VaLID: Variable-Length Input Diffusion for Novel View Synthesis — Supplementary Material —

Shijie Li[1,2*] Farhad G. Zanjani[2] Haitam Ben Yahia[2] Yuki Asano[2] Jürgen Gall[1]
Amirhossein Habibian[2]

[1]University of Bonn [2]Qualcomm AI Research[†]

lsj@uni-bonn.de, gall@iai.uni-bonn.de, {fzanjani, hyahia, asano, ahabibia}@qti.qualcomm.com

1. Conditioning Strategy

As mentioned in the main paper, Zero123 [1] 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 1 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.

2. Ablation Study

The importance of the stage 2 training is shown in Figure 2 (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 previously. As for Zero123 [1], 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.

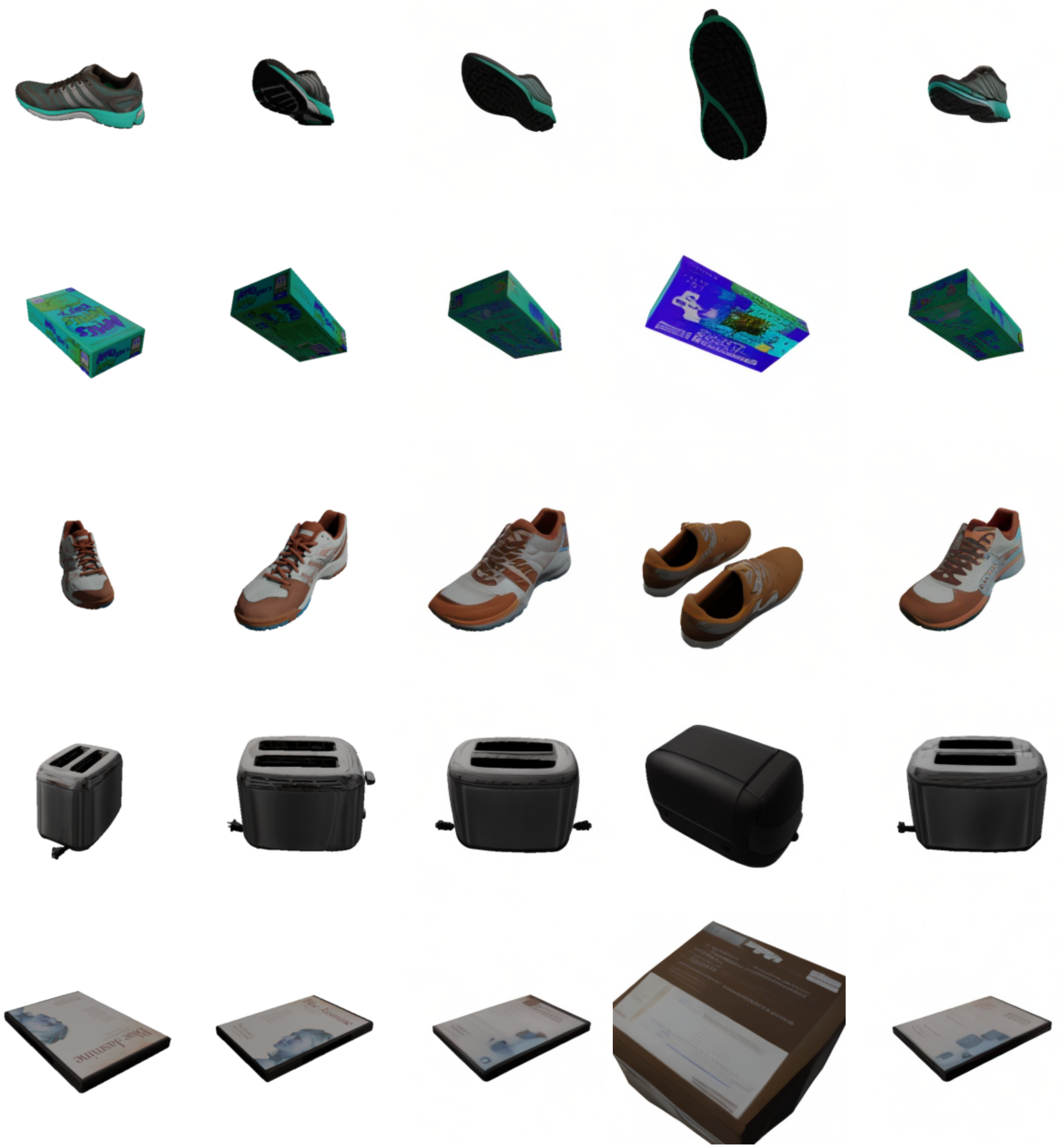
In inference time, we try to validate the robustness of the proposed method. This is achieved by feeding partial tokens into Multi-view Cross Former in the inference time. The results are shown in Figure 2 (b)-(d). The number is reported on 5 runs. It can be observed that when the number of available 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.

