

702
703
704
705
706
707
708
709
710
711
712
A IMPLEMENTATION DETAILS704
705
A.1 BASELINES IMPLEMENTATION DETAILS706
707
708
709
710
711
As described in Section 4.1, we compare our method against the following baselines: TIME (Jean-
neret et al., 2024), Stable Diffusion (Rombach et al., 2022) with EF-DDPM inversion (Huberman-
Spiegelglas et al., 2024), once using class names as text prompts, and once using learned textual
embeddings of a group of each class’ images through Textual Inversion (Gal et al., 2022). Lastly
we also compare to Concept Sliders (Gandikota et al., 2023). We used the following third-party
implementation in this project:713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
• TIME: (Jeanneret et al., 2024). Instead of using their classifier’s labels to group the data
into two classes, we used the ground-truth labels. Additionally, to perform the evaluations,
we used our own ensemble classifiers. [official implementation](#).732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
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
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
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
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
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
10010
10011
10012
10013
10014
10015
10016
10017
10018
10019
10020
10021
10022
10023
10024
10025
10026
10027
10028
10029
10030
10031
10032
10033
10034
10035
10036
10037
10038
10039
10040
10041
10042
10043
10044
10045
10046
10047
10048
10049
10050
10051
10052
10053
10054
10055
10056
10057
10058
10059
10060
10061
10062
10063
10064
10065
10066
10067
10068
10069
10070
10071
10072
10073
10074
10075
10076
10077
10078
10079
10080
10081
10082
10083
10084
10085
10086
10087
10088
10089
10090
10091
10092
10093
10094
10095
10096
10097
10098
10099
100100
100101
100102
100103
100104
100105
100106
100107
100108
100109
100110
100111
100112
100113
100114
100115
100116
100117
100118
100119
100120
100121
100122
100123
100124
100125
100126
100127
100128
100129
100130
100131
100132
100133
100134
100135
100136
100137
100138
100139
100140
100141
100142
100143
100144
100145
100146
100147
100148
100149
100150
100151
100152
100153
100154
100155
100156
100157
100158
100159
100160
100161
100162
100163
100164
100165
100166
100167
100168
100169
100170
100171
100172
100173
100174
100175
100176
100177
100178
100179
100180
100181
100182
100183
100184
100185
100186
100187
100188
100189
100190
100191
100192
100193
100194
100195
100196
100197
100198
100199
100200
100201
100202
100203
100204
100205
100206
100207
100208
100209
100210
100211
100212
100213
100214
100215
100216
100217
100218
100219
100220
100221
100222
100223
100224
100225
100226
100227
100228
100229
100230
100231
100232
100233
100234
100235
100236
100237
100238
100239
100240
100241
100242
100243
100244
100245
100246
100247
100248
100249
100250
100251
100252
100253
100254
100255
100256
100257
100258
100259
100260
100261
100262
100263
100264
100265
100266
100267
100268
100269
100270
100271
100272
100273
100274
100275
100276
100277
100278
100279
100280
100281
100282
100283
100284
100285
100286
100287
100288
100289
100290
100291
100292
100293
100294
100295
100296
100297
100298
100299
100200
100201
100202
100203
100204
100205
100206
100207
100208
100209
1002010
1002011
1002012
1002013
1002014
1002015
1002016
1002017
1002018
1002019
10020100
10020101
10020102
10020103
10020104
10020105
10020106
10020107
10020108
10020109
10020110
10020111
10020112
10020113
10020114
10020115
10020116
10020117
10020118
10020119
100201100
100201110
100201120
100201130
100201140
100201150
100201160
100201170
100201180
100201190
100201101
100201111
100201121
100201131
100201141
100201151
100201161
100201171
100201181
100201191
100201102
100201112
100201122
100201132
100201142
100201152
100201162
100201172
100201182
100201192
100201103
100201113
100201123
100201133
100201143
100201153
100201163
100201173
100201183
100201193
100201104
100201114
100201124
100201134
100201144
100201154
100201164
100201174
100201184
100201194
100201105
100201115
100201125
100201135
100201145
100201155
100201165
100201175
100201185
100201195
100201106
100201116
100201126
100201136
100201146
100201156
100201166
100201176
100201186
100201196
100201107
100201117
100201127
100201137
100201147
100201157
100201167
100201177
100201187
100201197
100201108
100201118
100201128
100201138
100201148
100201158
100201168
100201178
100201188
100201198
100201109
100201119
100201129
100201139
100201149
100201159
100201169
100201179
100201189
100201199
100201110
100201120
100201130
100201140
100201150
100201160
100201170
100201180
100201190
1002011010
1002011110
1002011210
1002011310
1002011410
1002011510
1002011610
1002011710
1002011810
1002011910
1002011020
1002011120
1002011220
1002011320
1002011420
1002011520
1002011620
1002011720
1002011820
1002011920
1002011030
1002011130
1002011230
