

A APPENDIX

A.1 IMPLEMENTATION DETAILS OF BASELINES

Textual Inversion. For this baseline method, we train the model for 2,000 iterations with a constant learning rate $2e-3$. We use the batch size of 4 to train the method. Other than the concept token embeddings, no parameters are updated during the training. We set the batch size for MLM as 25.

XTL. For this baseline method, we train the model for 1,500 iterations with a constant learning rate of $2e-3$. We use the batch size of 4 to train the method. Following the original method, a set of multiple concept embeddings is utilized to be aligned with the same concept image. We set the batch size for MLM as 12 due to the memory limit.

DreamBooth. For this baseline method, we train the model for 1,000 iterations with a constant learning rate of $1e-6$. We use the batch size of 2 to train the method. Following the original method, the prior preservation loss is adopted during the training. For this, we generate a set of 200 images by prompting with “a picture of [SUBJECT CLASS]”, by denoting the general class of the concept in the prompt. We update the parameters of both the CLIP text encoder and the diffusion U-Net. We set the batch size for MLM as 25.

CustomDiffusion. For this baseline method, we train the model for 5,00 iterations with a constant learning rate of $4e-5$. We use the batch size of 4 to train the method. Following the original method, we adopt prior preservation with generated images. During training only the Key/Value projection layers of the diffusion U-Net are updated during training. We set the batch size for MLM as 25.

A.2 DETAILS OF TEXT PROMPT SET CONSTRUCTION

To generate a contextually diverse prompt set with minimal human intervention, we utilize a large pretrained language model (LLM) OpenAI (2023). Based on whether the personal concept is classified as living or nonliving, we predefined context categories and query the LLM to generate relevant elements for each category. The predefined categories for the living personal concepts are as below,

1. **Human Interactive Prompts:** A set of prompts that involves diverse interaction between different human subjects (e.g., “*Albert Einstein is watching TV with [V]*”).
2. **Relative Position Prompts:** A set of prompts that involves different positioning words and different objects (e.g., “*a picture of [V] next to a red vase*”).
3. **Background Prompts:** A set of prompts that describes a scene with different backgrounds (e.g., “*a picture of [V] with Eiffel Tower in the background*”).
4. **Image Style Prompts:** A set of prompts that describe image style (e.g., “*a picture of [V] in Pop Art style*”).
5. **Attributes Changing Prompts:** A set of prompts that describe the target concept with different visual attributes (e.g., “*a picture of [V] in blue sailor outfit*”).

Similarly, for non-living objects, we construct a set of prompts in five different types of contexts,

1. **Human Interactive Prompts:** A set of prompts that involves diverse interaction between different human subjects (e.g., “*Albert Einstein is watching TV with [V]*”).
2. **Relative Position Prompts:** A set of prompts that involves different positioning words and different objects (e.g., “*a picture of [V] next to a red vase*”).
3. **Background Prompts:** A set of prompts that describes a scene with different backgrounds (e.g., “*a picture of [V] with Eiffel Tower in the background*”).
4. **Image Style Prompts:** A set of prompts that describe image style (e.g., “*a picture of [V] in Pop Art style*”).
5. **Attributes Changing Prompts:** A set of prompts that describe the target concept with different visual attributes (e.g., “*a picture of [V] in blue sailor outfit*”).

756 A.3 IMPLEMENTATION DETAILS OF CONTEXTUALIZER

757
758 Contextualizer constitutes four blocks of a self-attention layer and a feed-forward layer, followed
759 by a layer normalization layer, where each block learns the residuals of the input with the residual
760 connection. To train the contextualizer, we use a merged set of COCO caption dataset Chen et al.
761 (2015) and the prompt set that we constructed. For the manual prompt set, we replace the personal
762 concept token with the personal concept token to corresponding prior concept token. During training
763 we set the ratio of batch of the two prompt set to be 70 to 30. The contextualizer is pretrained for
764 100K iterations with a learning rate of $1e-4$, and batch size 150. We use the AdamW optimizer
765 Loshchilov (2017).

766 A.4 JUSTIFICATION - SEMANTIC ENHANCEMENT IN TEXTUAL SPACE

767 The proof of Proposition 1.

768 *Proof.* Given an attention map \mathbf{A} , with $\sum_j \mathbf{A}[i, j] = 1$, $\mathbf{A}[i, j] \geq 0$, and the value matrix \mathbf{V} , the
769 output of the attention layer is,

$$770 c_i = \sum_{j=1}^N \mathbf{A}[i, j] \mathbf{V}[j, :]. \quad (1)$$

771 The concept token at index j_* has the highest attention value, *i.e.*, $\mathbf{A}[i, j_*] \gg \mathbf{A}[i, j], \forall j \neq j_*$. We
772 have,

$$773 c_i = \sum_{j=1}^N \mathbf{A}[i, j] \mathbf{V}[j, :] = \sum_{j=1, j \neq j_*}^N \mathbf{A}[i, j] \mathbf{V}[j, :] + \mathbf{A}[i, j_*] \mathbf{V}[j_*, :] \approx \mathbf{A}[i, j_*] \mathbf{V}[j_*, :] \approx \mathbf{V}[j_*, :]. \quad (2)$$

