrate the utilization and fusion of multi-view input images by the proposed *VaLID* method, Figure 12 shows the generated images with or without stage 2 training. As it can be observed in these examples, after stage 1 training, the model has the ability to produce plausible outcomes (see column (b)). Although our model in stage 1 has the flexibility to receive variable-length input views, it has not been trained to fuse multi-view inputs. So, at this stage, the inference on multi-view inputs shows inconsistency in data fusion to generate reasonable output (see columns (c-e)). In other words, since the training inputs in stage 1 of training always contain a single view image, the multi-view Cross Former block has not been adapted to perform the multi-view fusion task. To empower the model to perform the multi-view fusion, in stage 2 of training where only Cross Former parameters are tuned, the variable number of views are introduced as inputs. With this efficient strategy, Multi-view Cross Former learns how to combine provided information from multiple images to generate a consistent output image. As the number of input source-view images increases, the quality of produced images will gradually increase (see columns (f-i)).

Finally, we show some qualitative results in the attached video to show our method can produce more consistent images compared to previous methods. These videos consist of generated images at a predefined camera trajectory. We can observe with the single source-view image as input, that our method already can produce more consistent outcomes. When more input views are available, the consistency is improved further.

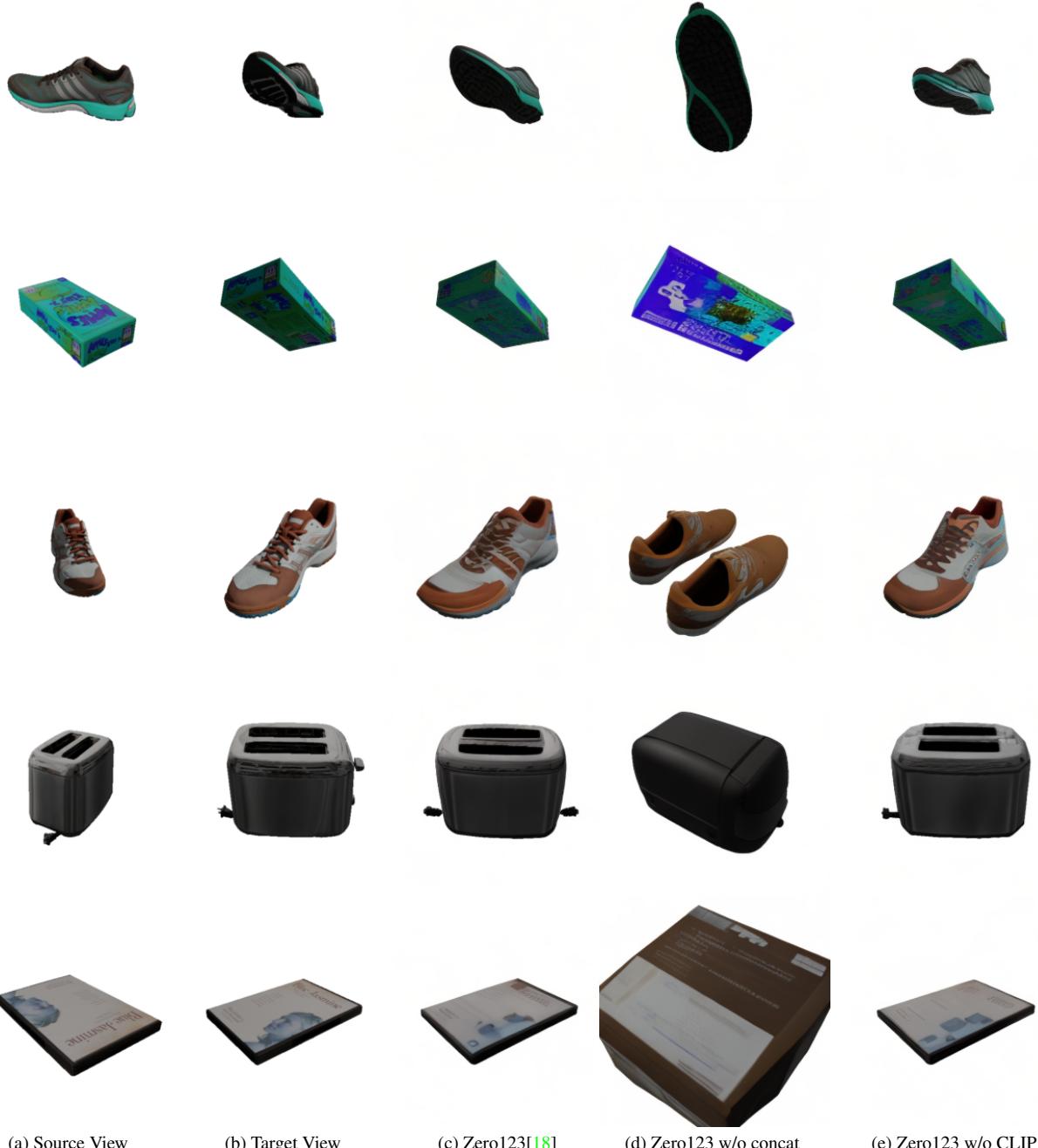


Figure 10. Zero123 conditioning strategy.

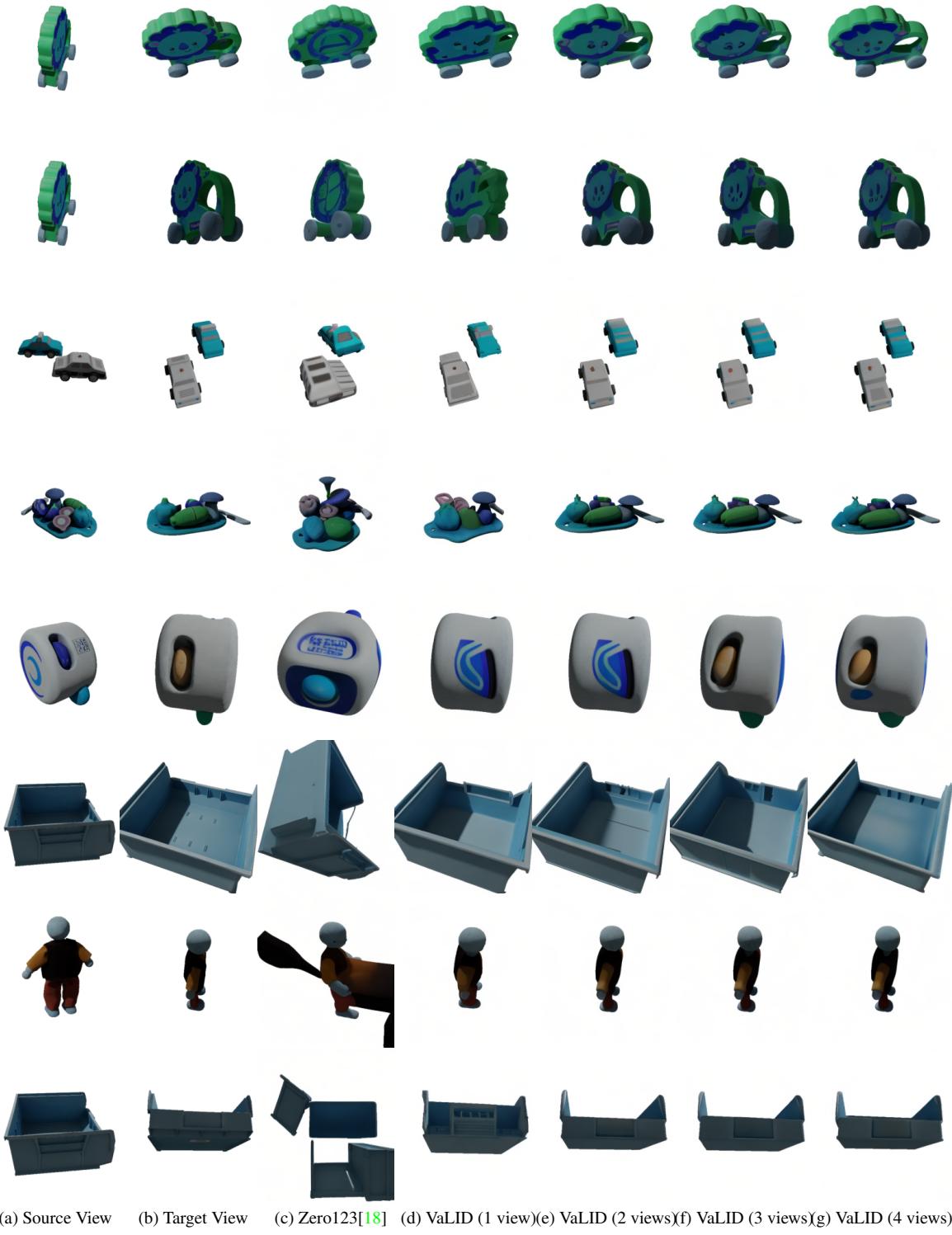


Figure 11. More qualitative examples.