1002011330
1002011430
1002011530
1002011630
1002011730
1002011830
1002011930
1002011040
1002011140
1002011240
1002011340
1002011440
1002011540
1002011640
1002011740
1002011840
1002011940
1002011050
1002011150
1002011250
1002011350
1002011450
1002011550
1002011650
1002011750
1002011850
1002011950
1002011060
1002011160
1002011260
1002011360
1002011460
1002011560
1002011660
1002011760
1002011860
1002011960
1002011070
1002011170
1002011270
1002011370
1002011470
1002011570
1002011670
1002011770
1002011870
1002011970
1002011080
1002011180
1002011280
1002011380
1002011480
1002011580
1002011680
1002011780
1002011880
1002011980
1002011090
1002011190
1002011290
1002011390
1002011490
1002011590
1002011690
1002011790
1002011890
1002011990
1002011100
1002011200
1002011300
1002011400
1002011500
1002011600
1002011700
1002011800
1002011900
10020110100
10020111100
10020112100
10020113100
10020114100
10020115100
10020116100
10020117100
10020118100
10020119100
10020110200
10020111200
10020112200
10020113200
10020114200
10020115200
10020116200
10020117200
10020118200
10020119200
10020110300
10020111300
10020112300
10020113300
10020114300
10020115300
10020116300
10020117300
10020118300
10020119300
10020110400
10020111400
10020112400
10020113400
10020114400
10020115400
10020116400
10020117400
10020118400
10020119400
10020110500
10020111500
10020112500
10020113500
10020114500
10020115500
10020116500
10020117500
10020118500
10020119500
10020110600
10020111600
10020112600
10020113600
10020114600
10020115600
10020116600
10020117600
10020118600
10020119600
10020110700
10020111700
10020112700
10020113700
10020114700
10020115700
10020116700
10020117700
10020118700
10020119700
10020110800
10020111800
10020112800
10020113800
10020114800
10020115800
10020116800
10020117800
10020118800
10020119800
10020110900
10020111900
10020112900
10020113900
10020114900
10020115900
10020116900
10020117900
10020118900
10020119900
10020111000
10020112000
10020113000
10020114000
10020115000
10020116000
10020117000
10020118000
10020119000
100201101000
100201111000
100201121000
100201131000
100201141000
100201151000
100201161000
100201171000
100201181000
100201191000
100201102000
100201112000
100201122000
100201132000
100201142000
100201152000
100201162000
100201172000
100201182000
100201192000
100201103000
100201113000
100201123000
100201133000
100201143000
100201153000
100201163000
100201173000
100201183000
100201193000
100201104000
100201114000
100201124000
100201134000
100201144000
100201154000
100201164000
100201174000
100201184000
100201194000
100201105000
100201115000
100201125000
100201135000
100201145000
100201155000
100201165000
100201175000
100201185000
100201195000
100201106000
100201116000
100201126000
100201136000
100201146000
100201156000
100201166000
100201176000
100201186000
100201196000
100201107000
100201117000
100201127000
100201137000
100201147000
100201157000
100201167000
100201177000
100201187000
100201197000
100201108000
100201118000
100201128000
100201138000
100201148000
100201158000
100201168000
100201178000
100201188000
100201198000
100201109000
100201119000
100201129000
100201139000
100201149000
100201159000
100201169000
100201179000
100201189000
100201199000
100201110000
100201120000
100201130000
100201140000
100201150000
100201160000
100201170000
100201180000
100201190000
1002011010000
1002011110000
1002011210000
1002011310000
1002011410000
1002011510000
1002011610000
1002011710000
1002011810000
1002011910000
1002011020000
1002011120000
1002011220000
1002011320000
1002011420000
1002011520000
1002011620000
1002011720000
1002011820000
1002011920000
1002011030000
1002011130000
1002011230000
1002011330000
1002011430000
1002011530000
1002011630000
1002011730000
1002011830000
1002011930000
1002011040000
1002011140000
1002011240000
1002011340000
1002011440000
1002011540000
1002011640000
1002011740000
1002011840000
1002011940000
1002011050000
1002011150000
1002011250000
1002011350000
1002011450000
1002011550000
1002011650000
1002011750000
1002011850000
1002011950000
1002011060000
1002011160000
1002011260000
1002011360000
1002011460000
1002011560000
1002011660000
1002011760000
1002011860000
1002011960000
1002011070000
1002011170000
1002011270000
1002011370000
1002011470000
1002011570000
1002011670000
1002011770000
1002011870000
1002011970000
1002011080000
1002011180000
1002011280000
1002011380000
1002011480000
1002011580000
1002011680000
1002011780000
1002011880000
1002011980000
1002011090000
1002011190000
1002011290000
1002011390000
1002011490000
1002011590000
1002011690000
1002011790000
1002011890000
1002011990000
1002011100000
1002011200000
1002011300000
1002011400000
1002011500000
1002011600000
1002011700000
1002011800000
1002011900000
10020110100000
10020111100000
10020112100000
10020113100000
10020114100000
10020115100000
10020116100000
10020117100000
10020118100000
10020119100000
10020110200000
10020111200000
10020112200000
10020113200000