3. Qualitative Results

Figure 3 shows more qualitative results. It can be observed that compared to Zero123, our method can produce high-quality images. In some examples, Zero123 produces

[*] Work was completed during an internship at Qualcomm AI Research.

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

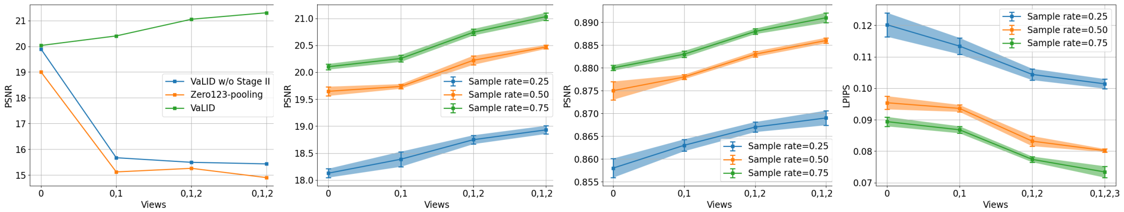


(a) Source View (b) Target View (c) Zero123 [1] (d) Zero123 w/o concat (e) Zero123 w/o CLIP

Figure 1. Zero123 conditioning strategy.

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 available, it can still outperform Zero123 qualitatively.

To further demonstrate the utilization and fusion of multi-view input images by the proposed *VaLID* method,



(a) Multi-view Comparison

(b) PSNR

(c) SSIM

(d) LPIPS

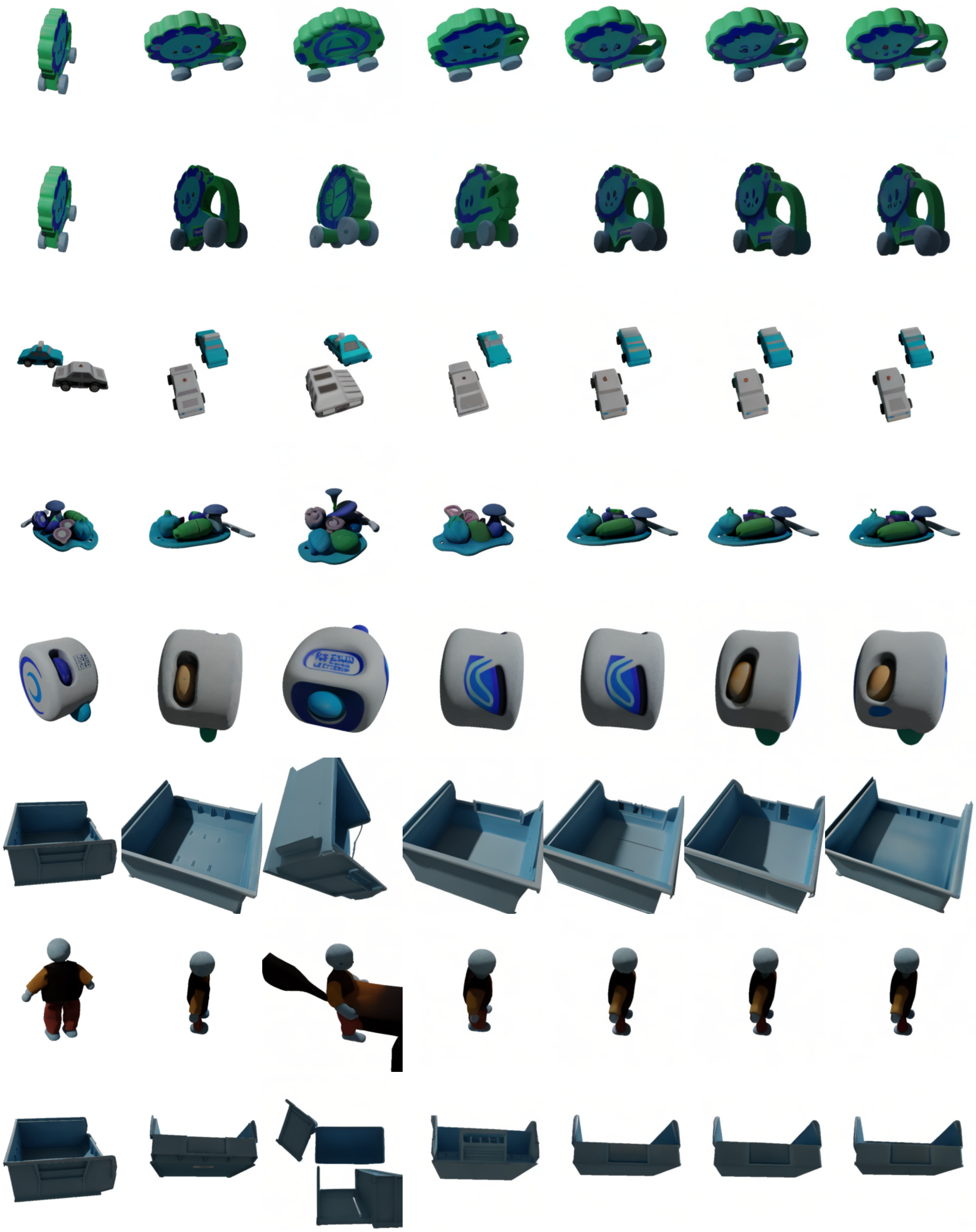
Figure 2. We show the necessity of stage 2 training in (a). The effect of token sampling at Inference time is shown in (b)-(d). Green line denotes sampling 75% tokens before Multi-view Cross Former. Red line and Blue line represent sample ratio equals to 50% and 25% respectively. We can observe as the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

Figure 4 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.

