

Table 1: Quantitative comparisons on CLIP similarity in NeRF generation.

Methods	ViT-B-32	ViT-L-14	ViT-g-14
VFDS(Poole et al., 2022)	32.13	31.85	31.78
VF-CSD(Yu et al., 2023)	32.46	31.57	32.02
VF-ISMLiang et al. (2023)	32.72	32.96	33.14
FlowDreamer	<b>34.96</b>	<b>34.19</b>	<b>34.58</b>

Table 2: Quantitative comparisons on CLIP similarity in 3D Gaussian splatting generation.

Methods	ViT-B-32	ViT-L-14	ViT-g-14
VFDS(Poole et al., 2022)	28.32	28.48	29.08
VF-CSD(Yu et al., 2023)	28.36	28.03	28.56
VF-ISMLiang et al. (2023)	29.56	29.52	29.87
FlowDreamer	<b>30.70</b>	<b>30.49</b>	<b>30.66</b>

## APPENDIX

The overview of the Appendix: We provide comparisons under a unified framework (see .1), some illustrations, including the effects of the Transformer Jacobian term, a comparison of SDS, VFDS and FlowDreamer with different steps, and an example of the initialization difference between NeRF and 3D GS models (see .2). Accordingly, we introduce the details of the experiments in .3. To better compare with other methods, we also conduct a user study to evaluate user preferences in .4. The derivation process of VF-ISM with the Rectified Flow prior is discussed in .5. Additionally, we discuss the steps of NFE and sampling methods in .6. More results include comparisons under 3D GS and NeRF generation settings and additional results from our FlowDreamer (see .7).

### .1 COMPARISONS UNDER A UNIFIED FRAMEWORK AND ABLATION STUDIES

#### .1.1 COMPARISONS UNDER A UNIFIED FRAMEWORK

In the baseline methods, SD2.1 was chosen, whereas Flow-based SD3 was used in our method. Taking the reviewers’ suggestions into account, we replaced SD2.1 in the comparative methods with SD3, as shown in Figure 2 and Figure 3. However, since the baseline methods were specifically designed for the Diffusion model or adjusted 3D model parameters within Diffusion model, directly transferring them to SD3 results in limited improvements and, in some cases, even worse performance (for example, the DreamGaussian results for the prompt ”an origami pig”). Forcing a direct transfer to SD3 for comparison also leads to unfairness.

Therefore, we continued the examples shown in Figure 7 of the original paper and transferred the loss functions into a unified framework. In the original paper, we had already provided VFDS and VF-ISM; now, we have further transferred the Consistent3D loss into the unified framework, referred to as VF-CSD below. However, the original paper did not include many examples or experiments. To address this, we have conducted comparative experiments for a variety of prompts. All experiments use SD3 as the prior, with the same random seeds, NeRF settings, and 3D GS settings; the only difference lies in the loss design. We have provided a large number of experimental images for the reviewers to compare. Please refer to Fig. 16 to Fig. 28. The results with orange borders correspond to 3D GS, while the results with green borders correspond to NeRF. After comparison, **Our results yield high-fidelity outputs with richer textual details compared to other baseline methods using the same SD3 prior.**

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

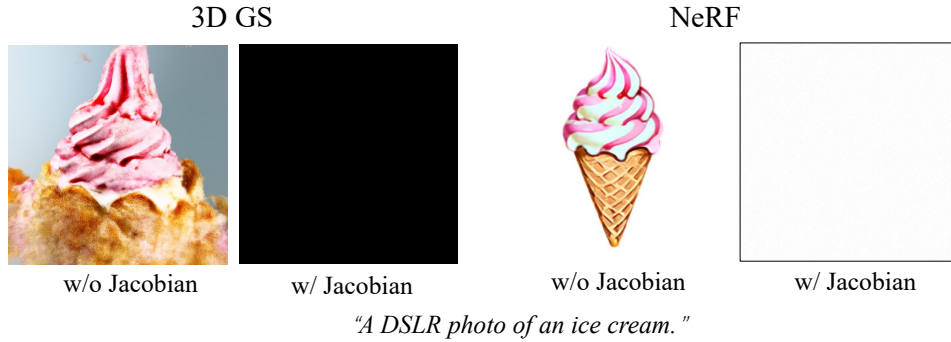


Figure 1: VFDS results of ignoring the Transformer Jacobian. We found that it is difficult to generate meaningful results.

### 1.2 QUANTITATIVE COMPARISONS ON CLIP SIMILARITY IN NeRF AND 3D GS GENERATION

All experiments use SD3 as the prior, with the same random seeds, NeRF settings, and 3D GS settings; the only difference lies in the loss design. We randomly select 12 images with the same viewpoints from the rendered images and employ three CLIP models from OpenCLIP, ViT-B-32, ViT-L-14, and ViT-g-14 to calculate the CLIP similarity. **Our FlowDreamer achieves superior CLIP similarity in both NeRF and 3D GS scenarios.**

### 1.3 ABLATION STUDIES

As shown in Figure 7, we also provide some examples for comparison with the results of VF-ISM with  $\eta_t$ , VF-ISM, and FlowDreamer. As shown in Figure 8, as the number of warm-up steps increases, the results improve, and at 1200 steps, the performance becomes highly stable. A comparison of DDIM inversion and Push-backward at different NFEs and CFG scales is shown in Figure 9.

## 2 SOME ILLUSTRATIONS

We provide some illustrations for a better understanding of our paper. Figure 1 illustrates the effects of the Transformer Jacobian term. It is shown that keeping this term leads to training crashes, making it very difficult to generate meaningful 3D objects. Therefore, we omit this term to achieve an effective gradient for optimization. Figure 10 illustrates an example of the initialization difference between NeRF and 3D GS models.

## 3 IMPLEMENTATION DETAILS

We adopt Stable Diffusion 3 (SDv3) (Esser et al., 2024) as our Rectified flow model. To facilitate a better comparison, we will categorize the methods into two types: one where the 3D model is NeRF, and the other where the 3D model is 3D GS. For NeRF results comparisons, unlike ProlificDreamer (Wang et al., 2024) and Consistent3D (Wu et al., 2024), which use multi-stage rendering with normal, geometric, and texture rendering, our method simply uses Instant-NGP (Müller et al., 2022) for rendering. It is optimized with a resolution of 256 for the first 5000 steps with a batch size of 1, and then 512 for the subsequent 3000 steps, also with a batch size of 1. For the warm-up strategy, we use VFDS to optimize over 1200 steps. The initial framework VFDS need more steps to optimize, which uses 5000 steps for both 256 and 512 resolutions, with the batch size always set to 1. For 3D Gaussian splatting comparisons, we utilize the pretrained PointE (Nichol et al., 2022) to initialize the locations of 3D Gaussians, while other properties of 3D Gaussians adopt random initialization. Our FlowDreamer uses a batch size of 4 and 3000 iterations, while VFDS uses the same batch size but with 4000 iterations. The CFG for VFDS is set to 100 on both NeRF and 3D GS, while it is set to 40 for FlowDreamer.



Figure 2: Comparison of SOTA methods using the Stable Diffusion 3 prior.

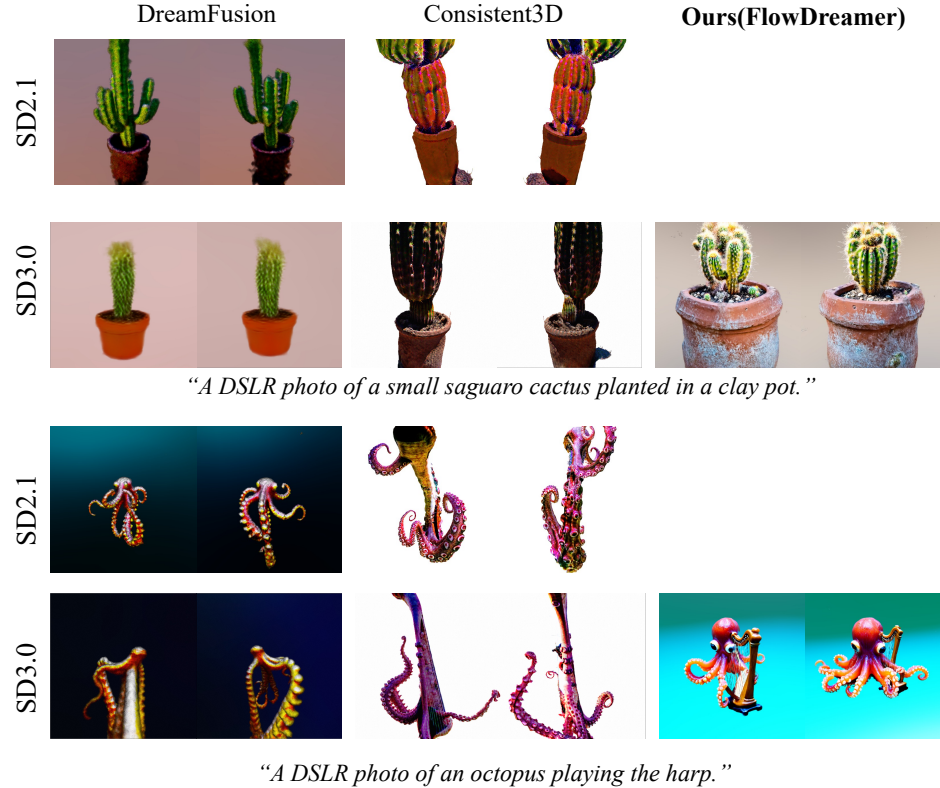


Figure 3: Comparison of SOTA methods using the Stable Diffusion 3 prior.