756
757

A.3 USER STUDY DETAILS

758
759
760
761
762
763
764

Group 1 studied a folder of unpaired images for 3 minutes. This folder contained images of both classes, with the label of the class written on top of the image. Group 2 studied the unpaired image folder for 1 minute, the counterfactuals generated by the best baseline from Class 0 to Class 1 for another minute, and finally the counterfactuals generated by the best baseline from Class 1 to class 0 for a minute. Group 3 followed the same protocol as Group 2, except with our counterfactuals instead. For both Butterfly and Black hole, the best baseline was TI + DDPM-EF, according to LPIPS. We only showed counterfactuals where the class flipped.

765
766
767
768

The participants of the user studies were undergraduates and graduates who volunteered in exchange for baked goods. The supplementary material contains videos of the study material for Group 3, for both the black hole and the butterfly dataset. No user had any prior knowledge about either of the datasets before studying the material and taking the test.

769
770

A.4 TRAINING DETAILS

771
772
773
774
775
776
777

We utilize the diffusion decoder from (Shakhmatov et al., 2023) and optionally fine-tune cross-attention weights using LoRA (Hu et al., 2021) on either subsets or the full dataset. For fine-tuning, we set the LoRA rank to 4, the LoRA scaling factor to $\alpha = 8$ and use a base learning rate of 0.003. Fine-tuning was conducted on a single NVIDIA A100 GPU, although the implementation supports multi-GPU training as well. We train for 4 epochs, and select the checkpoint with the optimal balance between LPIPS and classifier accuracy on generated counterfactuals as our final model for each dataset (or data subset).

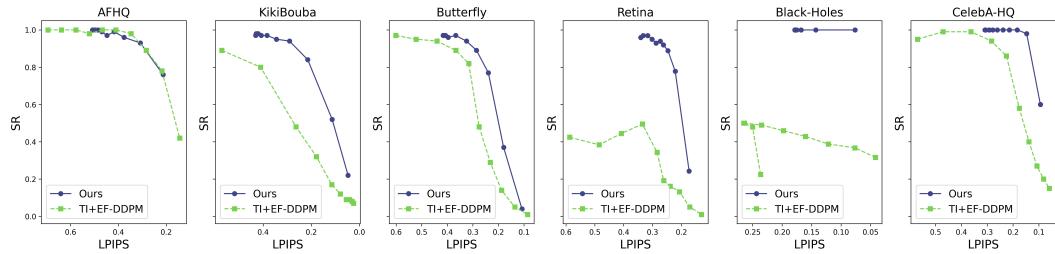
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

810
811 A.5 DATASETS

812 • AFHQ (Choi et al., 2020): [official implementation](#). Creative Commons BY-NC 4.0.
 813 • KikiBouba (Alper & Averbuch-Elor, 2024): [official implementation](#). MIT License.
 814 • Retina (Kermany et al., 2018): [official implementation](#). Creative Commons BY-NC 4.0.
 815 • Butterfly (Van Horn et al., 2018): [official implementation](#). MIT License.
 816 • CelebA-HQ (Lee et al., 2020): [official implementation](#). Creative Commons BY-NC 4.0.
 817