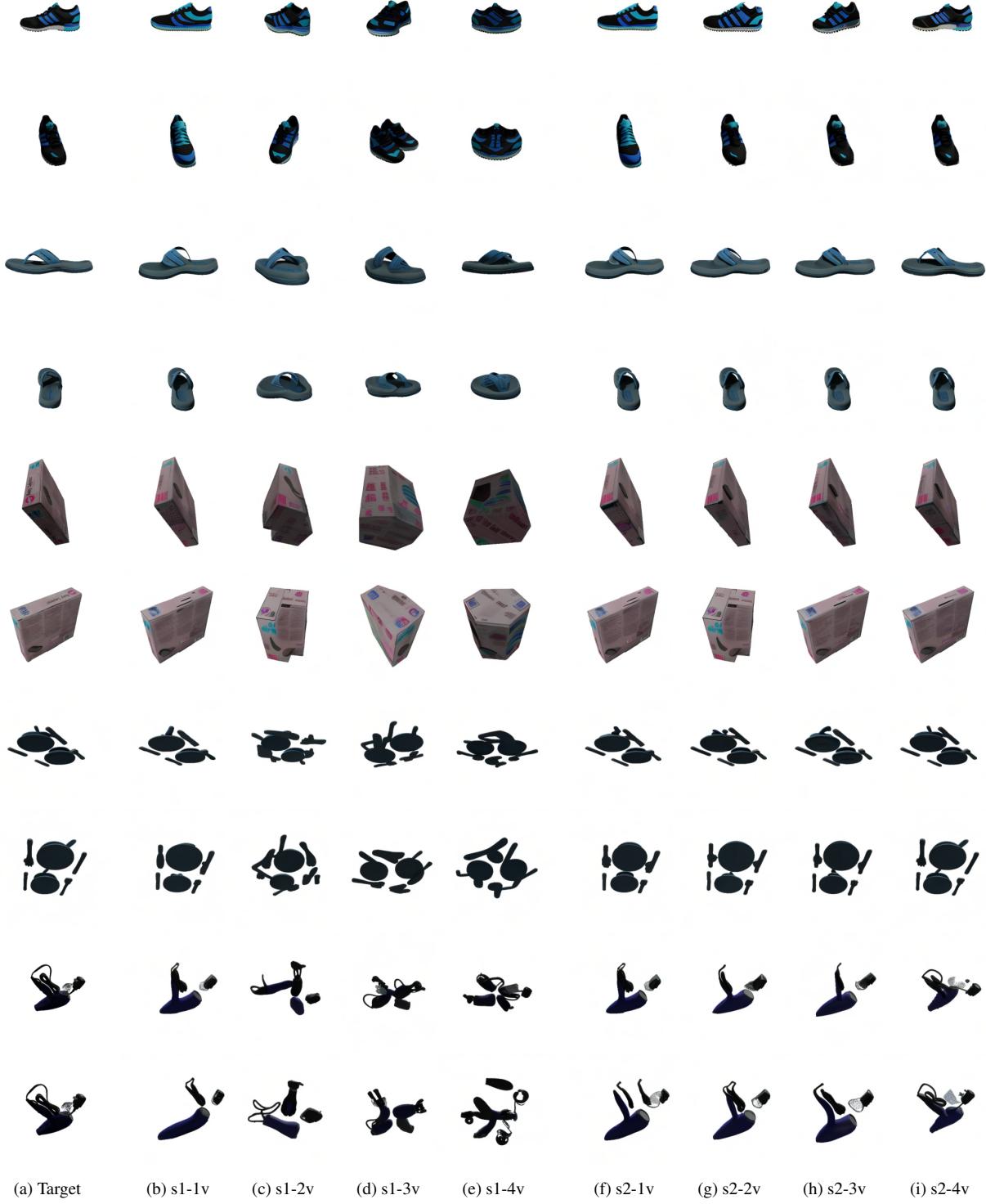


Figure 12. Impact of stage 2 training of the proposed VaLiD method. Columns (b)-(e) show the inference results on the variable number of input views (up to 4 views) after stage 1 training (s1). Columns (f)-(i) show the inference results after stage 2 training (s2) when the Cross Former parameters are tuned to perform multi-view image fusion.