#### 4 USER STUDY

To compare our method with other comparison methods based on human perception, we conducted a user study involving 36 participants using generated videos from 26 prompts in NeRF and 3D

Table 3: User study of 3D Gaussian splatting methods.

Methods	DreamGaussian (Tang et al., 2023)	GaussianDreamer (Yi et al., 2023)	LucidDreamer (Liang et al., 2023)	<b>Ours</b>
Scores	1.18	2.03	3.09	3.70

Table 4: User study of NeRF methods.

Methods	DreamFusion (Poole et al., 2022)	ProlificDreamer (Wang et al., 2024)	Consistent3D (Wu et al., 2024)	<b>Ours</b>
Scores	1.46	2.70	2.18	3.65

Gaussian splatting, respectively. Participants viewed four videos simultaneously and assigned scores of 1, 2, 3, and 4 to each video. In each test, the prompts were shown in the title, and participants were instructed to make their decisions based on the degree of alignment of the video with the text, the detail of the video, the color of the video, and the quality of the video. A higher score indicates that users believe the video is better. Our FlowDreamer achieves the highest scores among these methods, both in NeRF results and in 3D Gaussian splatting results.

## 5 VF-ISM DERIVATION

ISM uses DDIM inversion to predict the noisy latent  $x_s$  as below.

$$\begin{aligned} x_s &= \sqrt{\alpha_s} \hat{x}^{s-\delta_T} + \sqrt{1 - \alpha_s} \epsilon_\phi(x_{s-\delta_T}, s - \delta_T, \emptyset) \\ &= \sqrt{\alpha_s} (\hat{x}^{s-\delta_T} + \gamma(s) \epsilon_\phi(x_s, s, \emptyset)) \end{aligned} \quad (1)$$

where  $x^{s-\delta_T} = \frac{1}{\sqrt{\alpha_s}} x_{s-\delta_T} - \gamma(s - \delta_T) \epsilon_\phi(x_{s-\delta_T}, s - \delta_T, \emptyset)$ ,  $\gamma(t) = \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}$  and  $s = t - \delta_T$ . The  $\hat{x}^{s-\delta_T}$  is computed using DDIM inference, while  $x$  represents the rendered image from the 3D model, as shown below.

$$\begin{aligned} \hat{x}^{s-\delta_T} &= x - \gamma(\delta_T) [\epsilon_\phi(x_{\delta_T}, \delta_T, \emptyset) - \epsilon_\phi(x, 0, \emptyset)] \dots \\ &\quad - \gamma(s - \delta_T) [\epsilon_\phi(x_{s-\delta_T}, s - \delta_T, \emptyset) - \epsilon_\phi(x_{s-2\delta_T}, s - 2\delta_T, \emptyset)] \end{aligned} \quad (2)$$

Next, ISM computes  $x_t$  based on  $x_s$ .

$$\begin{aligned} x_t &= \sqrt{\alpha_t} \hat{x}^s + \sqrt{1 - \alpha_t} \epsilon_\phi(x_s, s, \emptyset) \\ \hat{x}^s &= x^{s-\delta_T} - \gamma(s) [\epsilon_\phi(x_s, s, \emptyset) - \epsilon_\phi(x_{s-\delta_T}, s - \delta_T, \emptyset)] \end{aligned} \quad (3)$$

After that, ISM computes the  $\tilde{x}^t$  with the denoising process.

$$\begin{aligned} \tilde{x}^t &= \frac{x_t}{\sqrt{\alpha_t}} - \gamma(t) \epsilon_\phi(x_t, t, y) + \gamma(s) [\epsilon_\phi(x_t, t, y) - \epsilon_\phi(x_s, s, y)] \\ &\quad + \dots + \gamma(\delta_T) [\epsilon_\phi(\tilde{x}_{2\delta_T}, 2\delta_T, y) - \epsilon_\phi(\tilde{x}_{\delta_T}, \delta_T, y)] \end{aligned} \quad (4)$$

ISM computes the  $x - \tilde{x}^t$  and with the DDIM inversion process.

$$\begin{aligned} x - \tilde{x}^t &= \gamma(t) [\epsilon_\phi(\tilde{x}_t, t, y) - \epsilon_\phi(x_s, s, \emptyset)] + \eta_t \\ \text{where, } \eta_t &= +\gamma(s) [\epsilon_\phi(\tilde{x}_s, s, y) - \epsilon_\phi(x_{s-\delta_T}, s - \delta_T, \emptyset)] - \gamma(s) [\epsilon_\phi(\tilde{x}_t, t, y) - \epsilon_\phi(x_s, s, \emptyset)] \\ &\quad + \dots \\ &\quad + \gamma(\delta_T) [\epsilon_\phi(\tilde{x}_{\delta_T}, \delta_T, y) - \epsilon_\phi(x, 0, \emptyset)] - \gamma(\delta_T) [\epsilon_\phi(\tilde{x}_{2\delta_T}, 2\delta_T, y) - \epsilon_\phi(x_{\delta_T}, \delta_T, \emptyset)] \end{aligned} \quad (5)$$

To derive VF-ISM, we adapt ISM to rectified flow. Concretly, we first conduct *push-backward* operation to compute the  $x_s$  with Euler sample method as below.

$$\begin{aligned} x_{\delta_T} &= x + v_\phi(x, 0, \emptyset) dt \\ x_{2\delta_T} &= x_{\delta_T} + v_\phi(x_{\delta_T}, \delta_T, \emptyset) dt \\ &\dots \\ x_s &= x_{s-\delta_T} + v_\phi(x_{s-\delta_T}, s - \delta_T, \emptyset) dt \end{aligned} \quad (6)$$



where,  $dt = \Delta_T$ .

Then we use the *push-backward* to compute the  $x_t$  with Euler method:

$$x_t = x_s + v_\phi(x_s, s, \emptyset)ds \quad (7)$$

where,  $ds = t - s$ .

Due to rectified flow, the  $\tilde{x}_t^t$  can be simply expressed as below:

$$\begin{aligned} \tilde{x}^t &= x_t - v_\phi(x_t, t, y)ds \\ &\quad - v_\phi(x_s, s, y)dt - v_\phi(x_{s-\delta_T}, s - \delta_T, y)dt \\ &\quad - \dots - v_\phi(x_{\delta_T}, \delta_T, y)dt \end{aligned} \quad (8)$$

Then we compute the  $x - \tilde{x}^t$  with *push-backward* with rectified flow prior:

$$\begin{aligned} x - \tilde{x}^t &= [v_\phi(x_t, t, y) - v_\phi(x_s, s, \emptyset)] ds + \eta_t \\ \text{where, } \eta_t &= [v_\phi(x_s, s, y) - v_\phi(x_{s-\delta_T}, s - \delta_T, \emptyset)] dt \\ &\quad + \dots \\ &\quad + [v_\phi(x_{2\delta_T}, 2\delta_T, y) - v_\phi(x_{\delta_T}, \delta_T, \emptyset)] dt \end{aligned} \quad (9)$$

Same as ISM, we ignore the  $\eta_t$  and use  $v_\phi(x_t, t, y) - v_\phi(x_s, s, \emptyset)$  as ISM with Rectified flow prior.

## 6 DISCUSSION OF THE STEPS OF NFE AND SAMPLE METHODS

As the number of iterations for *push-backward* increases, more training time are required. As NFE increases, the LEGO car exhibits more complex structures overall. When viewed vertically, each column represents results of different sampling methods with the same NFE steps. In terms of results for each column, the Euler sampling method demonstrates strong competitiveness, regardless of number of steps.

## 7 MORE RESULTS OF FLOWDREAMER

As shown in Figure 12 and Figure 13, our FlowDreamer can generate high-fidelity textures and shapes from pretrained rectified flow models both in 3D GS and NeRF. Our method can produce realistic objects, such as sweaters and wooden bowls, including fantastical ones that are rare in reality, like pumpkins with glowing runes and baby dragons.

Comparison with other methods in text-to-3D generation for NeRF and 3D GS, respectively, shows that our FlowDreamer creates 3D objects that match well with the input text prompts, exhibiting high fidelity and intricate details. The coupe is realistic in both color and shape, and the generated watch dial appears lifelike in 3D GS results (See Figure 14). The burrito’s shape and quality are more aligned with the prompts, and the generated tacos are even more realistic in NeRF results (See Figure 15).

