

Supplementary Materials: DQG: Database Question Generation for Exact Text-based Image Retrieval

Anonymous Authors

1 COMPUTATIONAL COMPLEXITY

In the training phase, using two NVIDIA GeForce RTX 2080 Ti GPUs and an Intel Core i9-10980XE CPU, one hour and six hours were required for loss convergence in MSCOCO and Visual Genome datasets, respectively. Also, in the test phase, 19.7 and 37.1 seconds were required for conducting the re-ranking in MSCOCO and Visual Genome datasets, respectively.

2 CODE OF OUR METHOD

In this supplemental material, we provide the code of our paper. The detail instruction of our code is described in “ReadMe.md” of “CodeForReview” folder.

3 COMPARISON WITH THE OTHER METHODS

With the limitation of the page lengths, we only provide the comparison with the representative state-of-the-art methods in the main paper. Therefore, in this supplementary material, we provide comparisons with the other state-of-the-art methods. Specifically, UVS [5], OE [9], DPC [11], VSEPP [1], RRF-Net [7], CMPM [10], SISM [3], GXN [2], PVSE [8], SAN [4], and VSRN [6] are used as the comparative methods.

Experimental results are shown in Tables 1, 2, 3, 4, respectively. In each experiment, we also reach the same conclusions in the main paper.

Table 1: Experimental results for R@k, mean rank, and median rank using MSCOCO dataset.

	R@1	R@10	Mean	Med
UVS	0.154	0.504	70.464	10
UVS+Ours w/o opt	0.278	0.651	51.114	7
UVS+Ours w/o $L^{\mathcal{D}}$	0.310	0.722	43.063	6
UVS+Ours	0.329	0.734	40.596	5
OE	0.187	0.576	64.592	9
OE+Ours w/o opt	0.298	0.684	56.661	7
OE+Ours w/o $L^{\mathcal{D}}$	0.321	0.734	50.660	7
OE+Ours	0.341	0.755	48.059	5
DPC	0.288	0.703	53.540	9
DPC+Ours w/o opt	0.338	0.781	40.006	6
DPC+Ours w/o $L^{\mathcal{D}}$	0.355	0.832	23.102	5
DPC+Ours	0.398	0.838	18.592	4
VSEPP	0.297	0.724	52.440	9
VSEPP+Ours w/o opt	0.354	0.791	38.994	7
VSEPP+Ours w/o $L^{\mathcal{D}}$	0.374	0.822	24.510	6
VSEPP+Ours	0.401	0.822	17.662	5
RRF-Net	0.294	0.709	47.300	9
RRF-Net+Ours w/o opt	0.340	0.773	34.555	6
RRF-Net+Ours w/o $L^{\mathcal{D}}$	0.403	0.823	17.663	4
RRF-Net+Ours	0.417	0.840	15.511	4
CMPM	0.303	0.752	30.223	7
CMPM+Ours w/o opt	0.349	0.804	22.533	5
CMPM+Ours w/o $L^{\mathcal{D}}$	0.426	0.844	17.544	4
CMPM+Ours	0.431	0.864	12.441	2
SISM	0.301	0.755	28.480	5
SISM+Ours w/o opt	0.354	0.819	20.115	3
SISM+Ours w/o $L^{\mathcal{D}}$	0.420	0.833	14.682	2
SISM+Ours	0.433	0.849	12.011	2
GXN	0.307	0.746	26.731	3
GXN+Ours w/o opt	0.374	0.792	19.931	3
GXN+Ours w/o $L^{\mathcal{D}}$	0.432	0.827	11.042	2
GXN+Ours	0.455	0.830	10.615	2
PVSE	0.324	0.759	20.463	3
PVSE+Ours w/o opt	0.393	0.809	15.195	2
PVSE+Ours w/o $L^{\mathcal{D}}$	0.479	0.860	9.771	1
PVSE+Ours	0.481	0.864	9.515	1
SAN	0.337	0.777	23.143	2
SAN+Ours w/o opt	0.397	0.814	14.293	2
SAN+Ours w/o $L^{\mathcal{D}}$	0.452	0.836	12.663	2
SAN+Ours	0.499	0.865	9.193	1
VSRN	0.403	0.701	15.323	1
VSRN+Ours w/o opt	0.455	0.825	9.995	1
VSRN+Ours w/o $L^{\mathcal{D}}$	0.503	0.854	7.524	1
VSRN+Ours	0.521	0.868	6.777	1

Table 2: Experimental results for R@k, mean rank, and median rank using the VG dataset.

	R@1	R@10	Mean	Med
UVS	0.0218	0.107	5842.051	588.5
UVS+Ours w/o opt	0.0331	0.156	3542.519	421
UVS+Ours w/o $L^{\mathcal{D}}$	0.0931	0.187	2699.894	183
UVS+Ours	0.0944	0.191	2659.331	180
OE	0.0248	0.110	4324.212	531.5
OE+Ours w/o opt	0.0348	0.162	2950.444	306
OE+Ours w/o $L^{\mathcal{D}}$	0.0998	0.210	2511.422	181
OE+Ours	0.102	0.211	2499.590	175
DPC	0.0263	0.118	3659.612	475
DPC+Ours w/o opt	0.0368	0.158	2958.345	275
DPC+Ours w/o $L^{\mathcal{D}}$	0.0911	0.197	2603.443	201
DPC+Ours	0.0939	0.201	2542.332	179
VSEPP	0.0266	0.118	3646.613	365
VSEPP+Ours w/o opt	0.0374	0.166	3002.231	246
VSEPP+Ours w/o $L^{\mathcal{D}}$	0.0992	0.206	2511.114	174
VSEPP+Ours	0.107	0.217	2485.237	168
RRF-Net	0.0268	0.119	3588.127	318
RRF-Net+Ours w/o opt	0.0392	0.168	2974.329	225
RRF-Net+Ours w/o $L^{\mathcal{D}}$	0.102	0.209	2605.666	201
RRF-Net+Ours	0.112	0.212	2500.776	177
CMPM	0.0270	0.122	3530.614	274
CMPM+Ours w/o opt	0.0387	0.172	2922.571	219
CMPM+Ours w/o $L^{\mathcal{D}}$	0.0943	0.199	2456.092	172
CMPM+Ours	0.105	0.229	2401.220	165
SISM	0.0275	0.121	3266.327	266
SISM+Ours w/o opt	0.0398	0.170	2901.392	212
SISM+Ours w/o $L^{\mathcal{D}}$	0.0951	0.201	2503.441	201
SISM+Ours	0.110	0.218	2478.531	171
GXN	0.0278	0.125	3185.392	271
GXN+Ours w/o opt	0.0409	0.174	2885.332	201
GXN+Ours w/o $L^{\mathcal{D}}$	0.110	0.219	2603.225	183
GXN+Ours	0.118	0.225	2500.032	165
PVSE	0.0280	0.129	2943.513	258
PVSE+Ours w/o opt	0.0431	0.179	2742.562	190
PVSE+Ours w/o $L^{\mathcal{D}}$	0.0985	0.221	2504.593	176
PVSE+Ours	0.104	0.234	2485.331	163
SAN	0.0286	0.130	2861.242	248
SAN+Ours w/o opt	0.0442	0.180	27694.322	185
SAN+Ours w/o $L^{\mathcal{D}}$	0.114	0.225	2634.507	164
SAN+Ours	0.122	0.242	2433.492	153
VSRN	0.0390	0.130	2861.242	248
VSRN+Ours w/o opt	0.0472	0.185	2800.293	176
VSRN+Ours w/o $L^{\mathcal{D}}$	0.120	0.233	2421.063	152
VSRN+Ours	0.124	0.248	2394.391	147

Table 3: Experimental results for DScore in MSCOCO and Visual Genome dataset. Since “Ours w/o opt” can be considered as upper limits of the other methods, we also show the margin between “Ours w/o opt” and the other methods.

