A. Sensitivity to Delta: GloVe Experiment

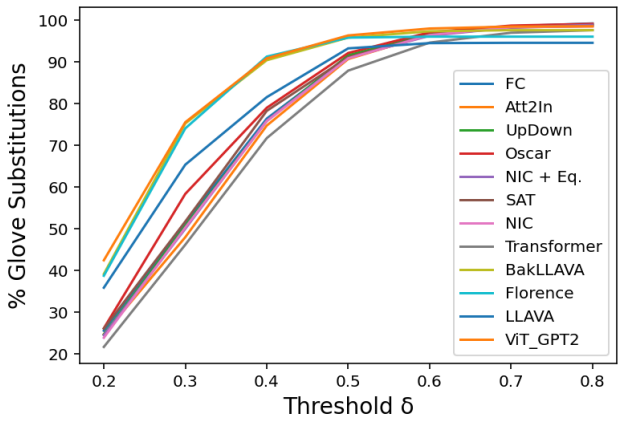

Figure 1. **Glove Thresholds:** Percentage of glove substitutions vs. the selected threshold δ across different models.

To select the optimal distance threshold δ for our experiments, we observed the number of contextual substitutions that occurred at different thresholds. We tested this using the GloVe word embedding model.

If δ is large, we allow weaker word substitutions. Consider the HGC: "a <gender> is sleeping". Assume the word "sleeping" is not present in V_{MGC}. If δ is large, "sleeping" can get replaced with an unrelated word like "book". If δ is small, we only allow very similar word substitutions. But we increase the risk of <unk> replacements.

In Figure 1, the x-axis shows different distance thresholds (δ), and the y-axis shows the % of contextual substitutions out of the total substitutions (contextual + <unk>).

For GloVe, we found $\delta = 0.4$ to be the best balance. Below 0.4, we had very less contextual substitutions ($< 60\%$). Above $\delta = 0.4$, we had a lot of contextual substitutions. However, a large number of contextual substitutions could be unrelated words (e.g., replacing sleeping with book).

B. Experiment Setup: Additional Details

For language-aware metrics (LIC and DBAC), we used an LSTM-unidirectional model as our sentence encoder (unless specified otherwise). We paired this LSTM with a 3-layer MLP to form our final attacker model.

For the controlled experiment 5.1, we used accuracy as our quality function (Q). For experiments 5.2, 5.3, and 5.4, we set $Q = 1/\textit{cross-entropy}$.

We performed vocabulary alignment in our experiments using GloVe at $\delta = 0.4$ (unless specified otherwise). We found $\delta = 0.4$ to be a suitable threshold based on the sensitivity experiment we performed in Section A. While FastText is better at vocabulary alignment, we used Glove to show that DBAC is superior even if we use a slightly weaker embedding model for alignment.

C. PseudoCode

Algorithm 1 Pseudocode for $DBAC_{A \to T}$

Initialize δ
Initialize $f^a_{s_a}$, $f^a_{\hat{s}_a}$
Fetch s_a, $\hat{s}_a$, a, t, $\hat{t}$
Calculate $P(t)$, $P(\hat{t})$, $P(a)$
$s_a \leftarrow Mask\_Attribute(s_a)$
$\hat{s}_a \leftarrow Mask\_Attribute(\hat{s}_a)$
$s_a, \hat{s}_a \leftarrow VocabAlign(s_a, \hat{s}_a, \delta)$
Train $(f^a_{s_a}, a)$
Train $(f^a_{\hat{s}_a}, a)$
$\omega_{H:A \to T} \leftarrow Q(f^a_{s_a}, a) \cdot \frac{P(t)}{P(a)}$
$\omega_{M:A \to T} \leftarrow Q(f^a_{\hat{s}_a}, a) \cdot \frac{P(\hat{t})}{P(a)}$
$DBAC_{A \to T} \leftarrow \frac{(\omega_{M:A \to T} - \omega_{H:A \to T})}{(\omega_{M:A \to T} + \omega_{H:A \to T} + \epsilon)}$