270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323

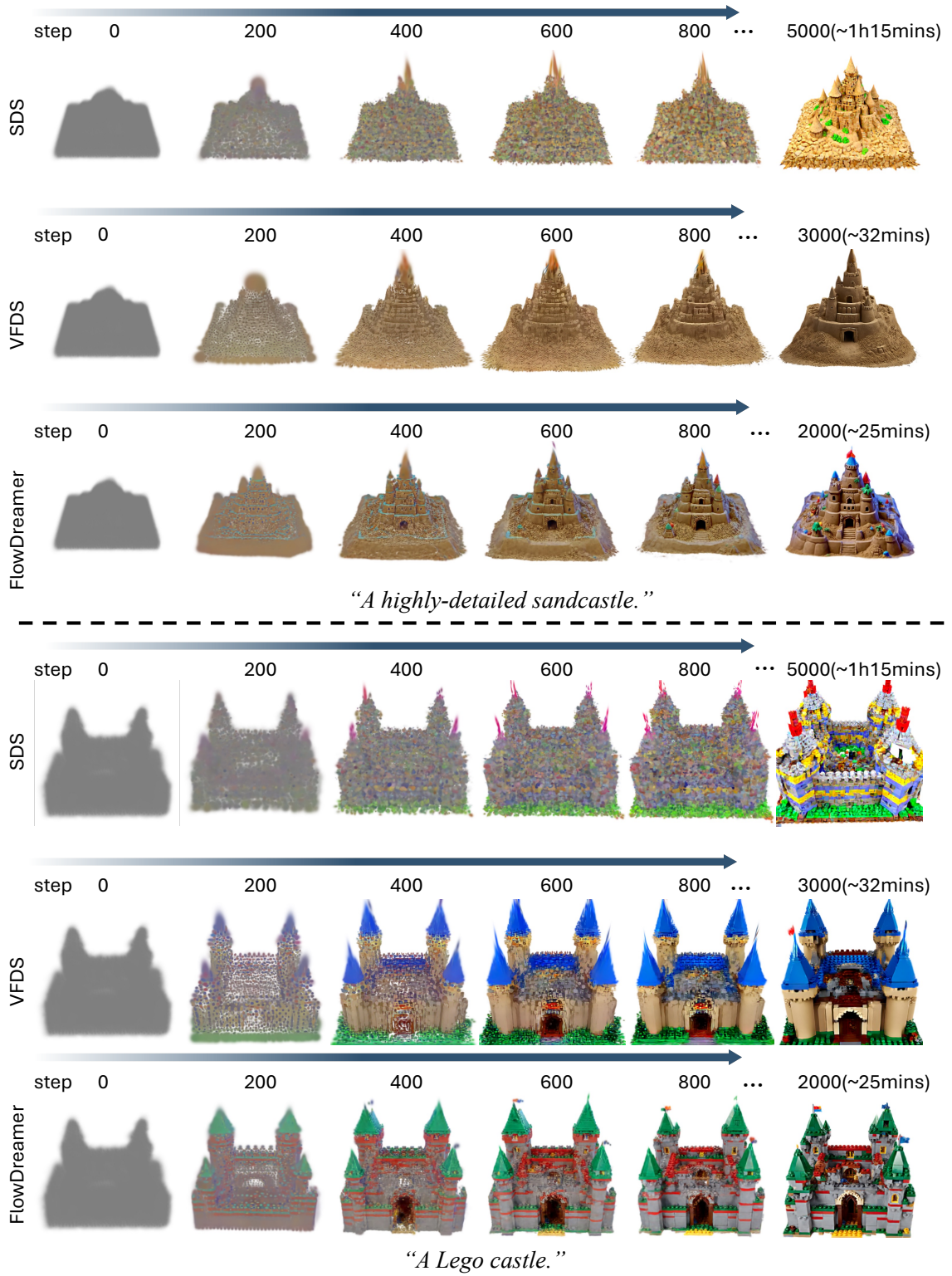


Figure 4: A comparison of our VFDS and FlowDreamer and SDS with different steps

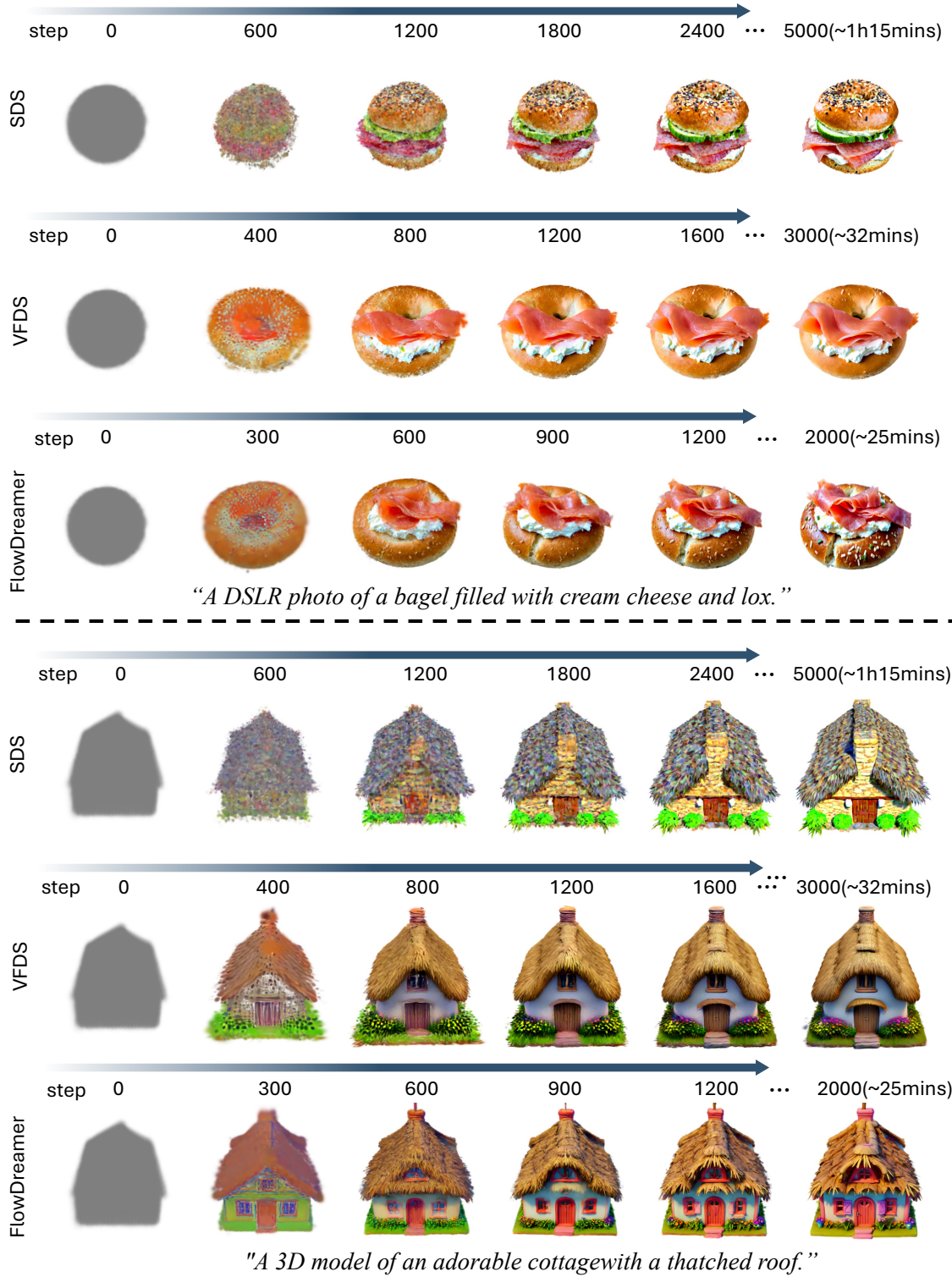
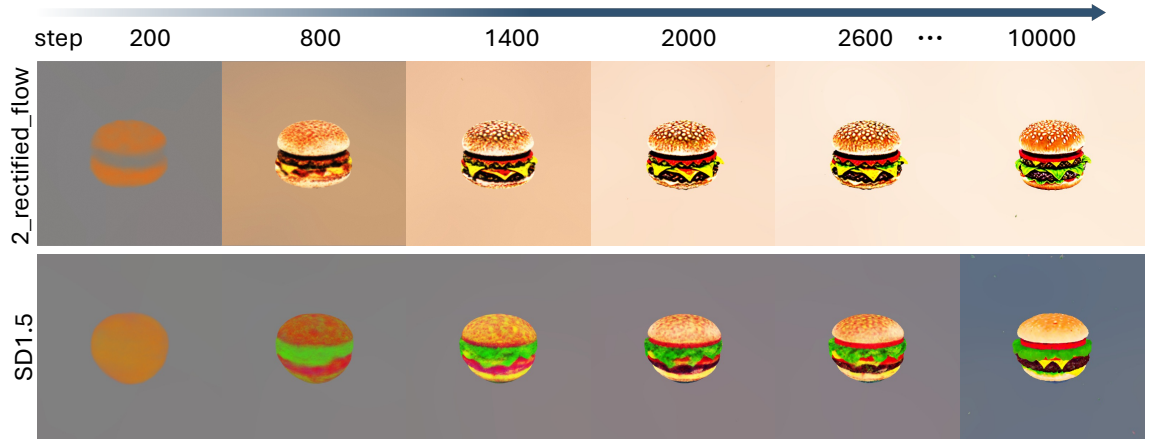


Figure 5: A comparison of our VFDS and FlowDreamer and SDS with different steps

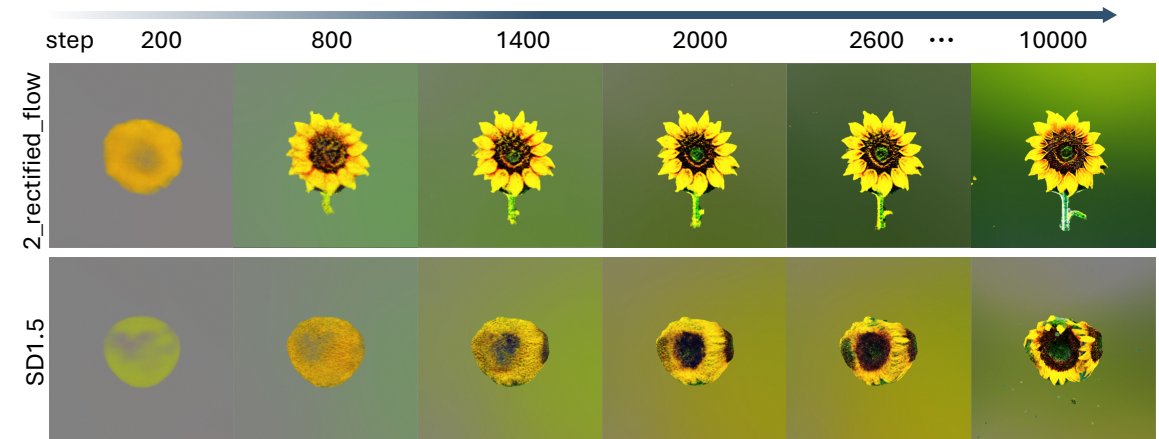
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431