	DScore in MSCOCO	DScore in VG
UVS+Ours w/o opt	0.978	0.949
UVS+Ours w/o $L^{\mathcal{D}}$	0.324 (-0.654)	0.239 (-0.710)
UVS+Ours	0.948 (-0.030)	0.917 (-0.032)
OE+Ours w/o opt	0.947	0.946
OE+Ours w/o $L^{\mathcal{D}}$	0.419 (-0.528)	0.429 (-0.517)
OE+Ours	0.900 (-0.047)	0.919 (-0.027)
DPC+Ours w/o opt	0.957	0.997
DPC+Ours w/o $L^{\mathcal{D}}$	0.461 (-0.496)	0.401 (-0.596)
DPC+Ours	0.955 (-0.002)	0.989 (-0.008)
VSEPP+Ours w/o opt	0.983	0.971
VSEPP+Ours w/o $L^{\mathcal{D}}$	0.319 (-0.664)	0.330 (-0.641)
VSEPP+Ours	0.979 (-0.004)	0.958 (-0.013)
RRF-Net+Ours w/o opt	0.990	0.985
RRF-Net+Ours w/o $L^{\mathcal{D}}$	0.310 (-0.680)	0.299 (-0.686)
RRF-Net+Ours	0.982 (-0.008)	0.970 (-0.015)
CMPM+Ours w/o opt	0.956	0.969
CMPM+Ours w/o $L^{\mathcal{D}}$	0.138 (-0.818)	0.172 (-0.797)
CMPM+Ours	0.954 (-0.002)	0.957 (-0.012)
SISM+Ours w/o opt	0.989	0.990
SISM+Ours w/o $L^{\mathcal{D}}$	0.429 (-0.560)	0.391 (-0.599)
SISM+Ours	0.979 (-0.010)	0.983 (-0.007)
GXN+Ours w/o opt	0.991	0.942
GXN+Ours w/o $L^{\mathcal{D}}$	0.310 (-0.681)	0.444 (-0.498)
GXN+Ours	0.988 (-0.003)	0.932 (-0.010)
PVSE+Ours w/o opt	0.901	0.948
PVSE+Ours w/o $L^{\mathcal{D}}$	0.293 (-0.708)	0.228 (-0.720)
PVSE+Ours	0.899 (-0.002)	0.930 (-0.018)
SAN+Ours w/o opt	0.999	0.985
SAN+Ours w/o $L^{\mathcal{D}}$	0.492 (-0.507)	0.300 (-0.685)
SAN+Ours	0.993 (-0.006)	0.984 (-0.001)
VSRN+Ours w/o opt	0.979	0.969
VSRN+Ours w/o $L^{\mathcal{D}}$	0.429 (-0.550)	0.325 (-0.644)
VSRN+Ours	0.968 (-0.011)	0.960 (-0.009)

Table 4: Experimental results for R@k, mean rank and median rank using the biased-MSOCO DB.

	R@1	R@10	Mean	Med
UVS	0.161	0.471	57.684	12
UVS+Ours w/o opt	0.185	0.492	50.338	10
UVS+Ours w/o $L^{\mathcal{D}}$	0.304	0.688	29.509	8
UVS+Ours	0.321	0.705	22.441	7
OE	0.194	0.532	39.144	10
OE+Ours w/o opt	0.205	0.588	31.115	8
OE+Ours w/o $L^{\mathcal{D}}$	0.303	0.719	21.442	7
OE+Ours	0.339	0.739	19.422	5
DPC	0.295	0.665	25.142	9
DPC+Ours w/o opt	0.310	0.698	23.684	8
DPC+Ours w/o $L^{\mathcal{D}}$	0.346	0.731	19.402	6
DPC+Ours	0.359	0.752	17.702	4
VSEPP	0.305	0.650	24.771	8
VSEPP+Ours w/o opt	0.343	0.697	20.190	6
VSEPP+Ours w/o $L^{\mathcal{D}}$	0.406	0.733	18.802	5
VSEPP+Ours	0.426	0.784	16.666	4
RRF-Net	0.314	0.635	19.324	8
RRF-Net+Ours w/o opt	0.329	0.652	18.492	7
RRF-Net+Ours w/o $L^{\mathcal{D}}$	0.384	0.693	17.994	7
RRF-Net+Ours	0.429	0.744	15.638	4
CMPM	0.326	0.687	18.567	6
CMPM+Ours w/o opt	0.342	0.702	16.441	5
CMPM+Ours w/o $L^{\mathcal{D}}$	0.450	0.780	12.885	3
CMPM+Ours	0.476	0.784	12.485	3
SISM	0.335	0.716	15.126	4
SISM+Ours w/o opt	0.349	0.752	13.092	4
SISM+Ours w/o $L^{\mathcal{D}}$	0.403	0.801	12.031	2
SISM+Ours	0.448	0.814	11.329	2
GXN	0.332	0.741	15.441	4
GXN+Ours w/o opt	0.354	0.746	14.441	4
GXN+Ours w/o $L^{\mathcal{D}}$	0.425	0.761	10.394	3
GXN+Ours	0.449	0.799	9.592	2
PVSE	0.343	0.761	15.810	2
PVSE+Ours w/o opt	0.359	0.779	14.104	2
PVSE+Ours w/o $L^{\mathcal{D}}$	0.438	0.844	12.204	1
PVSE+Ours	0.462	0.864	9.320	1
SAN	0.347	0.768	15.110	2
SAN+Ours w/o opt	0.352	0.793	14.392	2
SAN+Ours w/o $L^{\mathcal{D}}$	0.455	0.831	9.994	1
SAN+Ours	0.471	0.842	9.603	1
VSRN	0.379	0.793	14.332	2
VSRN+Ours w/o opt	0.381	0.803	12.331	2
VSRN+Ours w/o $L^{\mathcal{D}}$	0.499	0.877	7.506	1
VSRN+Ours	0.504	0.885	7.417	1

4 EVALUATING THE EFFECTS OF THE NUMBER OF QUESTIONS

In the main paper, we only provide the experimental results of the single round re-ranking. On the other hand, our method can be used for multiple round re-ranking, and the retrieval performance can be further enhanced by repeatedly conducting QA-based re-ranking. In this section, we provide additional experiments for confirming the effects of the multiple-round re-ranking (i.e., the effects of the number of questions).

Experimental results are shown in Tables 5, 6, 7, respectively. In each table, “UVS + Ours (1)” reveals 1st round re-ranking results using the method by UVS and our approach. Experimental results show that the retrieval performance gradually improves with the round increasing. From these results, the effectiveness of the multiple-round re-ranking is confirmed. However, compared with the performance improvement of the 1st round re-ranking, those of the 2nd and 3rd re-ranking are moderate. We consider that these results come from the proposed architecture that does not especially focus on multiple-round re-ranking. In our future work, we will consider its architecture for further enhancing retrieval performance.

Table 5: Experimental results for evaluating the effects of multiple-round re-ranking using MSCOCO dataset.