818
819 B ADDITIONAL EXPERIMENTS
820821 B.1 SR vs. LPIPS CURVES
822

823 In Figure 8 we plot the Success Ratio vs. LPIPS curves for our method compared to the best baseline
 824 - TI + EF-DDPM, rather than choosing a single set of parameters which is required to report Table 2.
 825 Since both use the same inversion technique, we create these curves by varying the T_{skip} parameter.
 826 A higher AUC generally indicates a better tradeoff between classifier flip-rate and similarity to
 827 input images for each dataset. Our method outperforms the best baseline across all datasets while
 828 achieving comparable performance on AFHQ.
 829



830 Figure 8: **Success Ratio (SR) vs. LPIPS curves.** As discussed in Section B.1, we fix the guidance
 831 scales for each method, and the manipulation scale ω for ours according to the implementation
 832 details in Section 3.2. We then vary T_{skip} in increments of 0.1 within the range [0.0, 0.9], where
 833 T_{skip} represents the percentage of timesteps skipped relative to the total denoising steps.
 834

835 B.2 RESULTS w/o DOMAIN TUNING
836

837 As discussed in Section 3.2, we begin by fine-tuning a LoRA adapter (Hu et al., 2021), applied to
 838 the cross-attention and their linear projection weights, using a simple loss at random timesteps t , i.e:
 839

$$\mathcal{L}_{simple} = \mathbb{E}_{x, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2]$$

840 where the prompts c are derived from image embeddings of training set examples. To obtain the
 841 best weights, we log SR and LPIPS scores on a small validation set at the end of each epoch. In
 842 certain datasets, fine-tuning has minimal to no impact on the overall results. This suggests that, in
 843 some cases, the prior learned by the pre-trained diffusion model is sufficiently strong to produce
 844 meaningful edits when conditioned on a manipulated CLIP image embedding. The results of this
 845 analysis are presented in Table 4, where we observe that domain tuning plays a crucial role in
 846 datasets like Retina and Butterfly, while having a lesser effect on others.
 847

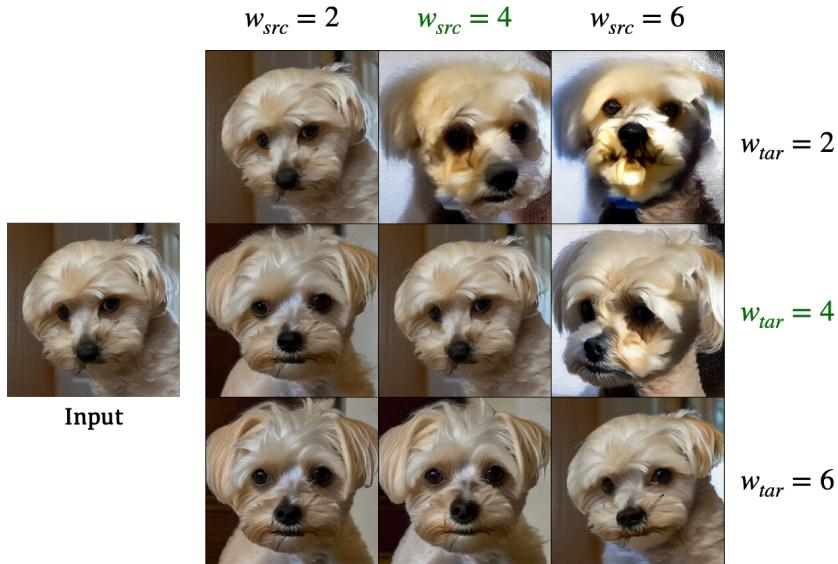
848 B.3 PERFECT INVERSION
849

850 Perfect reconstruction can be achieved when the same conditioning prompt is used during both inversion
 851 and sampling. In this case, we hope that the original image is fully reconstructed. However,
 852 DDIM (Song et al., 2022) introduces small errors at each timestep, making exact reconstruction chal-
 853 lenging, especially with a limited number of timesteps or within the classifier-free guidance frame-
 854 work (Ho & Salimans, 2022). Recent works (Huberman-Spiegelglas et al., 2024; Wu & la Torre,
 855 2022; Brack et al., 2024) focus on non-deterministic DDPM inversion and have demonstrated per-
 856 fect image reconstruction when applied to Stable Diffusion (Rombach et al., 2022). Since we are
 857

864
 865
 866
 867
 868
 869
 870
Table 4: Results without Domain Tuning. We evaluate our method without fine-tuning on each
 871 dataset, measuring performance using Success Ratio (SR) and perceptual distance (LPIPS). Com-
 872 pared to Table 2, we observe substantial improvements in both SR and similarity for the Retina and
 873 Butterfly datasets, as well as noticeable gains in reconstruction for Black-Holes and KikiBouba.
 874 However, the impact is minimal on the most natural-image datasets, AFHQ and CelebA-HQ. This
 875 suggests that the prior learned by the pre-trained diffusion model is strong enough to generate mean-
 876 ingful edits for these datasets without additional fine-tuning.