*"A hamburger."*



*"A photo of a bagel filled with cream cheese and lox."*



*"A photo of a vibrant sunflower."*

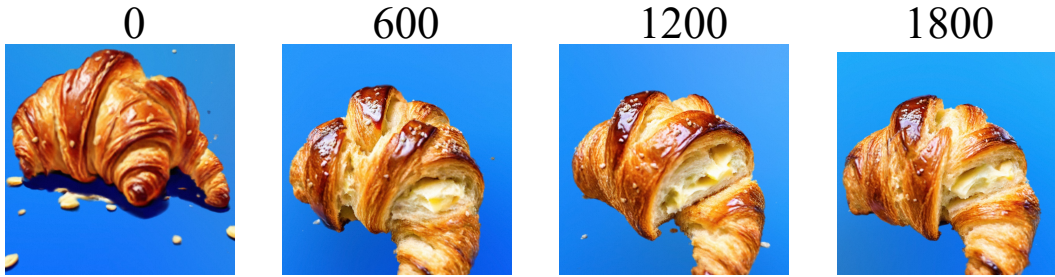
Figure 6: A comparison of VFDS and SDS with different steps. SDS uses SD.15 as the prior, while VFDS uses 2\_rectified\_flow from Instaflow as the prior.



432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485



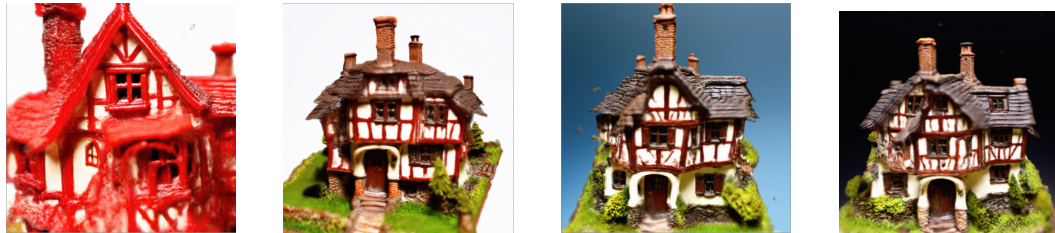
Figure 7: Ablation study for the ISM with  $\eta_t$ . The results show that the ISM with  $\eta_t$  becomes smoother.



*A delicious croissant.*



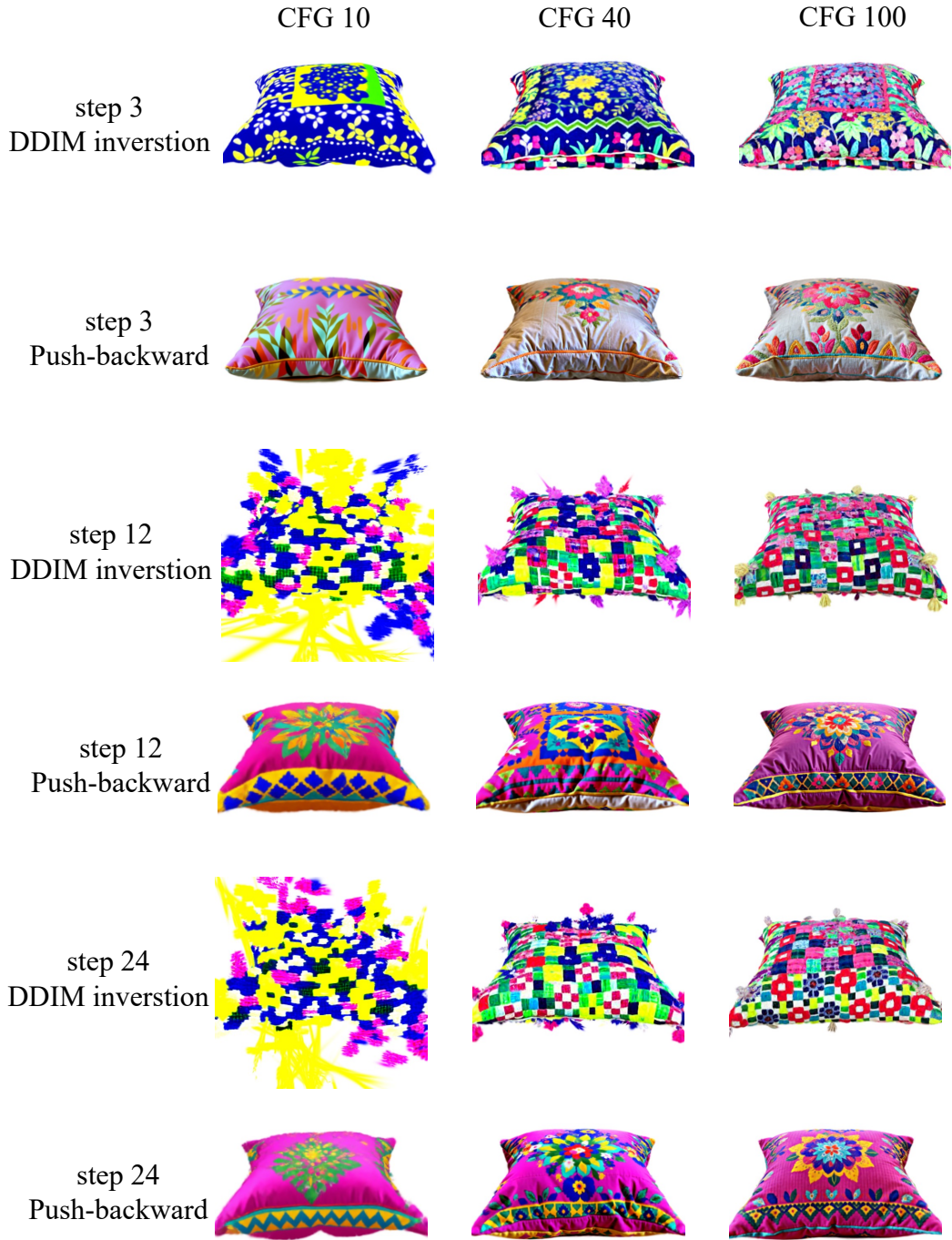
*Michelangelo style statue of dog reading news on a cellphone.*



*A model of a house in Tudor style.*

Figure 8: Ablation study for different warm up steps. As the number of warm-up steps increases, the results improve, and at 1200 steps, the performance becomes highly stable.

540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593



*“A decorative throw pillow.”*

Figure 9: A comparison of DDIM inversion and *Push-backward* in different NFEs and CFG scales.



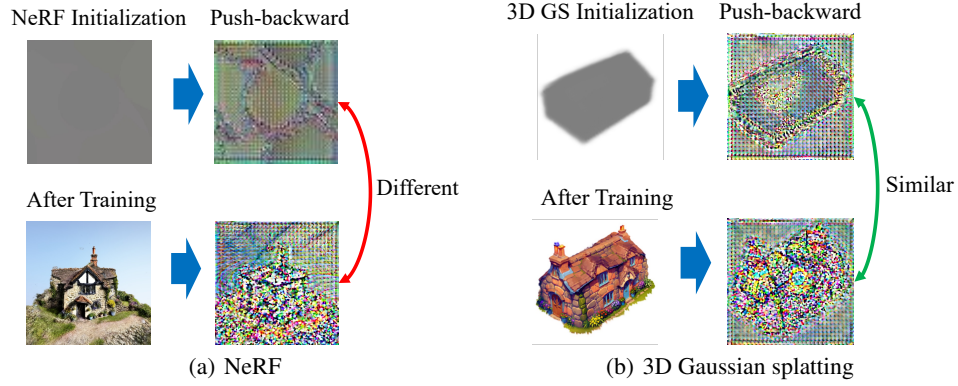


Figure 10: Comparison of initialization between NeRF and 3D GS models. (a) Images generated by NeRF models before and after training are different, leading to different push-backward results. (b) As for 3D Gaussian splatting model, the results are quite similar.