	R@1	R@10	Mean	Med
UVS	0.154	0.504	70.464	10
UVS+Ours (1)	0.329	0.734	40.596	5
UVS+Ours (2)	0.345	0.765	37.553	4
UVS+Ours (3)	0.364	0.787	33.183	4
OE	0.187	0.576	64.592	9
OE+Ours (1)	0.341	0.755	48.059	5
OE+Ours (2)	0.356	0.776	43.059	4
OE+Ours (3)	0.377	0.783	38.059	4
DPC	0.288	0.703	53.540	9
DPC+Ours (1)	0.398	0.838	18.592	4
DPC+Ours (2)	0.412	0.843	15.223	3
DPC+Ours (3)	0.422	0.847	14.332	3
VSEPP	0.297	0.724	52.440	9
VSEPP+Ours (1)	0.401	0.822	17.662	5
VSEPP+Ours (2)	0.421	0.833	16.542	4
VSEPP+Ours (3)	0.437	0.840	15.943	4
RRF-Net	0.294	0.709	47.300	9
RRF-Net+Ours (1)	0.417	0.840	15.511	4
RRF-Net+Ours (2)	0.423	0.843	14.594	4
RRF-Net+Ours (3)	0.430	0.845	14.031	4
CMPM	0.303	0.752	30.223	7
CMPM+Ours (1)	0.431	0.864	12.441	2
CMPM+Ours (2)	0.443	0.869	11.884	2
CMPM+Ours (3)	0.450	0.873	10.294	2
SISM	0.301	0.755	28.480	5
SISM+Ours (1)	0.433	0.849	12.011	2
SISM+Ours (2)	0.425	0.852	11.943	2
SISM+Ours (3)	0.419	0.853	11.543	2
GXN	0.307	0.746	26.731	3
GXN+Ours (1)	0.455	0.830	10.615	2
GXN+Ours (2)	0.463	0.834	9.925	1
GXN+Ours (3)	0.478	0.840	8.961	1
PVSE	0.324	0.759	20.463	3
PVSE+Ours (1)	0.481	0.864	9.515	1
PVSE+Ours (2)	0.490	0.869	9.090	1
PVSE+Ours (3)	0.499	0.873	8.760	1
SAN	0.337	0.777	23.143	2
SAN+Ours (1)	0.499	0.865	9.193	1
SAN+Ours (2)	0.499	0.865	9.193	1
SAN+Ours (3)	0.499	0.865	9.193	1
VSRN	0.403	0.701	15.323	1
VSRN+Ours (1)	0.521	0.868	6.777	1
VSRN+Ours (2)	0.530	0.875	6.057	1
VSRN+Ours (3)	0.539	0.879	5.872	1
PCME	0.352	0.765	25.322	3
PCME+Ours (1)	0.441	0.851	9.551	1
PCME+Ours (2)	0.464	0.860	9.403	1
PCME+Ours (3)	0.477	0.864	9.001	1
SDE	0.379	0.735	20.632	3
SDE+Ours (1)	0.475	0.858	8.339	1
SDE+Ours (2)	0.489	0.860	8.011	1
SDE+Ours (3)	0.490	0.864	7.694	1
CLIP	0.378	0.722	26.421	3
CLIP+Ours (1)	0.436	0.846	9.744	1
CLIP+Ours (2)	0.454	0.865	9.504	1
CLIP+Ours (3)	0.476	0.870	9.041	1
BLIP	0.402	0.753	18.773	2
BLIP+Ours (1)	0.531	0.867	8.631	1
BLIP+Ours (2)	0.544	0.870	8.302	1
BLIP+Ours (3)	0.554	0.873	8.001	1

Table 6: Experimental results for evaluating the effects of multiple-round re-ranking using the VG dataset.

	R@1	R@10	Mean	Med
UVS	0.0218	0.107	5842.051	588.5
UVS+Ours (1)	0.0944	0.191	2659.331	180
UVS+Ours (2)	0.102	0.199	2607.550	175
UVS+Ours (3)	0.105	0.210	2579.660	169
OE	0.0248	0.110	4324.212	531.5
OE+Ours (1)	0.102	0.211	2499.590	175
OE+Ours (2)	0.109	0.219	2403.430	170
OE+Ours (3)	0.113	0.224	2367.706	167
DPC	0.0263	0.118	3659.612	475
DPC+Ours (1)	0.0939	0.201	2542.332	179
DPC+Ours (2)	0.101	0.232	2340.450	170
DPC+Ours (3)	0.114	0.234	2304.533	165
VSEPP	0.0266	0.118	3646.613	365
VSEPP+Ours (1)	0.107	0.217	2485.237	168
VSEPP+Ours (2)	0.113	0.223	2405.432	163
VSEPP+Ours (3)	0.119	0.226	2349.409	156
RRF-Net	0.0268	0.119	3588.127	318
RRF-Net+Ours (1)	0.112	0.212	2500.776	177
RRF-Net+Ours (2)	0.118	0.219	2495.439	170
RRF-Net+Ours (3)	0.122	0.223	2465.053	164
CMPM	0.0270	0.122	3530.614	274
CMPM+Ours (1)	0.105	0.229	2401.220	165
CMPM+Ours (2)	0.119	0.235	2349.090	162
CMPM+Ours (3)	0.125	0.242	2300.001	154
SISM	0.0275	0.121	3266.327	266
SISM+Ours (1)	0.110	0.218	2478.531	171
SISM+Ours (2)	0.114	0.229	2403.293	165
SISM+Ours (3)	0.123	0.234	2392.211	159
GXN	0.0278	0.125	3185.392	271
GXN+Ours (1)	0.118	0.225	2500.032	165
GXN+Ours (2)	0.123	0.234	24788.653	160
GXN+Ours (3)	0.128	0.239	2400.549	154
PVSE	0.0280	0.129	2943.513	258
PVSE+Ours (1)	0.104	0.234	2485.331	163
PVSE+Ours (2)	0.114	0.243	2403.231	160
PVSE+Ours (3)	0.125	0.249	2385.654	154
SAN	0.0286	0.130	2861.242	248
SAN+Ours (1)	0.122	0.242	2433.492	153
SAN+Ours (2)	0.126	0.249	2400.210	150
SAN+Ours (3)	0.134	0.252	2349.329	143
VSRRN	0.0390	0.130	2861.242	248
VSRRN+Ours (1)	0.124	0.248	2394.391	147
VSRRN+Ours (2)	0.139	0.258	2268.549	140
VSRRN+Ours (3)	0.156	0.262	2194.549	138
PCME	0.0341	0.131	2822.341	253
PCME+Ours (1)	0.102	0.229	2499.323	164
PCME+Ours (2)	0.114	0.234	2349.432	154
PCME+Ours (3)	0.122	0.243	2302.320	149
SDE	0.0304	0.142	2755.495	249
SDE+Ours (1)	0.115	0.236	2423.391	154
SDE+Ours (2)	0.120	0.242	2402.223	150
SDE+Ours (3)	0.123	0.245	2304.534	143
CLIP	0.0405	0.151	2742.491	231
CLIP+Ours (1)	0.112	0.247	2313.441	146
CLIP+Ours (2)	0.121	0.253	2232.322	140
CLIP+Ours (3)	0.132	0.260	2021.201	132
BLIP	0.0454	0.173	2652.311	227
BLIP+Ours (1)	0.121	0.264	2223.486	128
BLIP+Ours (2)	0.127	0.272	2185.433	120
BLIP+Ours (3)	0.134	0.275	2100.212	114

Table 7: Experimental results for evaluating the effects of multiple-round re-ranking using the biased-MSOCO DB.