774 The L2 norm between the text embeddings of the concept token c_{i_*} and context tokens c_i is,

$$775 \begin{aligned} & \|c_i - c_{i_*}\|_2 \\ &= \left\| \sum_{j=1, j \neq j_*}^N (\mathbf{A}[i, j] - \mathbf{A}[i_*, j]) \mathbf{V}[j, :] + (\mathbf{A}[i, j_*] - \mathbf{A}[i_*, j_*]) \mathbf{V}[j_*, :]\right\|_2 \\ &\leq \left\| \sum_{j=1, j \neq j_*}^N (\mathbf{A}[i, j] - \mathbf{A}[i_*, j]) \mathbf{V}[j, :]\right\|_2 + \left\| (\mathbf{A}[i, j_*] - \mathbf{A}[i_*, j_*]) \mathbf{V}[j_*, :]\right\|_2 \\ &\leq \sum_{j=1, j \neq j_*}^N \|\mathbf{A}[i, j] - \mathbf{A}[i_*, j]\|_2 \|\mathbf{V}[j, :]\|_2 + \|\mathbf{A}[i, j_*] - \mathbf{A}[i_*, j_*]\|_2 \|\mathbf{V}[j_*, :]\|_2. \end{aligned} \quad (3)$$

776 Suppose $\mathbf{A}[i, j_*] = 1 - \delta_{ij_*}$ and $\mathbf{A}[i_*, j_*] = 1 - \delta_{i_*j_*}$, where $\mathbf{0} \leq \delta_{ij_*} < \delta$ and $\mathbf{0} \leq \delta_{i_*j_*} < \delta$.
777 $\mathbf{A}[i, j] = \delta_{ij}, \forall j \neq j_*$, $\mathbf{A}[i_*, j] = \delta_{i_*j}, \forall j \neq j_*$, $0 \leq \delta_{ij} < \delta$ and $0 \leq \delta_{i_*j} < \delta$, where δ is a small
778 value. We have $\|\delta_{ij} - \delta_{i_*j}\|_2 < \delta$ and $\|\delta_{i_*j_*} - \delta_{ij_*}\|_2 < \delta$. Thus,

$$779 \begin{aligned} & \|c_i - c_{i_*}\|_2 \\ &\leq \sum_{j=1, j \neq j_*}^N \|\delta_{ij} - \delta_{i_*j}\|_2 \|\mathbf{V}[j, :]\|_2 + \|\delta_{i_*j_*} - \delta_{ij_*}\|_2 \|\mathbf{V}[j_*, :]\|_2 \\ &\leq \delta \sum_{j=1, j \neq j_*}^N \|\mathbf{V}[j, :]\|_2 + \delta \|\mathbf{V}[j_*, :]\|_2. \end{aligned} \quad (4)$$

780 Since $\|\mathbf{V}[j, :]\|_2$ is bounded, we have,

$$781 \|c_i - c_{i_*}\|_2 \leq \delta_{\mathbf{V}}, \quad (5)$$

782 where $\delta_{\mathbf{V}} = \delta \sum_{j=1}^N \|\mathbf{V}[j, :]\|_2$. □

783 The proof of Proposition 2.

810 *Proof.* Suppose $\|c_b - \hat{c}_b\|_2$ is a small value, using the Taylor series, we have,

$$\begin{aligned} 811 \mathcal{L}_{\text{MLM}}(c_b) &= \mathcal{L}_{\text{MLM}}(\hat{c}_b) + (c_b - \hat{c}_b)^T \text{grad}(\mathcal{L}_{\text{MLM}}(\hat{c}_b)) + \mathcal{O}(c_b - \hat{c}_b) \\ 812 &\approx \mathcal{L}_{\text{MLM}}(\hat{c}_b) + (c_b - \hat{c}_b)^T \text{grad}(\mathcal{L}_{\text{MLM}}(\hat{c}_b)), \end{aligned} \quad (6)$$

813 where $\text{grad}(\cdot)$ is the first-order derivative. Using Cauchy-Schwartz inequality, we have,

$$814 (c_b - \hat{c}_b)^T \text{grad}(\mathcal{L}_{\text{MLM}}(\hat{c}_b)) \leq \|c_b - \hat{c}_b\|_2 \cdot \|\text{grad}(\mathcal{L}_{\text{MLM}}(\hat{c}_b))\|_2. \quad (7)$$