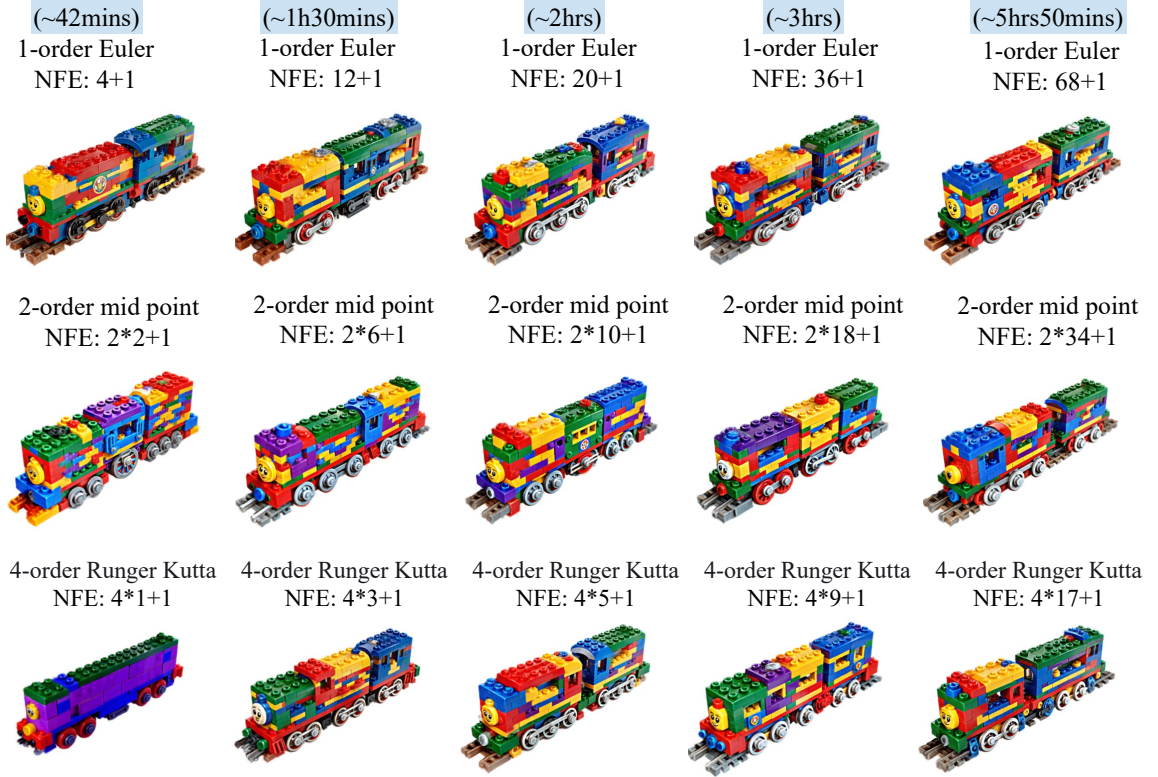


Figure 11: The results of different NFEs and sample methods.

648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701



Figure 12: More results under 3D GS generation setting. Please zoom in for details



Figure 13: More results under NeRF generation setting. Please zoom in for details

702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

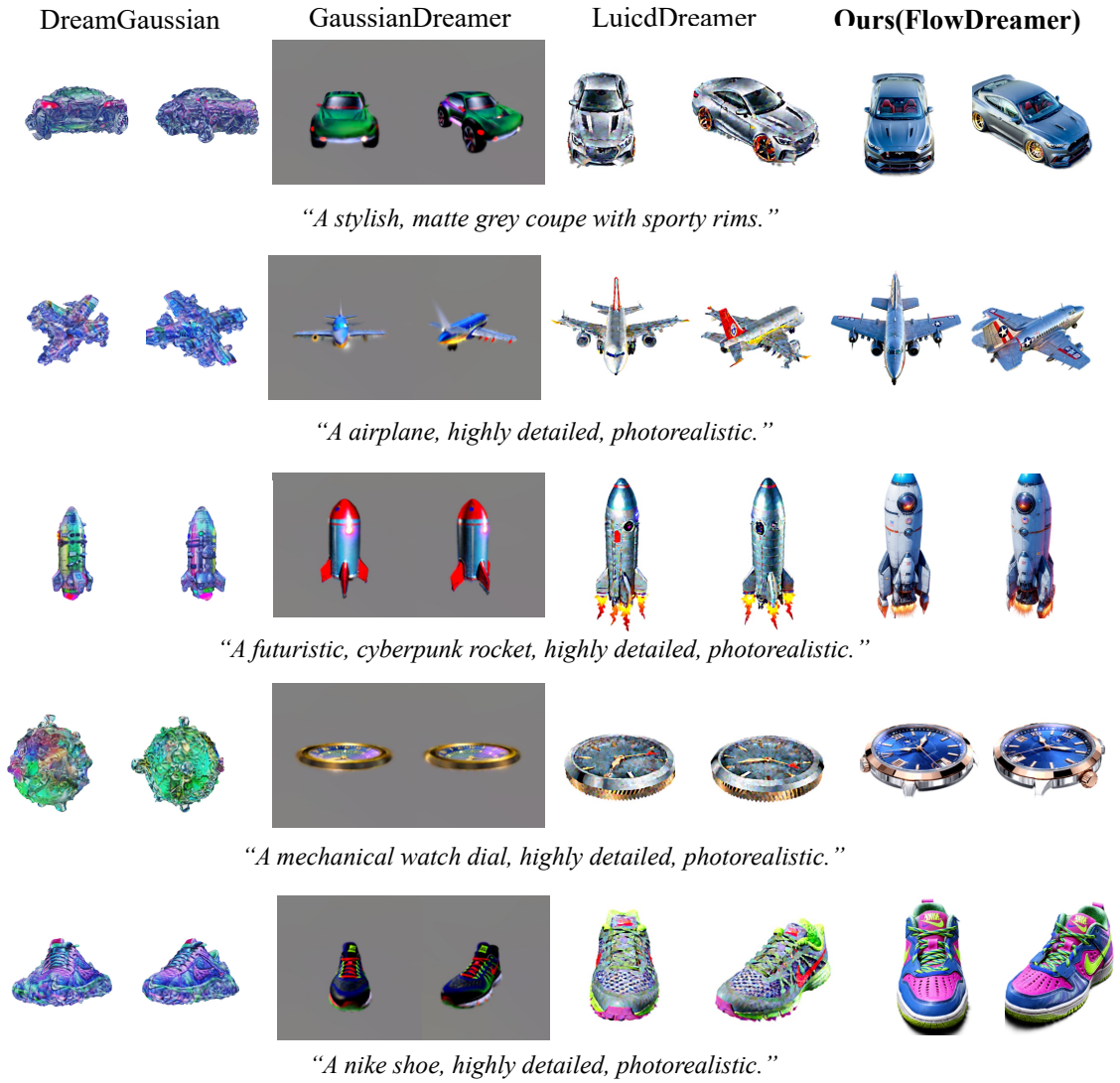


Figure 14: More qualitative comparison under 3D GS generation setting.

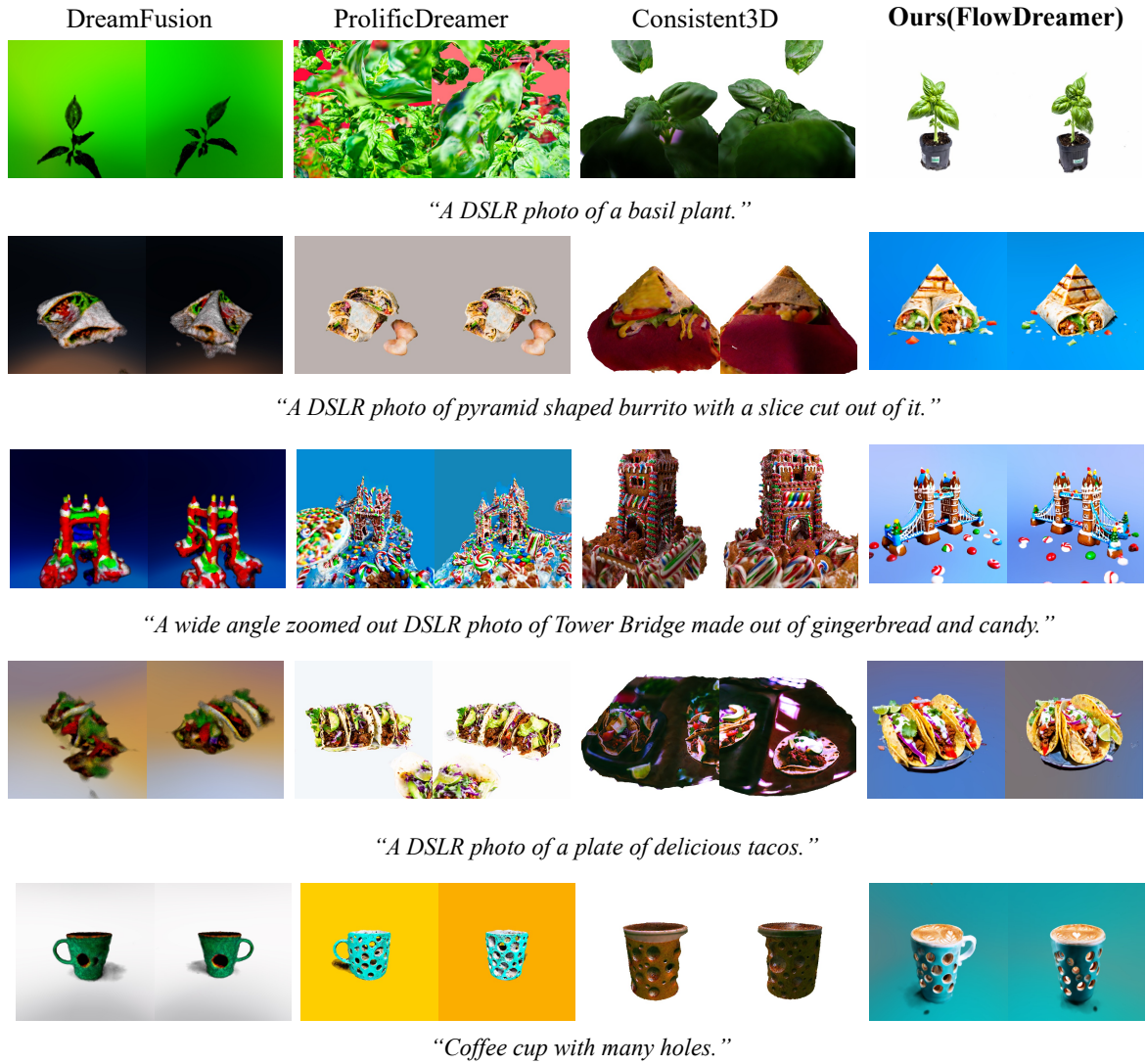


Figure 15: More qualitative comparison under NeRF generation setting.