	R@1	R@10	Mean	Med
UVS	0.161	0.471	57.684	12
UVS+Ours (1)	0.321	0.705	22.441	7
UVS+Ours (2)	0.334	0.734	21.322	6
UVS+Ours (3)	0.345	0.756	20.002	4
OE	0.194	0.532	39.144	10
OE+Ours (1)	0.339	0.739	19.422	5
OE+Ours (2)	0.343	0.754	18.534	5
OE+Ours (3)	0.345	0.765	17.653	5
DPC	0.295	0.665	25.142	9
DPC+Ours (1)	0.359	0.752	17.702	4
DPC+Ours (2)	0.365	0.767	15.423	3
DPC+Ours (3)	0.377	0.774	12.332	3
VSEPP	0.305	0.650	24.771	8
VSEPP+Ours (1)	0.426	0.784	16.666	4
VSEPP+Ours (2)	0.434	0.798	15.432	3
VSEPP+Ours (3)	0.443	0.802	14.633	3
RRF-Net	0.314	0.635	19.324	8
RRF-Net+Ours (1)	0.429	0.744	15.638	4
RRF-Net+Ours (2)	0.434	0.754	14.533	3
RRF-Net+Ours (3)	0.444	0.765	13.330	3
CMPM	0.326	0.687	18.567	6
CMPM+Ours (1)	0.482	0.784	12.485	3
CMPM+Ours (2)	0.485	0.792	11.533	3
CMPM+Ours (3)	0.492	0.802	11.001	3
SISM	0.335	0.716	15.126	4
SISM+Ours (1)	0.448	0.814	11.329	2
SISM+Ours (2)	0.453	0.822	10.634	2
SISM+Ours (3)	0.466	0.827	9.978	2
GXN	0.332	0.741	15.441	4
GXN+Ours (1)	0.449	0.799	9.592	2
GXN+Ours (2)	0.452	0.807	8.645	2
GXN+Ours (3)	0.463	0.822	8.001	2
PVSE	0.343	0.761	15.810	2
PVSE+Ours (1)	0.462	0.864	9.320	1
PVSE+Ours (2)	0.477	0.872	8.654	1
PVSE+Ours (3)	0.485	0.884	8.043	1
SAN	0.347	0.768	15.110	2
SAN+Ours (1)	0.471	0.842	9.603	1
SAN+Ours (2)	0.479	0.855	9.032	1
SAN+Ours (3)	0.521	0.859	8.543	1
VSRRN	0.379	0.793	14.332	2
VSRRN+Ours (1)	0.504	0.885	7.417	1
VSRRN+Ours (2)	0.512	0.889	7.065	1
VSRRN+Ours (3)	0.527	0.893	6.767	1
PCME	0.348	0.749	18.551	3
PCME+Ours (1)	0.464	0.853	11.002	1
PCME+Ours (2)	0.486	0.861	10.390	1
PCME+Ours (3)	0.499	0.873	9.765	1
SDE	0.349	0.751	17.422	3
SDE+Ours (1)	0.469	0.861	10.338	1
SDE+Ours (2)	0.478	0.875	9.877	1
SDE+Ours (3)	0.499	0.879	8.995	1
CLIP	0.354	0.752	17.531	3
CLIP+Ours (1)	0.474	0.834	10.021	1
CLIP+Ours (2)	0.487	0.844	9.888	1
CLIP+Ours (3)	0.492	0.851	9.432	1
BLIP	0.385	0.763	14.212	2
BLIP+Ours (1)	0.489	0.874	8.411	1
BLIP+Ours (2)	0.494	0.898	8.096	1
BLIP+Ours (3)	0.504	0.902	7.653	1

5 EVALUATING THE EFFECT OF HYPERPARAMETERS

For further understanding of our approach, we extensively conducted various extensive ablation studies. Specifically, we examined the effects of the hyperparameters α , β , γ , and δ . Each experiment follows the same settings in the main paper, and we report the experimental results of changing each hyperparameter. For the experiments with α , β , γ , and δ , we show the experimental results with the MSCOCO dataset using R@1 and R@10.

Effect of α . The hyperparameter α balances the importance of each rank in the initial retrieval results. The lower (resp. higher) α leads to focus on the higher (resp. entire) ranks of the initial retrieval results. Experimental results regarding α are shown in Figs. 1 and 2. In each result, lower α indicates better performance. It implies that focusing on the higher ranks of the initial retrieval results is important for generating effective questions.

Effect of β . The hyperparameter β balances the effect of the initial retrieval and our re-ranking. The lower (resp. higher) β leads to mainly focus on initial retrieval (resp. our re-ranking). Experimental results regarding β are shown in Figs. 3 and 4. In each result, the higher β indicates better performance. It means that our re-ranking contributes to the improvement of the cross-modal image retrieval performance.

Effect of γ . The hyperparameter γ balances the importance of the re-ranking loss L^{rank} and the discriminative loss L^{D} . The lower (resp. higher) γ leads to focus on the re-ranking loss (resp. the discriminative loss). Experimental results regarding γ are shown in Figs. 5 and 6. In each result, the higher γ indicates better performance. Similar to the experimental results in the main paper, these results mean that introducing L^{D} is also effective for enhancing retrieval performance. Further analysis of these results can lead to further improvement of the retrieval performance, and then we will tackle them in our future works.

Effect of δ . The hyperparameter δ is a margin hyperparameter of the re-ranking loss. Its parameter determines the acceptable distance between the paired and the non-paired samples. Experimental results regarding δ are shown in Figs. 7 and 8. In each result, the performance suddenly decreases in $\gamma \geq 0.7$. We consider that these results come from the hyperparameter β . In this experiment, we set $\beta = 0.6$, and then our approach cannot make the distance between the paired and the non-paired samples by the value of 0.7. The relationships between β and γ will be further analyzed in our future works.