Algorithm 2 Pseudocode for $DBAC_{T \to A}$

Initialize δ
Initialize $f^t_{s_t}$, $f^t_{\hat{s}_t}$
Fetch s_t, $\hat{s}_t$, t, a, $\hat{a}$
Calculate $P(a)$, $P(\hat{a})$, $P(t)$
$s_t \leftarrow Mask\_Task(s_t)$
$\hat{s}_t \leftarrow Mask\_Task(\hat{s}_t)$
$s_t, \hat{s}_t \leftarrow VocabAlign(s_t, \hat{s}_t, \delta)$
Train $(f^t_{s_t}, t)$
Train $(f^t_{\hat{s}_t}, t)$
$\omega_{H:T \to A} \leftarrow Q(f^t_{s_t}, t) \cdot \frac{P(a)}{P(t)}$
$\omega_{M:T \to A} \leftarrow Q(f^t_{\hat{s}_t}, t) \cdot \frac{P(\hat{a})}{P(t)}$
$DBAC_{T \to A} \leftarrow \frac{\omega_{M:T \to A} - \omega_{H:T \to A}}{\omega_{M:T \to A} + \omega_{H:T \to A} + \epsilon}$

D. Sensitivity to sentence encoders: Bias amplification scores

For gender, we reported $\text{DBAC}_{A \to T}$ scores on six sentence encoders trained from scratch on COCO captions, in Table 10. We reported the corresponding LIC scores in Table 11. We reported $\text{DBAC}_{A \to T}$ scores on the four pre-trained encoders in Table 12a. We reported the corresponding LIC scores in Table 12b.

For race, we reported $\text{DBAC}_{A \to T}$ scores on six sentence encoders trained from scratch, in Table 13. We reported the corresponding LIC scores in Table 14. We reported $\text{DBAC}_{A \to T}$ scores on four pre-trained encoders in Table 15a. We reported the corresponding LIC scores in Table 15b.

E. Race results for constant vs. contextual substitution

In column 2 of Table 16, for race HGCs, we showed the METEOR similarity scores for constant substitution. Columns 3 and 4 show the METEOR scores for contextual substitution with Glove and FastText, respectively. The percentage values in green indicate the % increase in METEOR score achieved by contextual substitution, relative to constant substitution.

For race, contextual substitution better retained the original meaning of the HGCs. Across all captioning models, contextual substitution produced higher similarity scores than constant substitution. On average, contextual substitution improved the METEOR sentence similarity score by 1.6% with Glove, and 2.2% with FastText.

In column 2 of Table 17a, we showed DBAC scores of the HGCs for race ($\omega_{H:A \to T}$) with constant substitution. Columns 3 and 4 show the DBAC scores for contextual substitution with Glove and FastText, respectively. We created an identical table for LIC scores in Table 17b.

For race, DBAC and LIC reported higher bias in the HGCs with contextual substitution. On average, contextual substitution increased DBAC scores by 5.2% with Glove, and 4.4% with FastText (Table 17a). Contextual substitution increased LIC scores by 3.0% with Glove, and 3.9% with FastText (Table 17b).

Models	LSTM	LSTM Bidirectional	RNN	RNN Bidirectional	Transformer (1 head)	Transformer (5 head)
Att2In	0.0590	0.0778	0.0540	0.0620	0.1109	0.2669
BakLLAVA	−0.0298	−0.0637	−0.0259	−0.0396	−0.0554	−0.0848
BLIP	−0.0998	−0.1102	−0.1344	−0.0545	−0.1395	−0.1442
FC	0.0645	0.0836	0.0105	0.0775	0.1269	0.1822
Florence	−0.1025	−0.1558	−0.0600	−0.1316	−0.1060	−0.1876
LLAVA	−0.0676	−0.1005	−0.0761	−0.1298	−0.0892	−0.0855
Oscar	0.1119	0.1407	0.1052	0.1013	0.1138	0.1641
SAT	0.0761	0.0856	0.0830	0.0832	0.0717	0.1175
NIC	0.2743	0.2467	0.1877	0.2694	0.1623	0.3122
NIC+Equal	0.2552	0.2261	0.1417	0.2837	0.1843	0.3511
Transformer	0.1215	0.1358	0.1325	0.1511	0.1525	0.2335
UpDn	0.1172	0.1193	0.1121	0.1091	0.1511	0.2357
Vit_GPT2	−0.0206	−0.0123	−0.0841	−0.0051	−0.0279	−0.0452

Table 10. DBAC scores for gender, on the six encoders trained from scratch on COCO captions.