Method	AFHQ		KikiBouba		Retina		Black-Holes		Butterfly		Celeba-HQ-Smile	
	SR ↑	LPIPS ↓	SR ↑	LPIPS ↓	SR ↑	LPIPS ↓	SR ↑	LPIPS ↓	SR ↑	LPIPS ↓	SR ↑	LPIPS ↓
Ours	1.0	0.249	0.98	0.2014	0.515	0.454	0.980	0.119	0.31	0.344	1.0	0.123

877 using an image-conditioned diffusion decoder from the Kandinsky model family (Razhigaev et al.,
 878 2023), we first explore the choice of guidance scales required to achieve perfect reconstruction while
 879 using CFG (Ho & Salimans, 2022) in both inversion and generation. While perfect reconstruction
 880 does not necessarily guarantee a useful editing space, poor reconstruction from the start is likely
 881 to cause significant deviations from the source image, which is an undesirable outcome when gen-
 882 erating counterfactuals. Figure 9 illustrates this effect, showing that using equal guidance terms
 883 in inversion and sampling results in good reconstruction, which starts to degrade when inversion
 884 guidance scale, w_{src} , and target guidance scale, w_{tar} , are larger than 4.



904
 905
Figure 9: Perfect Inversion. We choose our inversion and sampling guidance scales by first recon-
 906 structing the original image with CFG. Then, we use these guidance scales for steering.

B.4 VARYING DATASET SIZE RESULTS

907
 908
 909
 910 As described in Section 4.4, and in Figure 4a, we demonstrate the effect of varying the number of
 911 images we have access to for applying DIFFusion. In this section, we show examples of generated
 912 counterfactuals per number of images in access, N , as shown in Figures 10,11,12,13,14. We show
 913 the grids with increasing number of images so long as the results continue to improve.