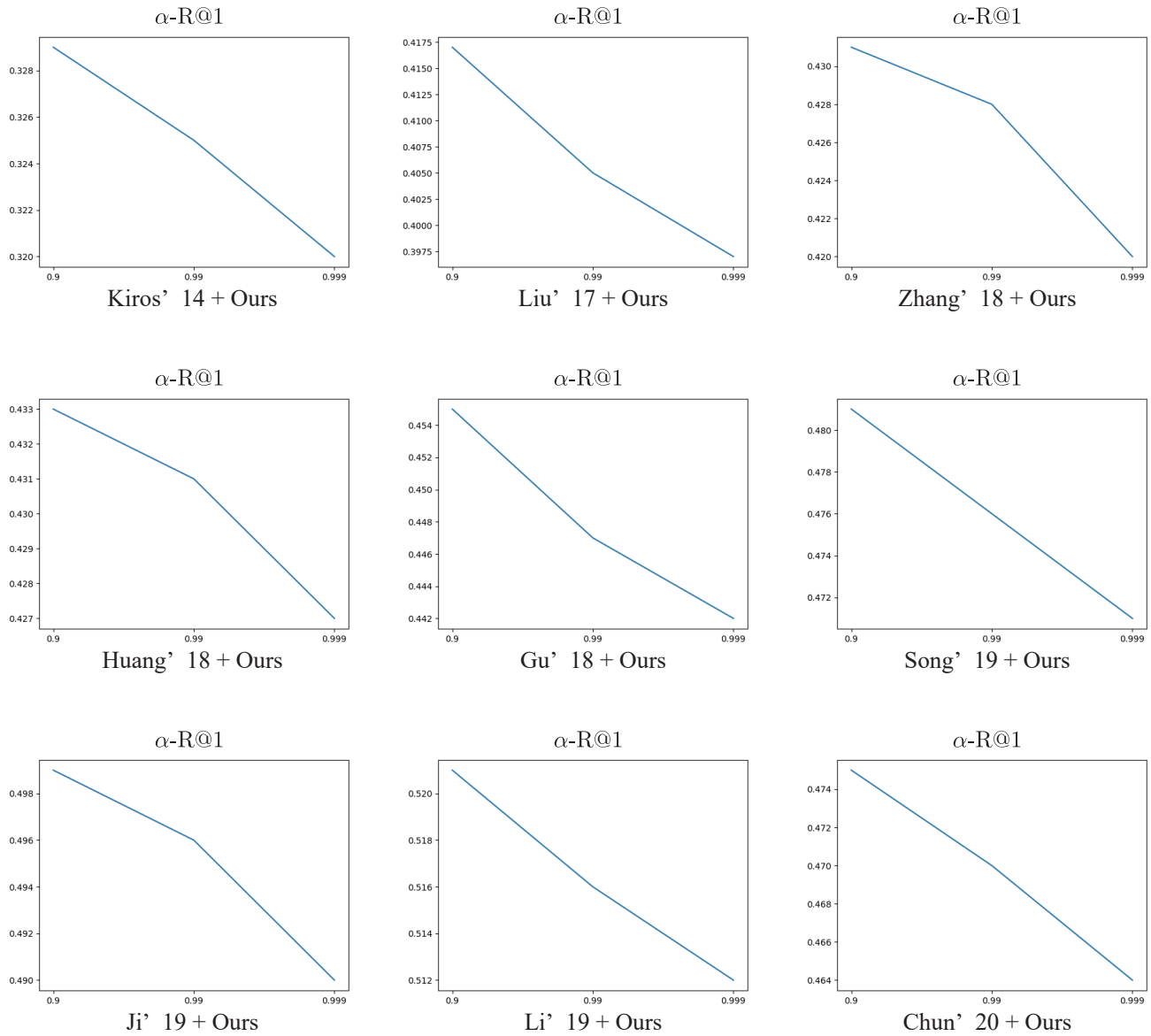


Figure 1: Experimental results for evaluating the effect of α on MSCOCO dataset using R@1. The higher value indicates better retrieval performance.

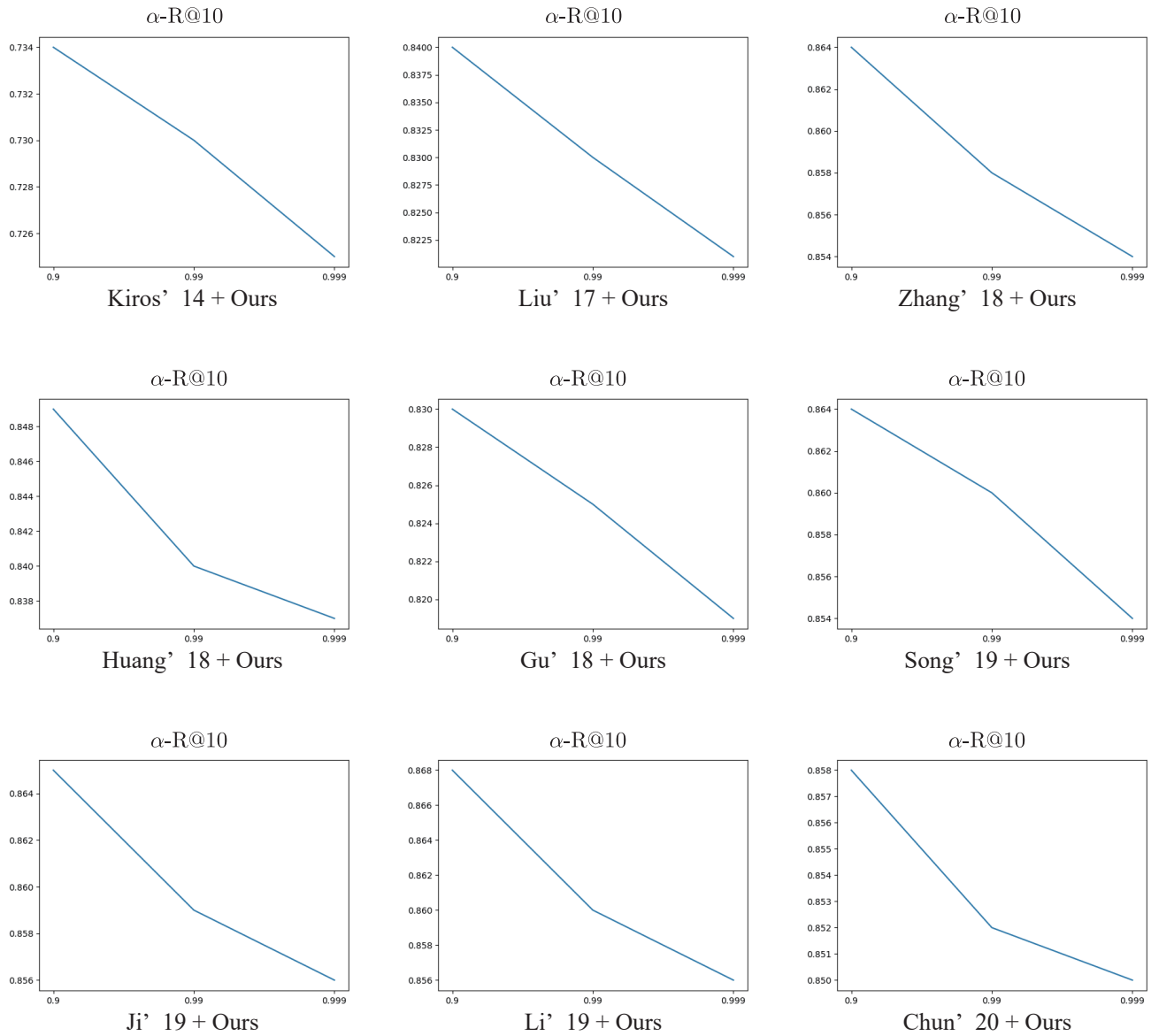


Figure 2: Experimental results for evaluating the effect of α on MSCOCO dataset using R@10. The higher value indicates better retrieval performance.