Models	LSTM	LSTM Bidirectional	RNN	RNN Bidirectional	Transformer (1 head)	Transformer (5 head)
Att2In	0.0001	0.0064	−0.0119	−0.0294	−0.0005	0.0004
BakLLAVA	0.0726	0.0160	0.0232	0.0000	−0.0004	−0.0019
BLIP	0.0649	0.0194	0.0000	−0.0213	−0.0005	0.0001
FC	0.0669	−0.0240	−0.0225	−0.0307	−0.0009	0.0001
Florence	−0.0236	0.0161	−0.0065	−0.0068	−0.0002	−0.0005
LLAVA	0.0021	0.0383	−0.0002	0.0077	0.0010	0.0002
Oscar	−0.0092	0.0385	−0.0154	−0.0153	0.0003	−0.0026
SAT	0.0405	0.0162	0.0099	0.0159	−0.0001	0.0002
NIC	0.0107	0.0242	0.0337	−0.0018	−0.0019	−0.0006
NIC+Equal	0.0408	0.0009	−0.0462	−0.0305	0.0005	0.0003
Transformer	0.0310	0.0195	0.0401	0.0259	−0.0001	0.0000
UpDn	0.0549	0.0000	−0.0114	0.0163	−0.0020	0.0002
Vit_GPT2	0.0263	0.0295	−0.0307	0.0000	−0.0015	0.0003

Table 11. LIC scores for gender, on the six encoders trained from scratch on COCO captions.

Models	DistilBERT	DistilRoBERTa	MiniLM	MPNet
Att2In	0.0222	0.0206	0.0186	0.0145
BakLLAVA	0.0680	0.0774	0.0804	0.0475
BLIP	−0.0044	−0.0026	−0.0004	−0.0016
FC	−0.0231	−0.0162	−0.0016	−0.0154
Florence	0.1060	0.1221	0.1182	0.0797
LLAVA	0.0979	0.0740	0.1038	0.0581
Oscar	0.0442	0.0368	0.0457	0.0328
SAT	0.0183	0.0202	0.0047	0.0277
NIC	0.0850	0.0166	0.0185	0.0103
NIC+Equal	−0.0214	−0.0040	−0.0014	−0.0081
Transformer	0.0369	0.0434	0.0340	0.0463
UpDn	0.0137	0.0200	0.0196	0.0189
Vit_GPT2	0.0154	0.0175	0.0334	0.0216

(a) DBAC scores for gender on the pre-trained encoders

Models	DistilBERT	DistilRoBERTa	MiniLM	MPNet
Att2In	−0.0013	0.0004	−0.0015	−0.0014
BakLLAVA	0.0013	−0.0005	−0.0002	−0.0001
BLIP	−0.0004	−0.0001	−0.0001	−0.0005
FC	−0.0001	−0.0002	0.0002	−0.0003
Florence	−0.0028	−0.0001	0.0000	0.0007
LLAVA	−0.0032	0.0003	0.0066	−0.0006
Oscar	0.0018	0.0003	−0.0010	0.0002
SAT	−0.0008	0.0008	0.0019	−0.0006
NIC	0.0007	0.0001	−0.0034	0.0001
NIC+Equal	−0.0001	−0.0006	0.0001	−0.0006
Transformer	0.0000	0.0000	0.0011	−0.0006
UpDn	0.0009	0.0004	0.0019	−0.0007
Vit_GPT2	0.0001	0.0011	−0.0012	−0.0008

(b) LIC scores for gender on the pre-trained encoders

Table 12. DBAC and LIC scores for gender, on the four pre-trained encoders

Models	LSTM	LSTM Bidirectional	RNN	RNN Bidirectional	Transformer (1 head)	Transformer (5 head)
Att2In	0.1572	0.2254	0.1224	0.1864	0.1941	0.1010
BakLLAVA	−0.0945	−0.0670	−0.1804	−0.1203	−0.0810	−0.0803
BLIP	0.1557	0.0446	0.1125	0.1683	0.2367	0.1894
FC	0.2754	0.3091	0.2207	0.1572	0.2689	0.3115
Florence	−0.4296	−0.4052	−0.4295	−0.3960	−0.5565	−0.5210
LLAVA	−0.0676	−0.1005	−0.0761	−0.1298	−0.0892	−0.0855
Oscar	0.1975	0.1655	0.2962	0.1461	0.3409	0.3939
SAT	0.1048	0.1610	0.1163	0.1211	0.1252	0.0959
NIC	0.1104	0.1140	0.0835	0.1753	0.1961	0.1677
NIC+Equal	0.2248	0.2585	0.3363	0.2077	0.3046	0.3759
Transformer	0.2083	0.2046	0.1894	0.1905	0.2861	0.3129
UpDn	0.1223	0.1668	0.1477	0.1480	0.2165	0.1611
Vit_GPT2	0.2521	0.1965	0.3782	0.1201	0.3610	0.3735

Table 13. DBAC scores for race, on the six encoders trained from scratch on COCO captions.