815 Since \hat{c}_b is near the optimal value, which is achieved by optimizing the contextualizer, we have $\text{grad}(\mathcal{L}_{\text{MLM}}(\hat{c}_b)) \leq \delta_g$, where δ_g is a small value. Therefore, we have

$$816 \mathcal{L}_{\text{MLM}}(c_b) - \mathcal{L}_{\text{MLM}}(\hat{c}_b) \leq \delta_g \|c_b - \hat{c}_b\|_2. \quad (8)$$

817 \square

818 A.5 JUSTIFICATION - SEMANTIC ENHANCEMENT IN IMAGE SPACE

819 The proof of Proposition 4.

820 *Proof.* The image embedding \mathbf{z} and text embedding \mathbf{C} are projected as $\mathbf{Q}_{\mathcal{I}} = \mathbf{z}\mathbf{W}_Q$, $\mathbf{K}_{\mathcal{T}} = \mathbf{C}\mathbf{W}_K$. For text embeddings at indices i and j , we have,

$$821 \mathbf{K}_{\mathcal{T}}[i, :] = c_i \mathbf{W}_K \quad (9)$$

$$822 \mathbf{K}_{\mathcal{T}}[j, :] = c_j \mathbf{W}_K. \quad (10)$$

823 The relation map is $\mathbf{M} = \mathbf{Q}_{\mathcal{I}}\mathbf{K}_{\mathcal{T}}^T$, and $\mathbf{M}[:, i] = \mathbf{Q}_{\mathcal{I}}\mathbf{K}_{\mathcal{T}}[i, :]$, $\mathbf{M}[:, j] = \mathbf{Q}_{\mathcal{I}}\mathbf{K}_{\mathcal{T}}[j, :]$. Thus,

$$\begin{aligned} 824 \|\mathbf{M}[:, i] - \mathbf{M}[:, j]\|_2 &= \|\mathbf{Q}_{\mathcal{I}}(\mathbf{K}_{\mathcal{T}}[i, :] - \mathbf{K}_{\mathcal{T}}[j, :])\|_2 \\ 825 &\leq \|\mathbf{Q}_{\mathcal{I}}\|_F \|\mathbf{K}_{\mathcal{T}}[i, :] - \mathbf{K}_{\mathcal{T}}[j, :]\|_2 \\ 826 &= \|\mathbf{Q}_{\mathcal{I}}\|_F \|(c_i - c_j)\mathbf{W}_K\|_2 \\ 827 &\leq \|\mathbf{Q}_{\mathcal{I}}\|_F \|\mathbf{W}_K\|_F \|c_i - c_j\|_2 \\ 828 &= \alpha \|c_i - c_j\|_2, \end{aligned} \quad (11)$$

829 where $\|\cdot\|_F$ is Frobenius norm, and $\alpha = \|\mathbf{Q}_{\mathcal{I}}\|_F \|\mathbf{W}_K\|_F$. \square

830 A.6 ADDITIONAL QUALITATIVE EXAMPLES

831 We provide additional qualitative results of our approach combined with each baseline method, TI (Gal et al., 2022), XTI (Voynov et al., 2023), DB (Ruiz et al., 2023) and CD (Kumari et al., 2023). We provide two types of generation results, living or non-living objects (Figures 1, 2, 3, 4, 5, 6, 7, 8)

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



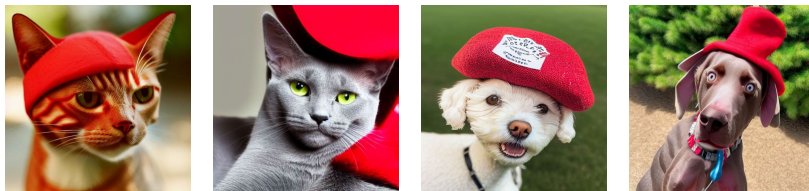
Input Images



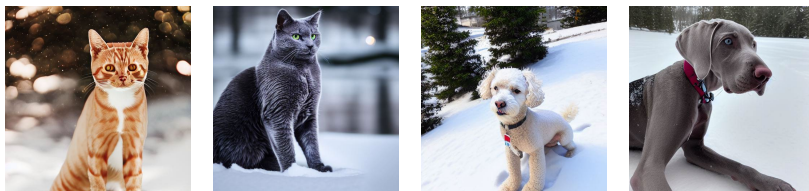
In the jungle



On a cobble stone street



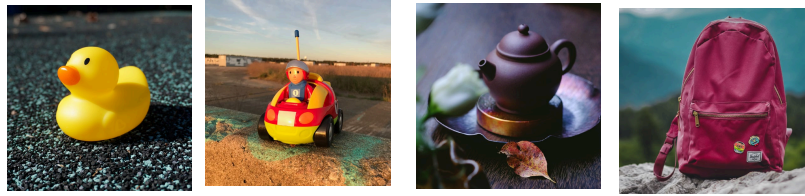
wearing a red hat



In the snow

Figure 1: Additional Qualitative Result of TI - Living Objects.