810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863



Figure 16: More qualitative comparison under a unified framework

864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917



Figure 17: More qualitative comparison under a unified framework

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

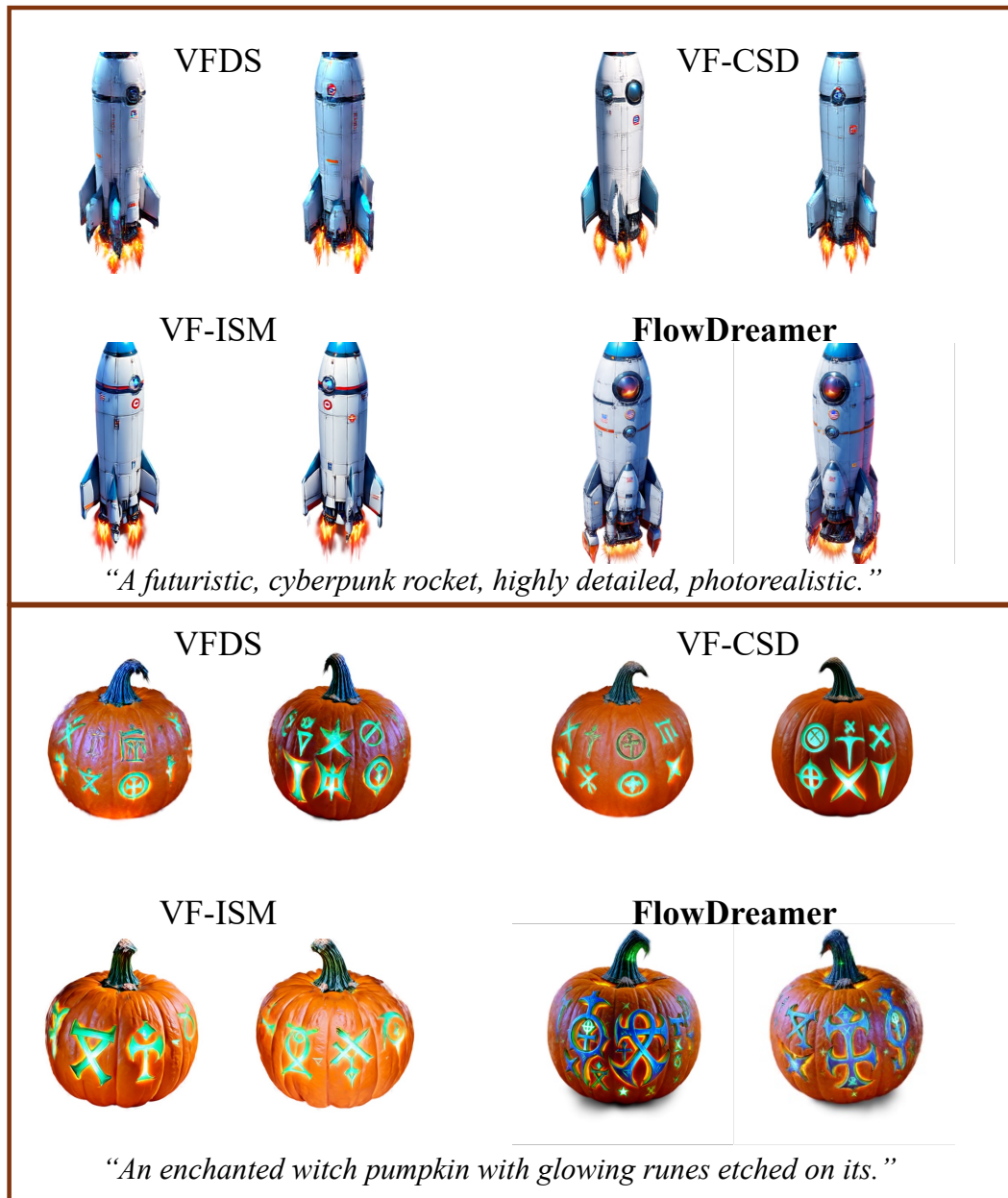


Figure 18: More qualitative comparison under a unified framework



972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025



Figure 19: More qualitative comparison under a unified framework

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079



Figure 20: More qualitative comparison under a unified framework

1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133

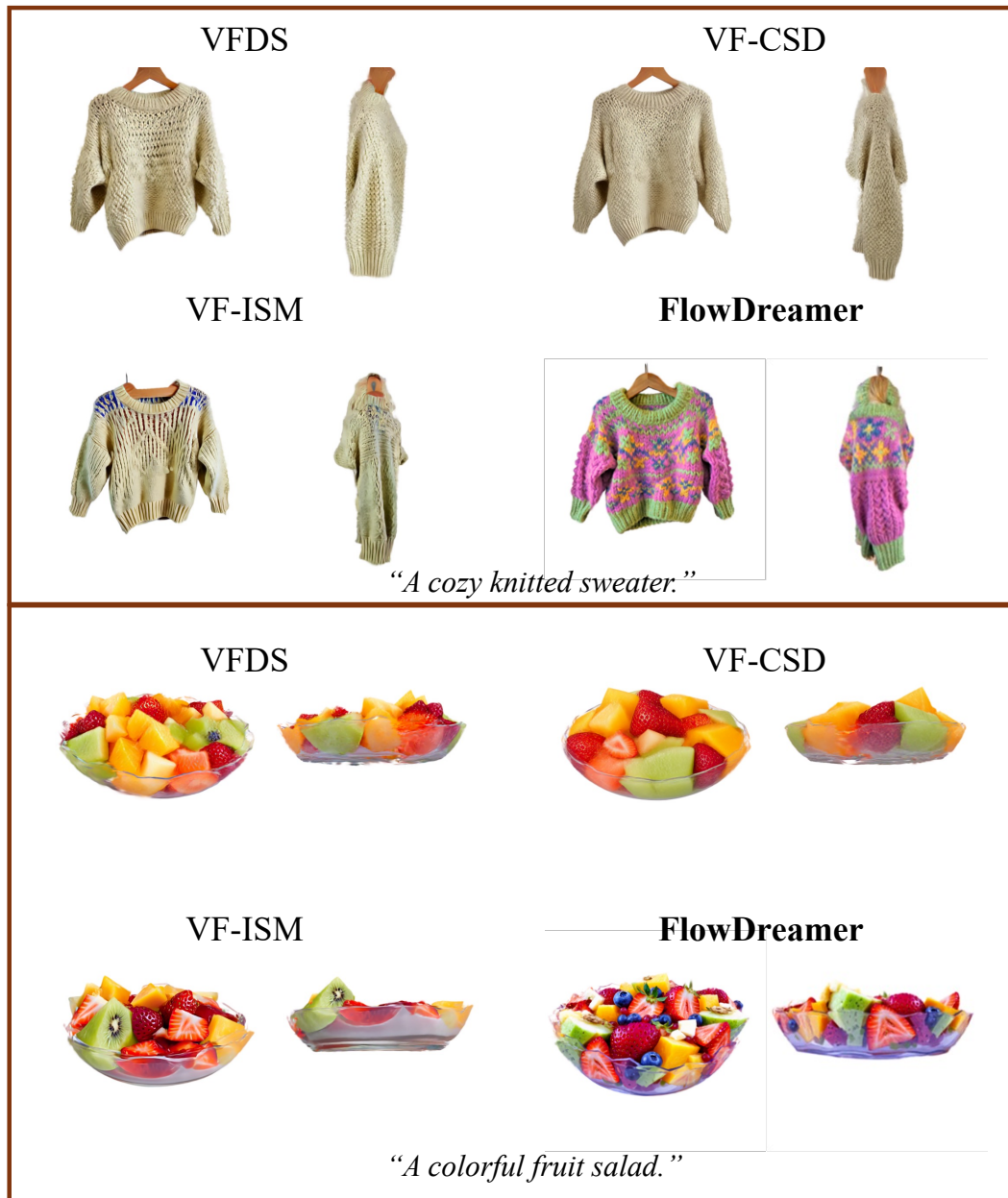


Figure 21: More qualitative comparison under a unified framework