Models	LSTM	LSTM Bidirectional	RNN	RNN Bidirectional	Transformer (1 head)	Transformer (5 head)
Att2In	0.0133	0.0127	0.0111	−0.0332	0.0298	0.0629
BakLLAVA	0.0293	0.0272	0.0232	−0.0555	0.0304	0.0373
BLIP	0.0111	0.0046	0.0112	0.0213	0.0066	0.0036
FC	−0.0404	−0.0206	−0.0444	−0.0444	−0.0263	−0.0200
Florence	0.0379	0.0364	0.0333	0.0254	0.0458	0.0734
LLAVA	0.0221	0.0413	0.0333	0.0077	0.0472	0.0502
Oscar	0.0200	0.0373	0.0333	0.0167	0.0430	0.0618
SAT	0.0419	0.0391	0.0333	0.0111	0.0400	0.0581
NIC	0.0036	0.0032	0.0000	0.0111	0.0097	0.0245
NIC+Equal	0.0176	0.0167	0.0111	−0.0665	0.0157	0.0342
Transformer	0.0105	0.0084	0.0000	−0.0444	0.0146	0.0340
UpDn	0.0106	0.0092	0.0000	0.0111	0.0108	0.0114
Vit_GPT2	−0.0104	−0.0265	−0.0222	−0.0212	−0.0080	−0.0062

Table 14. LIC scores for race, on the six encoders trained from scratch on COCO captions.

Models	DistilBERT	DistilRoBERTa	MiniLM	MPNet
Att2In	0.0432	0.0626	0.0587	0.0519
BakLLAVA	−0.1794	−0.1696	−0.1460	−0.1715
BLIP	−0.0084	−0.0093	−0.0268	−0.0116
FC	0.0162	0.0218	0.0297	0.0238
Florence	−0.5405	−0.5185	−0.4997	−0.5155
LLAVA	−0.1507	−0.1789	−0.1632	−0.1805
Oscar	0.0366	0.0127	0.0107	0.0320
SAT	0.0108	0.0121	0.0260	0.0232
NIC	0.0337	0.0103	0.0273	0.0231
NIC+Equal	0.0109	0.0044	0.0230	0.0223
Transformer	0.0205	0.0127	0.0265	0.0182
UpDn	0.0214	0.0159	0.0287	0.0329
Vit_GPT2	0.0433	0.0374	0.0380	0.0166

(a) DBAC scores for race on the pre-trained encoders

Models	DistilBERT	DistilRoBERTa	MiniLM	MPNet
Att2In	0.0011	0.0002	0.0004	0.0001
BakLLAVA	0.0003	−0.0005	−0.0002	−0.0003
BLIP	0.0011	0.0002	0.0003	−0.0001
FC	0.0003	−0.0001	0.0005	−0.0003
Florence	0.0005	0.0004	0.0014	0.0000
LLAVA	0.0049	−0.0018	0.0007	−0.0031
Oscar	0.0012	0.0000	0.0000	0.0004
SAT	−0.0012	−0.0004	0.0003	−0.0001
NIC	0.0003	−0.0003	0.0004	−0.0001
NIC+Equal	0.0002	−0.0002	0.0002	−0.0002
Transformer	0.0023	0.0000	−0.0001	0.0000
UpDn	0.0006	−0.0002	0.0008	0.0005
Vit_GPT2	0.0002	0.0001	0.0002	−0.0001

(b) LIC scores for race on the pre-trained encoders

Table 15. DBAC and LIC scores for race, on the four pre-trained encoders.