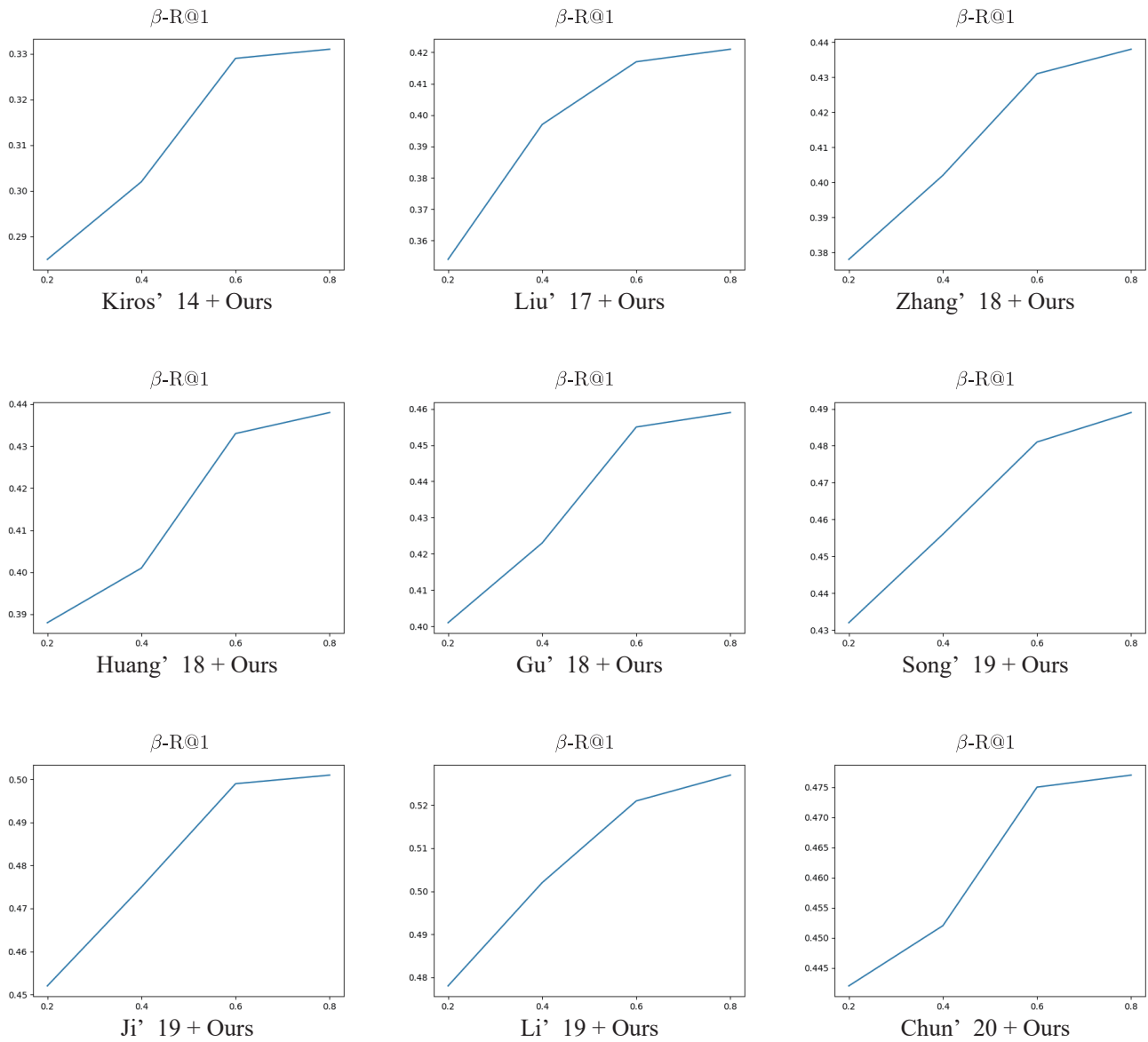


Figure 3: Experimental results for evaluating the effect of β on MSCOCO dataset using R@1. The higher value indicates better retrieval performance.

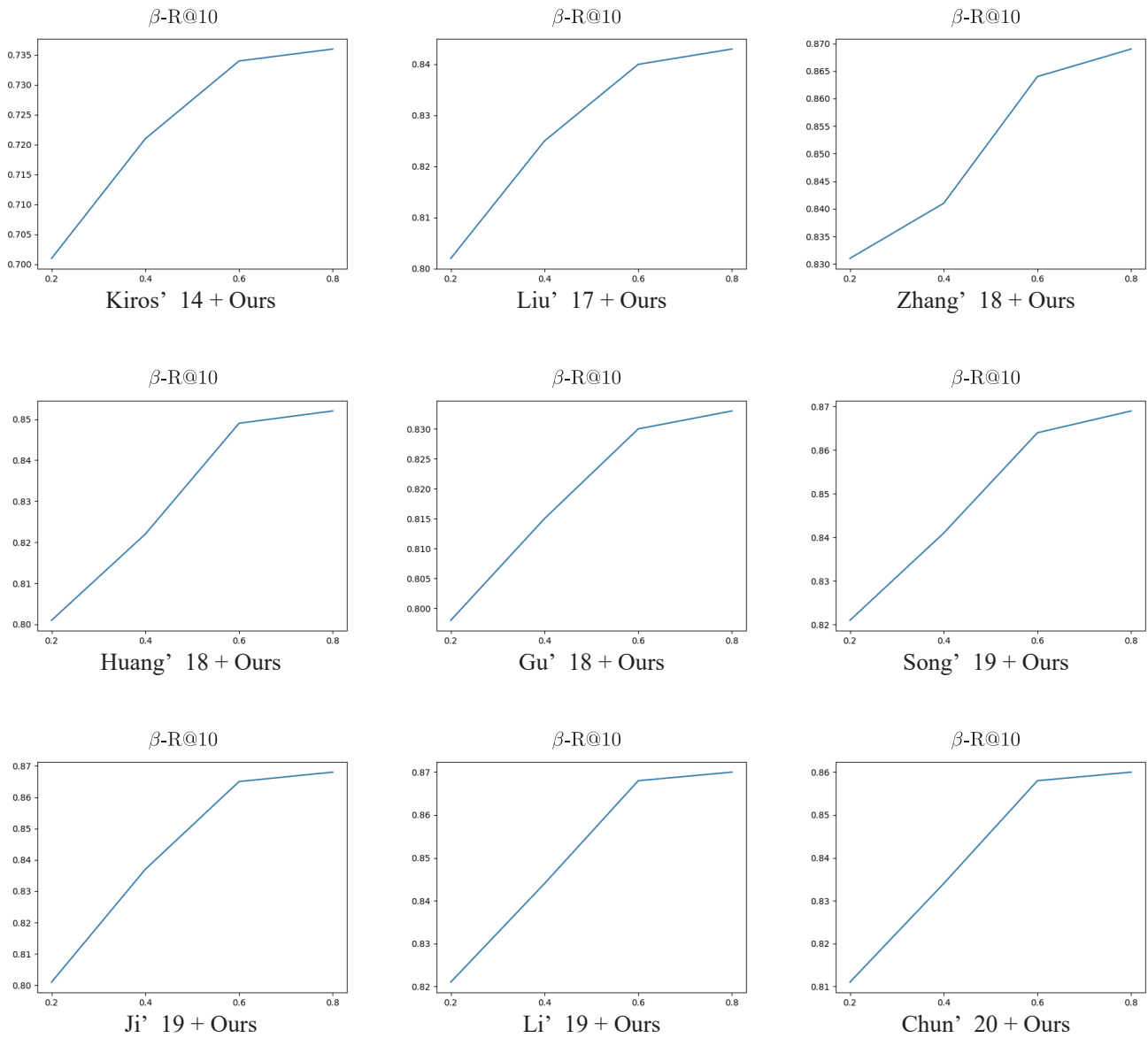


Figure 4: Experimental results for evaluating the effect of β on MSCOCO dataset using R@10. The higher value indicates better retrieval performance.