1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187

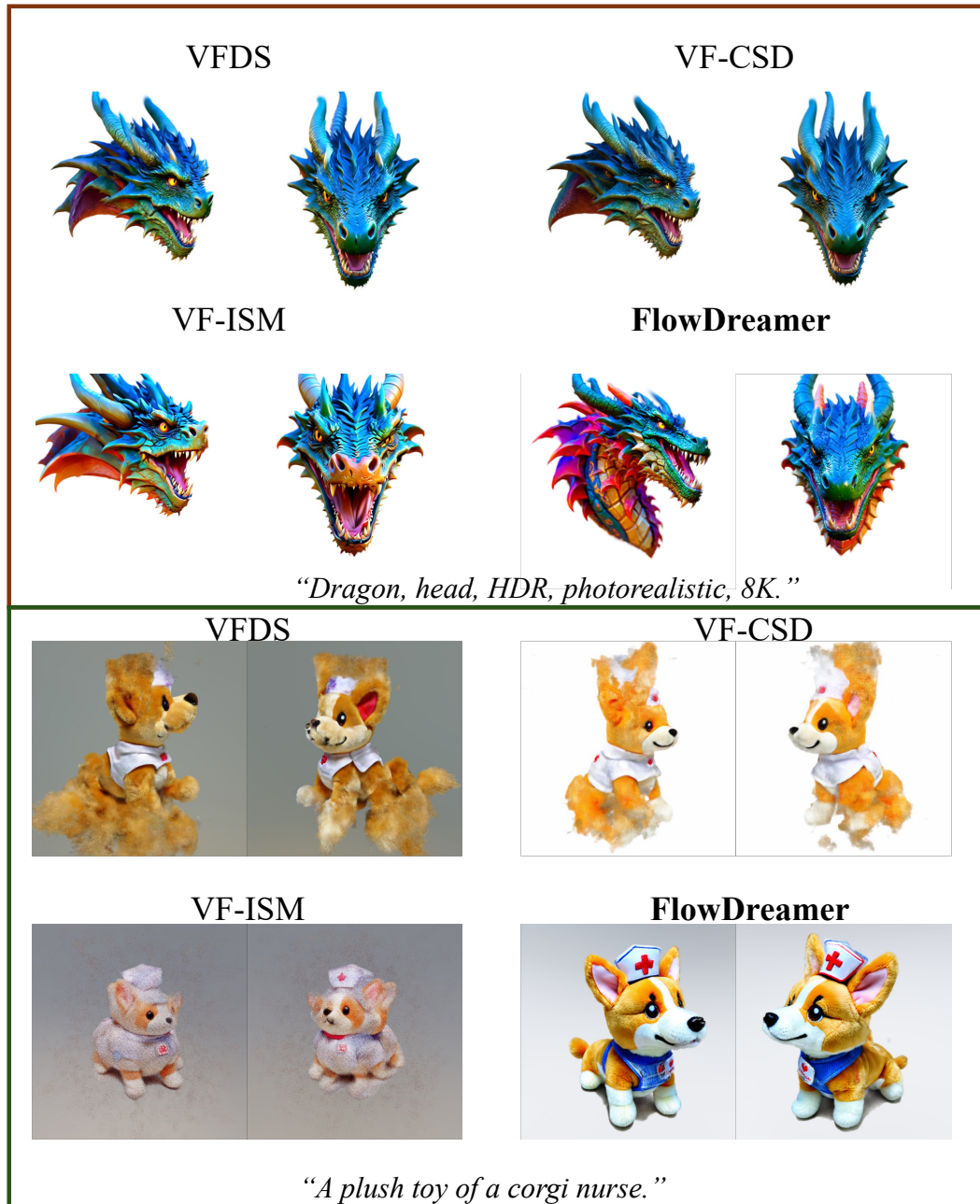


Figure 22: More qualitative comparison under a unified framework

1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241



Figure 23: More qualitative comparison under a unified framework



1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295

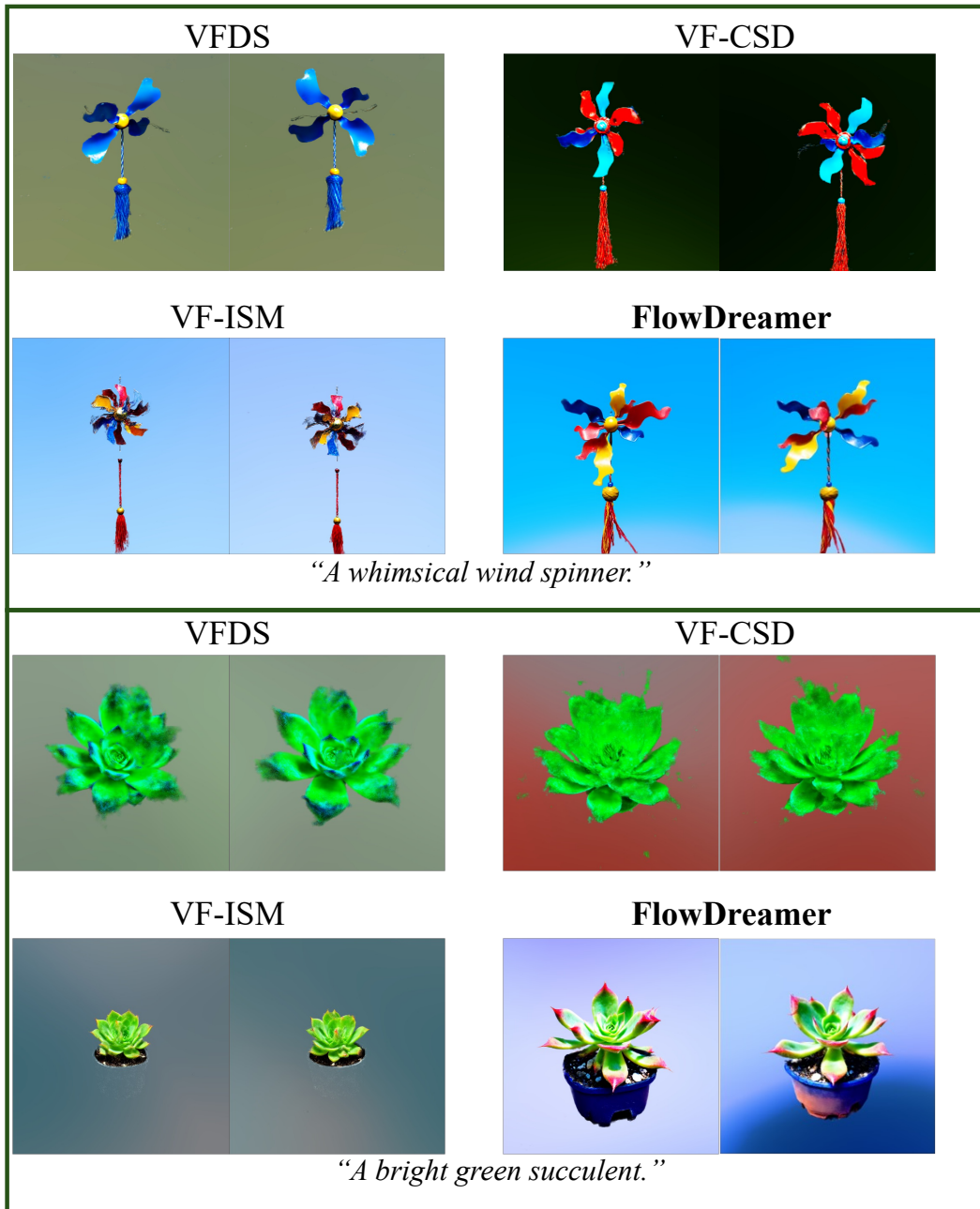


Figure 24: More qualitative comparison under a unified framework

1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349

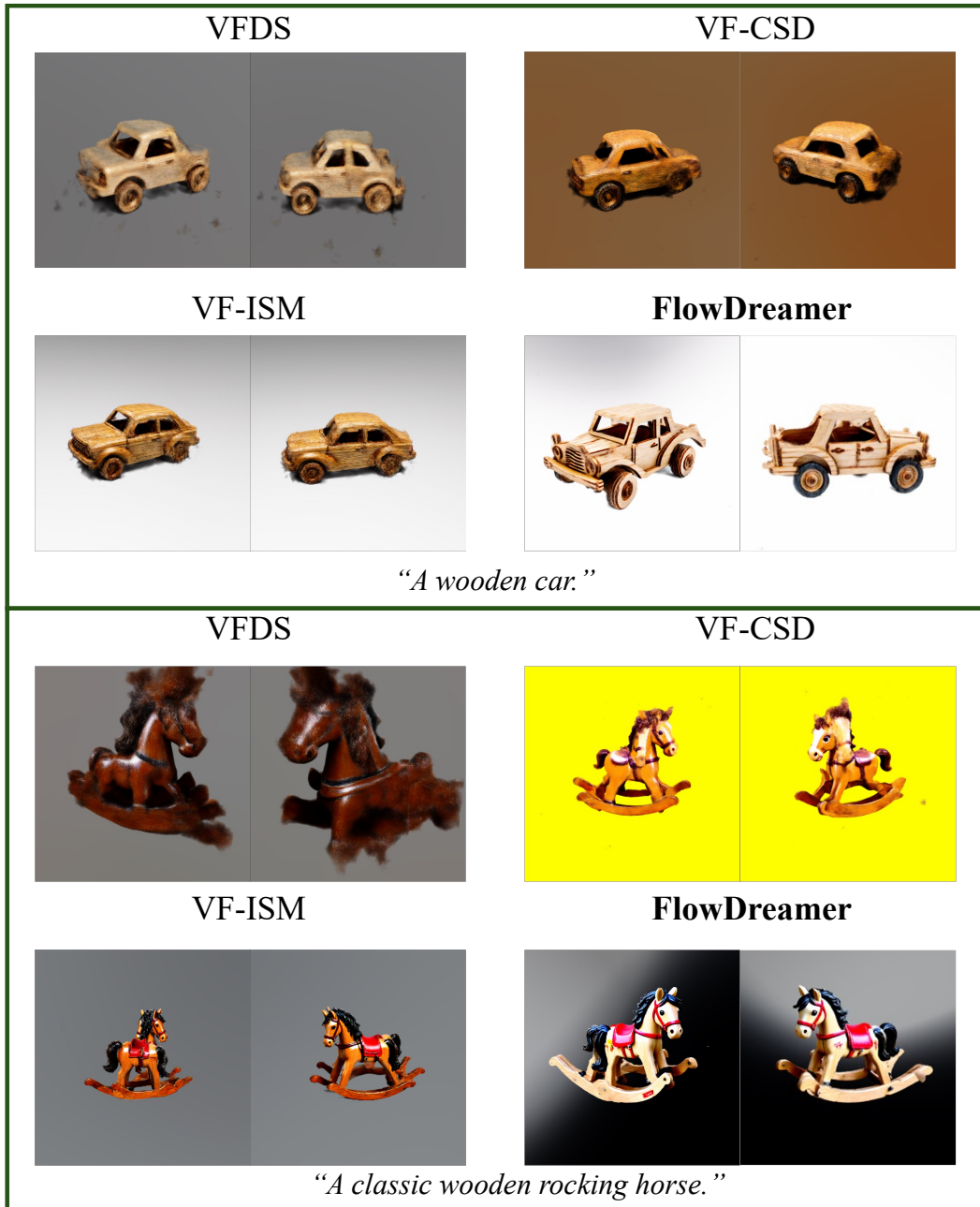


Figure 25: More qualitative comparison under a unified framework



1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403



Figure 26: More qualitative comparison under a unified framework

1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457