Models	Constant	Contextual	
		Glove	FastText
Att2In	0.709	0.732 (3.22% ↑)	0.741 (4.44% ↑)
BakLLAVA	0.901	0.901 (0.01% ↑)	0.901 (0.01% ↑)
BLIP	0.855	0.856 (0.08% ↑)	0.855 (0.01% ↑)
FC	0.654	0.680 (3.90% ↑)	0.686 (4.94% ↑)
Florence	0.881	0.881 (0.00% ↑)	0.881 (0.00% ↑)
LLAVA	0.891	0.891 (0.01% ↑)	0.891 (0.00% ↑)
Oscar	0.706	0.724 (2.66% ↑)	0.731 (3.56% ↑)
SAT	0.790	0.793 (0.38% ↑)	0.805 (1.86% ↑)
NIC	0.804	0.805 (0.19% ↑)	0.806 (0.24% ↑)
NIC+Equal.	0.772	0.794 (2.84% ↑)	0.805 (4.22% ↑)
Transformer	0.837	0.850 (1.56% ↑)	0.853 (1.91% ↑)
UpDown	0.735	0.764 (3.90% ↑)	0.769 (4.63% ↑)
VIT_GPT2	0.815	0.829 (1.66% ↑)	0.839 (2.92% ↑)
Average	0.796	0.808 (1.57% ↑)	0.812 (2.21% ↑)

Table 16. METEOR scores for constant vs. contextual substitution (race results): Values in green indicate % increase in METEOR scores with contextual substitution, relative to constant substitution. For all captioning models, contextual substitution achieved a higher METEOR score than constant substitution.

Models	Constant	Contextual	
		Glove	FastText
Att2In	0.290	0.300 (3.49% ↑)	0.309 (6.45% ↑)
BakLLAVA	0.281	0.283 (0.93% ↑)	0.284 (1.10% ↑)
BLIP	0.287	0.291 (1.11% ↑)	0.291 (1.29% ↑)
FC	0.284	0.291 (2.68% ↑)	0.286 (0.74% ↑)
Florence	0.281	0.307 (9.52% ↑)	0.314 (12.09%↑)
LLAVA	0.274	0.278 (1.57% ↑)	0.275 (0.37% ↑)
Oscar	0.305	0.342 (12.38%↑)	0.329 (7.88% ↑)
SAT	0.294	0.299 (1.84% ↑)	0.307 (4.50% ↑)
NIC	0.265	0.290 (9.44% ↑)	0.269 (1.70% ↑)
NIC+Equal.	0.300	0.302 (0.93% ↑)	0.302 (0.67% ↑)
Transformer	0.263	0.291 (10.76%↑)	0.280 (6.46% ↑)
UpDown	0.273	0.294 (7.70% ↑)	0.288 (5.42% ↑)
VIT_GPT2	0.300	0.316 (5.36% ↑)	0.326 (8.73% ↑)
Average	0.284	0.299 (5.21% ↑)	0.297 (4.42% ↑)

(a) DBAC race results for the HGCs: $\omega_{H:A\rightarrow T}$

Models	Constant	Contextual	
		Glove	FastText
Att2In	0.518	0.539 (4.07% ↑)	0.545 (5.15% ↑)
BakLLAVA	0.521	0.528 (1.27% ↑)	0.539 (3.47% ↑)
BLIP	0.581	0.609 (4.70% ↑)	0.591 (1.67% ↑)
FC	0.522	0.540 (3.33% ↑)	0.530 (1.44% ↑)
Florence	0.517	0.532 (2.86% ↑)	0.532 (2.84% ↑)
LLAVA	0.544	0.555 (1.93% ↑)	0.548 (0.74% ↑)
Oscar	0.522	0.536 (2.70% ↑)	0.545 (4.41% ↑)
SAT	0.536	0.545 (1.64% ↑)	0.558 (4.18% ↑)
NIC	0.545	0.563 (3.30% ↑)	0.580 (6.41% ↑)
NIC+Equal.	0.545	0.563 (3.26% ↑)	0.562 (3.15% ↑)
Transformer	0.518	0.527 (1.72% ↑)	0.528 (1.97% ↑)
UpDown	0.519	0.534 (2.87% ↑)	0.547 (5.37% ↑)
VIT_GPT2	0.496	0.521 (5.17% ↑)	0.548 (10.47%↑)
Average	0.530	0.545 (2.99% ↑)	0.550 (3.94% ↑)

(b) LIC race results for the HGCs

Table 17. DBAC and LIC scores for constant vs. contextual substitution (race results): Values in green indicate the % increase in HGC's bias reported by contextual substitution relative to constant substitution. Across all captioning models, both DBAC and LIC consistently reported higher bias in the HGCs with contextual substitution.