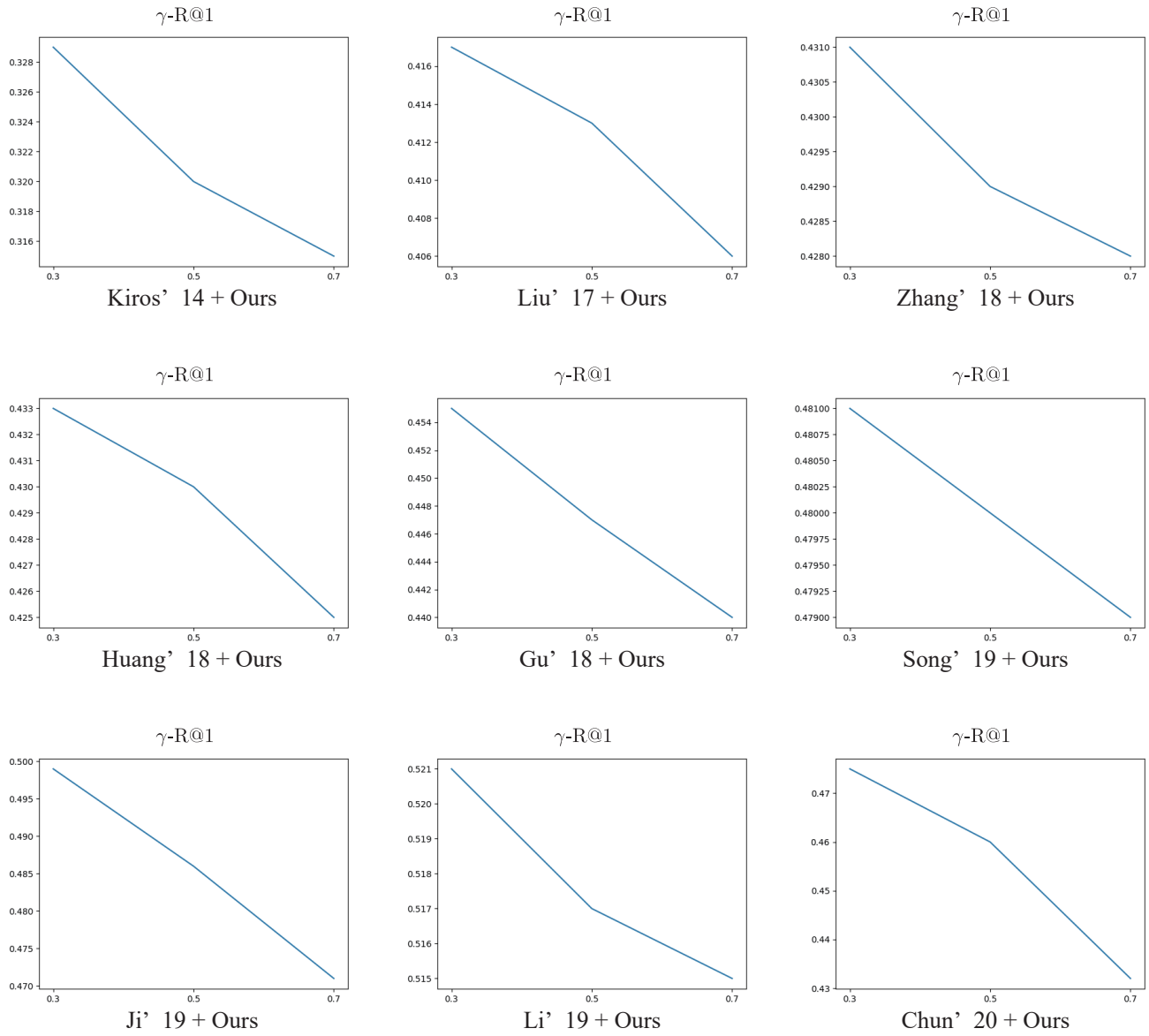


Figure 5: Experimental results for evaluating the effect of γ on MSCOCO dataset using R@1. The higher value indicates better retrieval performance.

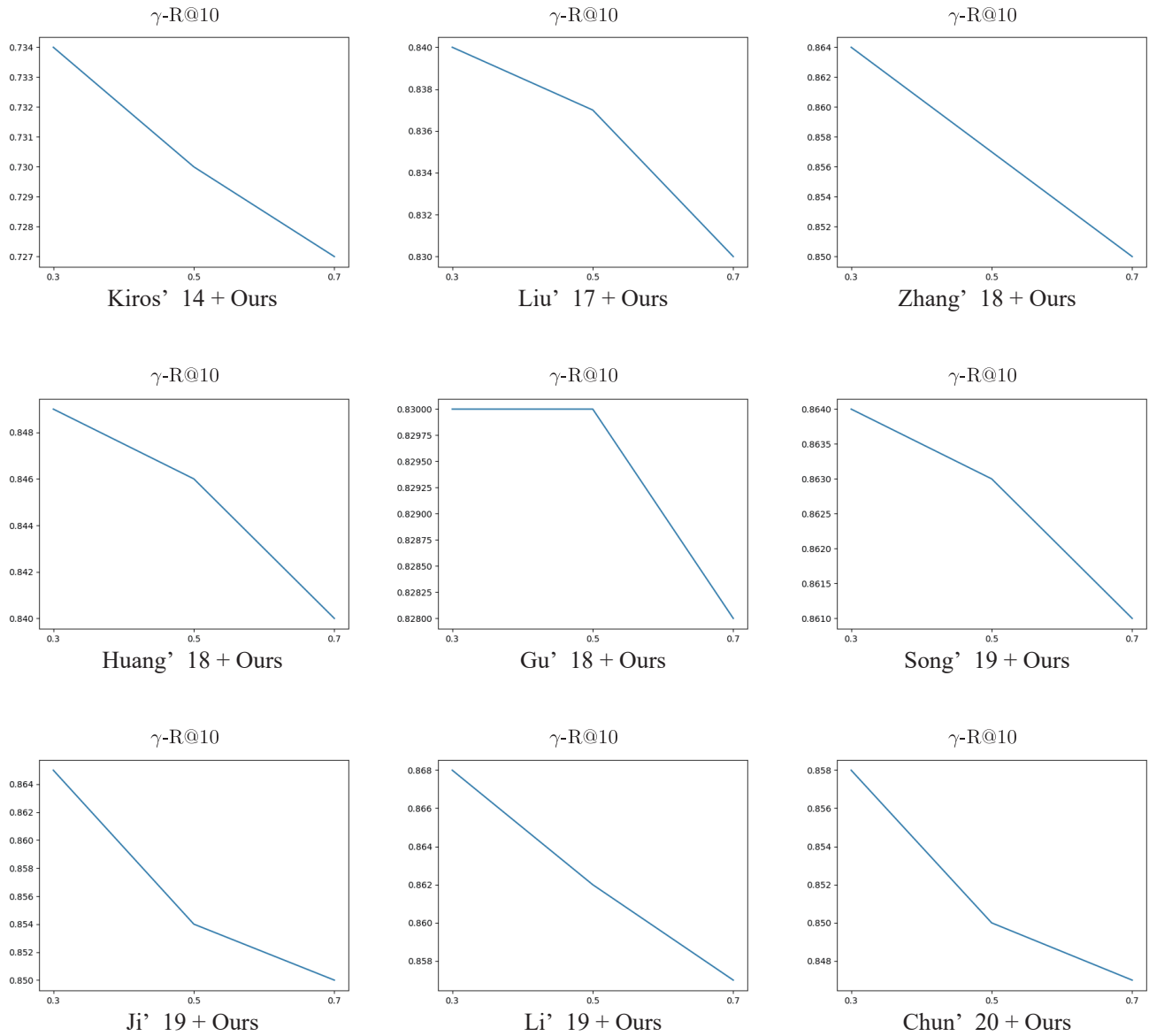


Figure 6: Experimental results for evaluating the effect of γ on MSCOCO dataset using R@10. The higher value indicates better retrieval performance.