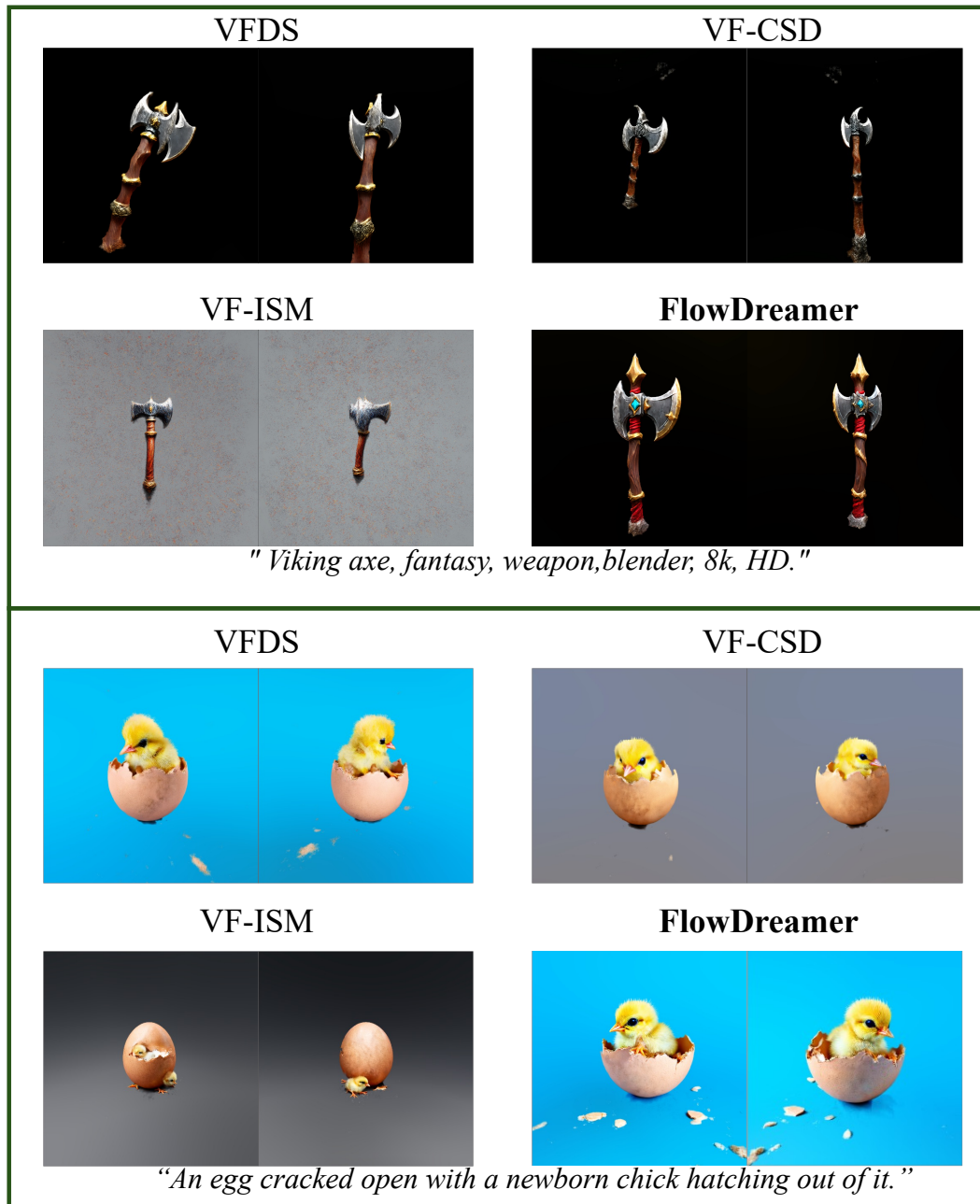


Figure 27: More qualitative comparison under a unified framework

1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

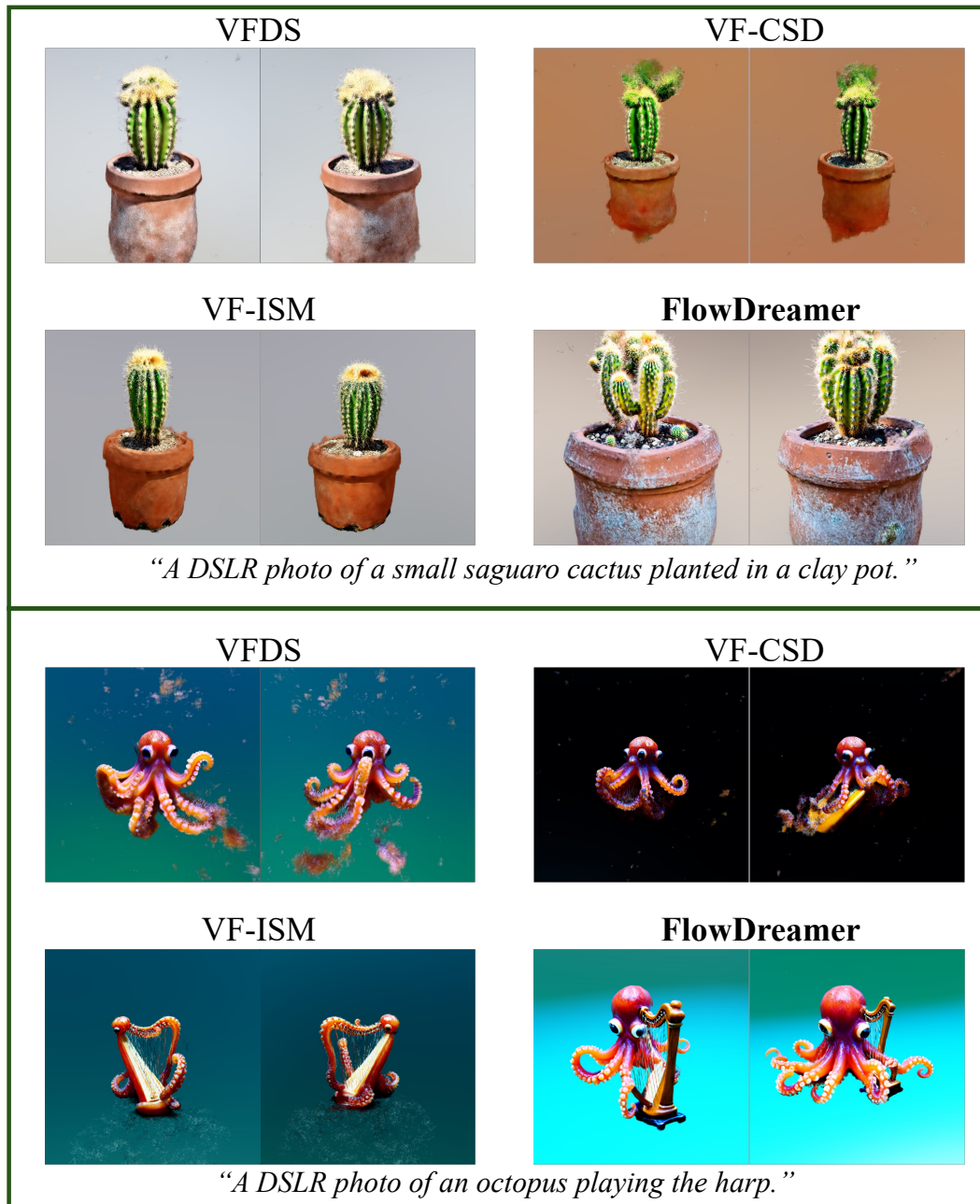


Figure 28: More qualitative comparison under a unified framework

---

## REFERENCES

- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Lucid-dreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.
- Alex Nichol, Heewoo Jun, Pratul Dharwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.
- Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022.
- Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.
- Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.
- Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9892–9902, 2024.
- Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023.
- Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023.