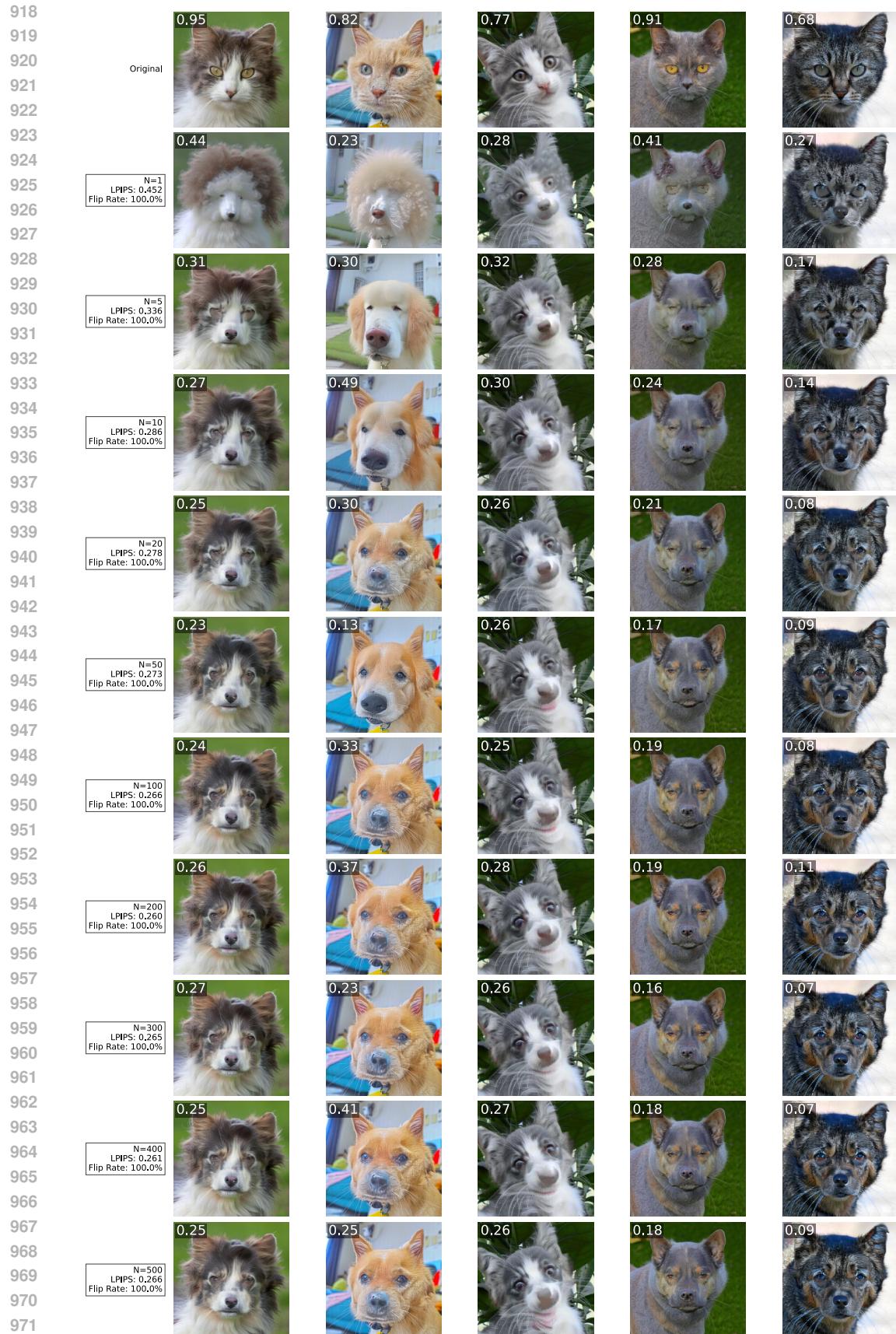


Figure 10: Varying Number of Images for AFHQ (Choi et al., 2020).