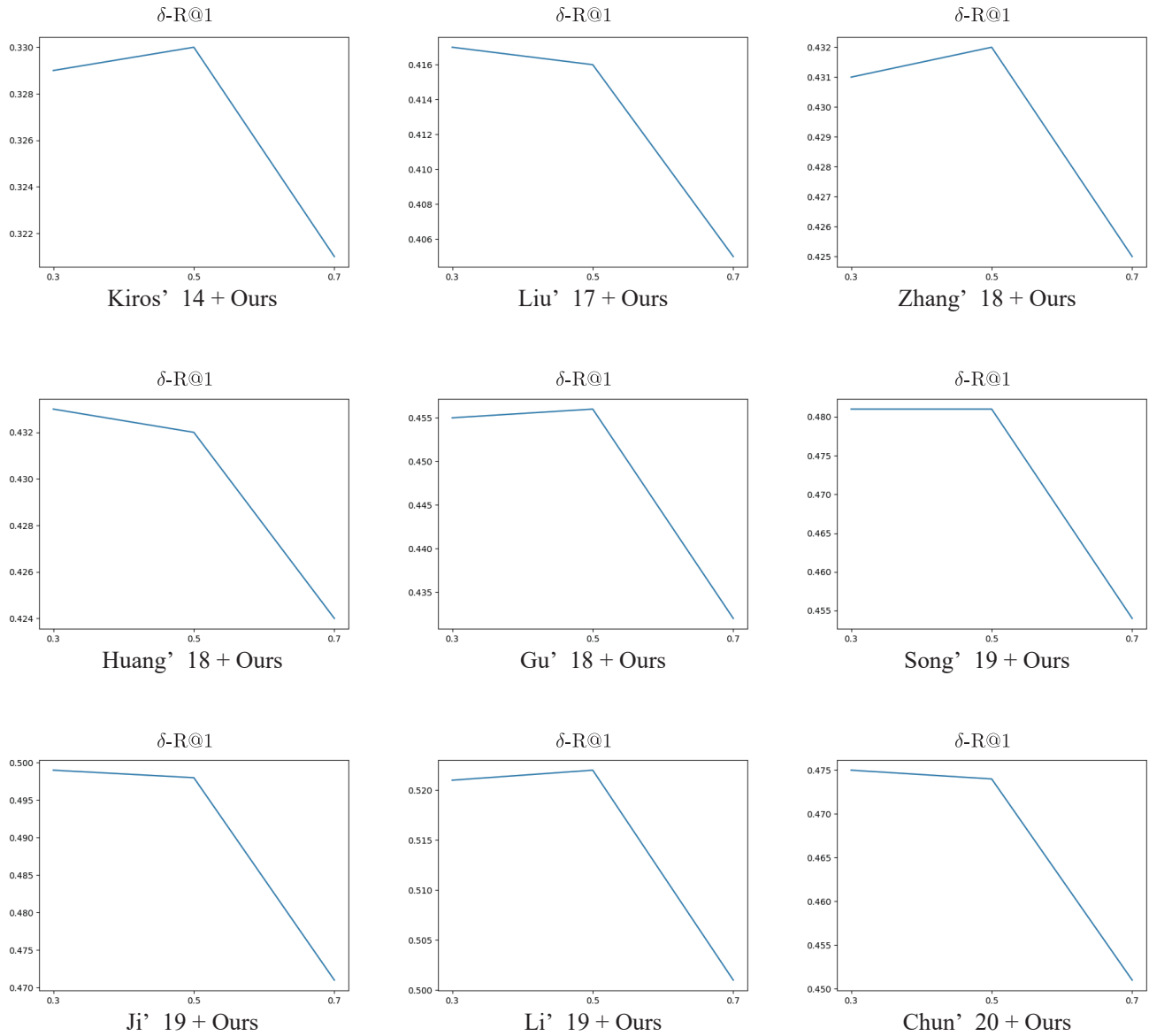


Figure 7: Experimental results for evaluating the effect of δ on MSCOCO dataset using R@1. The higher value indicates better retrieval performance.

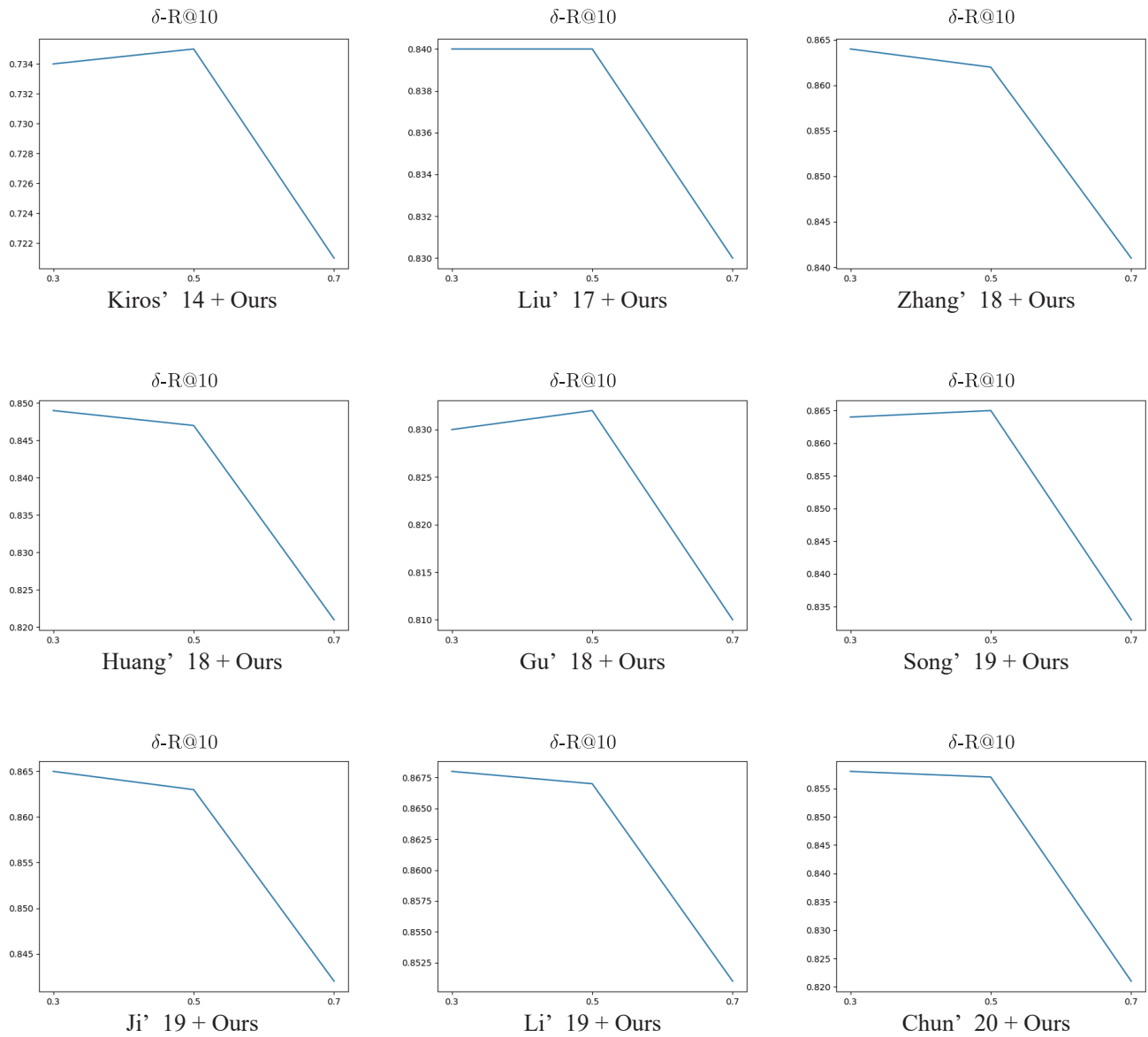


Figure 8: Experimental results for evaluating the effect of δ on MSCOCO dataset using R@10. The higher value indicates better retrieval performance.

REFERENCES

[1] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, Google Brain Toronto, and Sanja Fidler. 2017. VSE ++ : Improving visual-semantic embeddings with hard negatives. *arXiv:1707.05612* (2017).

[2] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7181–7189.

[3] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6163–6171.

[4] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. 2019. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 5754–5763.

[5] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539* (2014).

[6] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 4654–4662.

[7] Yu Liu, Yanming Guo, Erwin M. Bakker, and Michael S. Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 4107–4116.

[8] Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1979–1978.

[9] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *Proceedings of the International Conference on Learning Representations*. 1–12.

[10] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the IEEE European Conference on Computer Vision*. 686–701.

[11] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embedding with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2020).