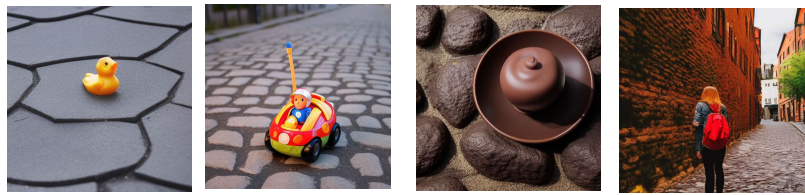
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971



Input Images



With Eiffel Tower at the Background



On a cobble stone



On top of a dirt road



Mountain in the background

Figure 2: Additional Qualitative Result of TI - Non-living Objects.

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



Input Images



In a firefighter outfit



Wearing a red hat



On top of pink fabric



In a purple wizard outfit

Figure 3: Additional Qualitative Result of XTI - Living Objects.

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



Input Images



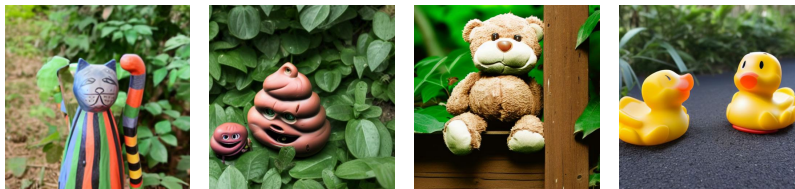
With Eiffel Tower in the background



In the snow



On the beach



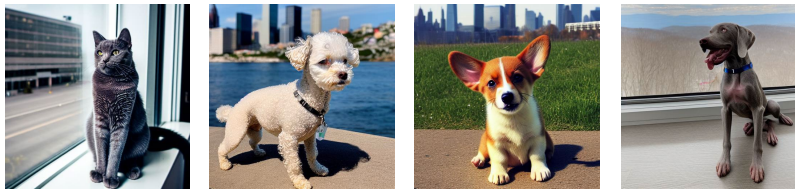
In the jungle

Figure 4: Additional Qualitative Result of XTI - Non-living Objects.

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



Input Images



With a city in the background



Wearing a red hat



In a police outfit



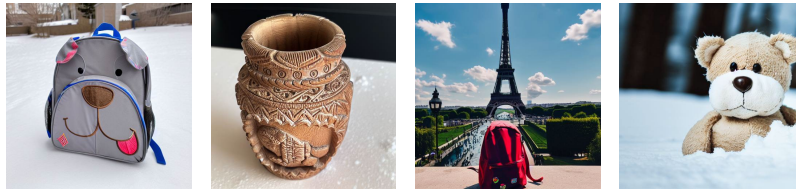
On top of a wooden floor

Figure 5: Additional Qualitative Result of DB - Living Objects.

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



Input Images



With Eiffel Tower in the background



With a wheat field in the background



Green grass with sunflowers around it



on a cobble stone street

Figure 6: Additional Qualitative Result of DB - Non-living Objects.

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



Input Images



On top of a purple rug in a forest



Wearing a rainbow scarf



Wearing a black top hat and a monocle



On a cobblestone street

Figure 7: Additional Qualitative Result of CD - Living Objects.

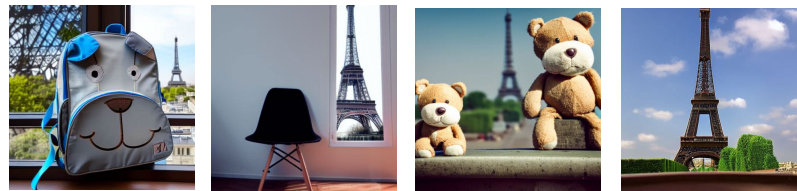
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295



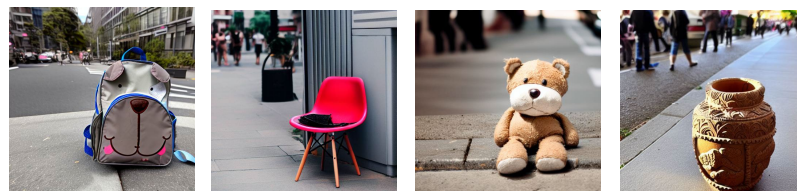
Input Images



With a city in the background



With Eiffel Tower in the background



On top of the sidewalk in a crowded street



In the snow

Figure 8: Additional Qualitative Result of CD - Non-living Objects.