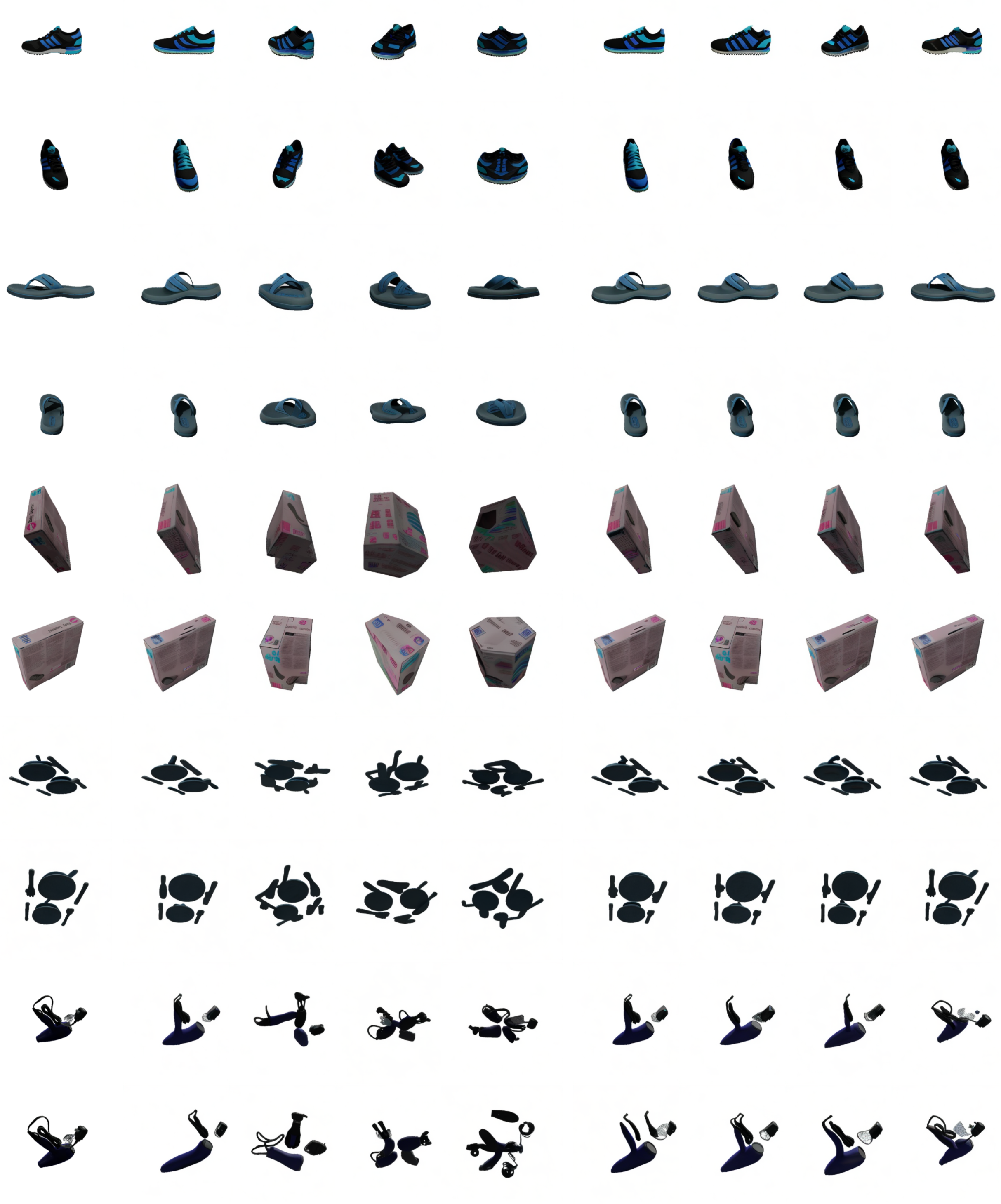
References

[1] 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.



(a) Source View (b) Target View (c) Z123 [1] (d) VaLID (1 view) (e) VaLID (2 views) (f) VaLID (3 views) (g) VaLID (4 views)

Figure 3. More qualitative examples.



(a) Target (b) s1-1v (c) s1-2v (d) s1-3v (e) s1-4v (f) s2-1v (g) s2-2v (h) s2-3v (i) s2-4v

Figure 4. 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.