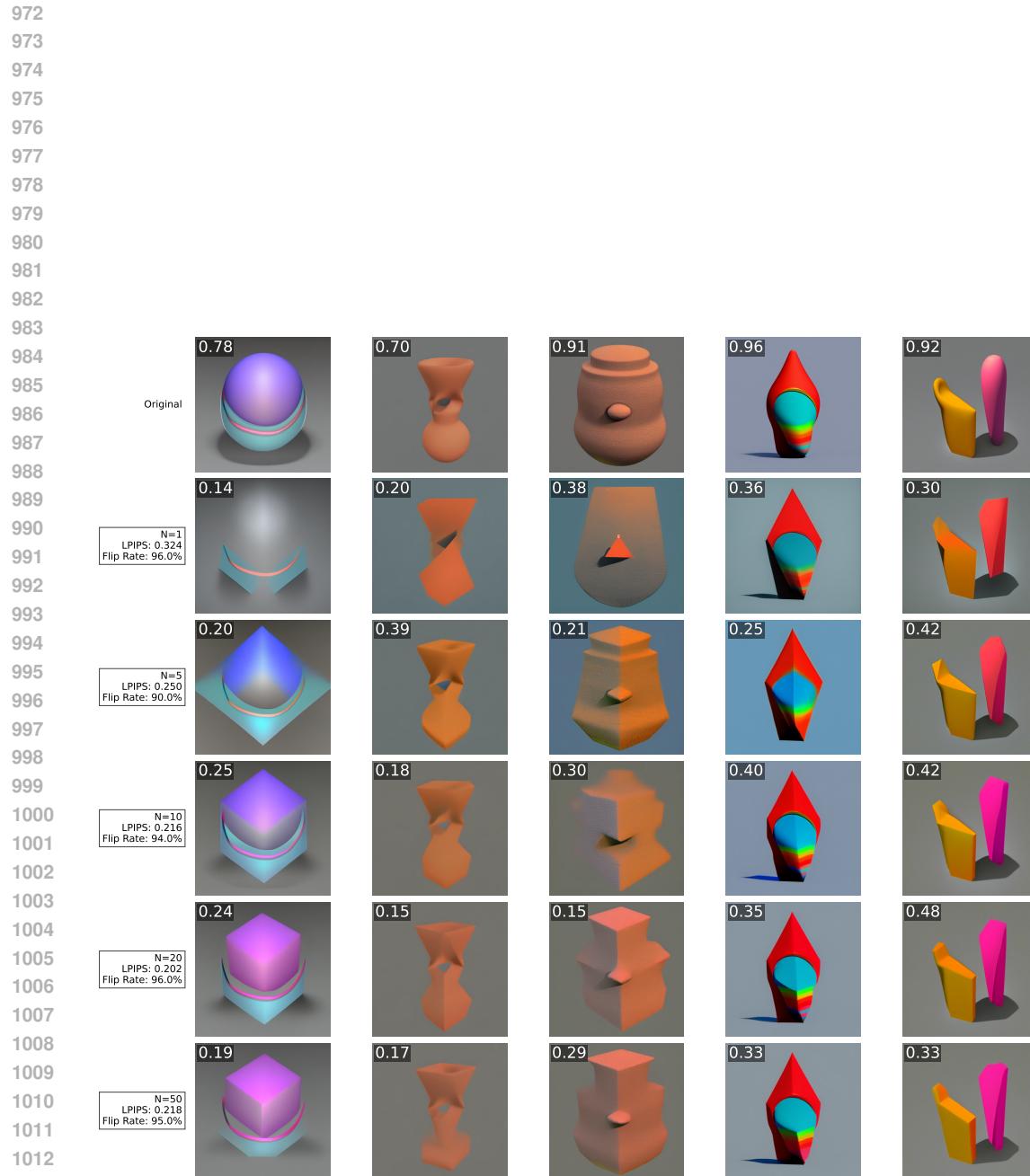


Figure 11: Varying Number of Images for KikiBouba (Alper & Averbuch-Elor, 2024).



Figure 12: Varying Number of Images for Butterfly (Van Horn et al., 2018).

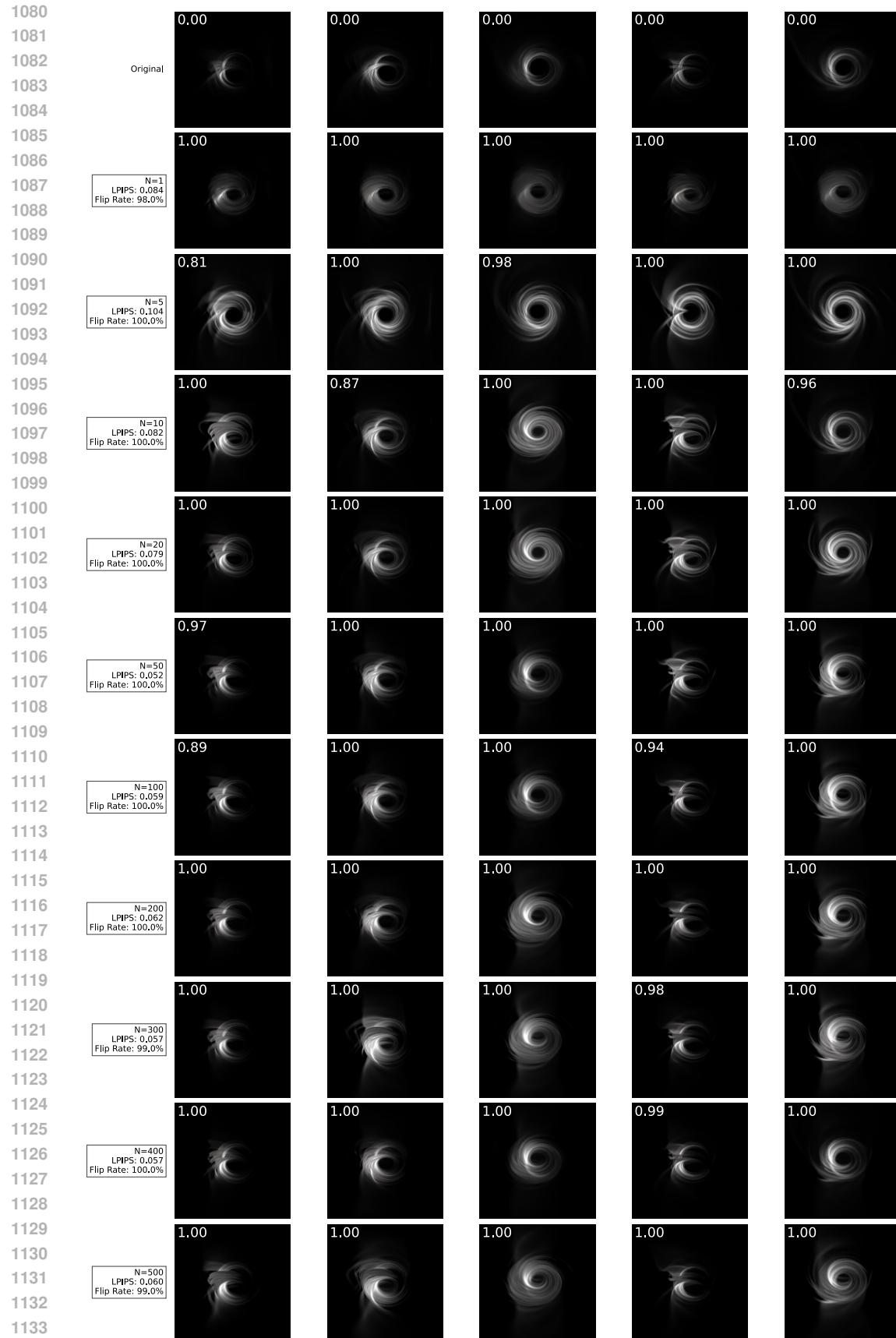


Figure 13: Varying Number of Images for Black-Holes.

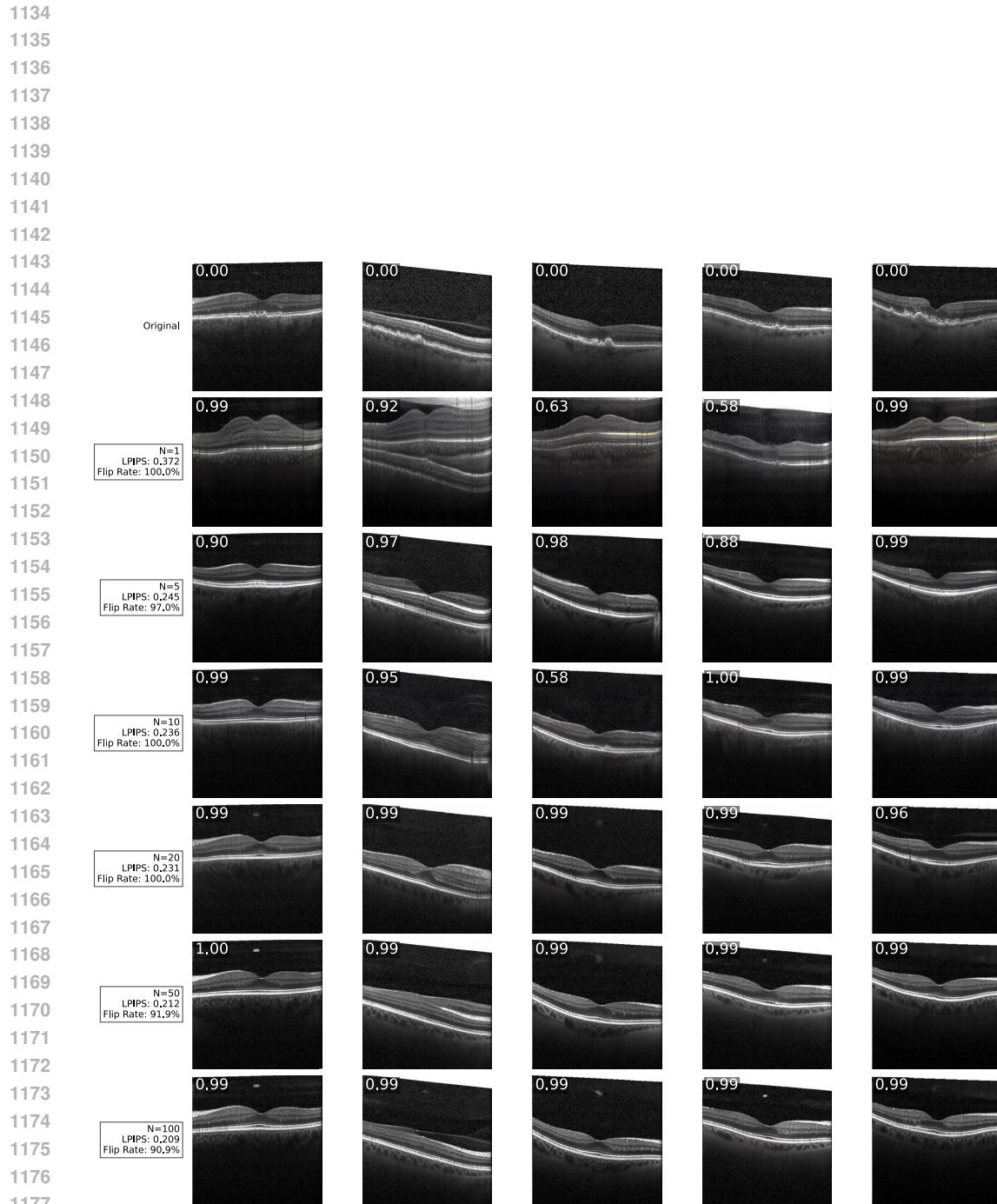


Figure 14: Varying Number of Images for Retina (Kermany et al., 2018).