A Datasets

All style data used for neuron identification are obtained from publicly available datasets. We applied the following preprocessing to the raw data: (1) removing sentences longer than 120 characters; (2) eliminating duplicate sentences; and (3) removing sentences containing a large number of special symbols. Table 7 provides detailed statistics of the preprocessed corpus.

B Classifiers used in Each Benchmark

To evaluate the accuracy of style transfer, we use open-source classifiers on the six benchmarks we evaluated. The sources of these classifiers are detailed in Table 8.

C JSD Distance between Layers

To verify whether the style-specific layers selected in Section 3.3 encode stylistic information, we calculate the Jensen-Shannon Divergence (JSD) distances between the final layer and all previous layers for the TST task of transfer from informal style text to formal style text. The results, shown in Table 9, led to the following findings: (1) For most of the early layers, from layer 0 to 26, the distances between the final layer and these layers remain almost constant or change very little, indicating that the information encoded in these layers is very similar. However, for the last few layers, from layer 27 to 31, the JSD distance from the final layer is smaller compared to the earlier layers, but the distance between different layers increases. This suggests that the last few layers are processing style-related information, consistent with the distribution characteristics of the style layers discussed in Section 3.3. (2) Some words associated with the formal style (target-side style), highlighted in bold in the Table 9, show a larger distance difference in the last few layers. This aligns with our expectation that words representing the target style are more likely to be activated in the style layer, increasing their probability of being selected as candidates for token generation in the style layer.

D Effectiveness of different model sizes

To verify the effectiveness of our method on a larger model, we conduct experiments using the 70B version of LLaMA-3. The results, presented in Table 10, indicate that our method is also effective on the larger model and consistent with the conclusions drawn from the 7B model (See Table 2 in Section 5 for more details).

Style		Dola	Our
Formality	informal→formal	78.14	80.80
	formal→informal	12.63	14.40
Toxicity	toxic→neutral	49.25	55.36
	neutral→toxic	25.41	31.98
Politics	democratic→republican	36.26	37.81
	republican→democratic	46.25	50.30
Politeness	impolite→polite	76.58	80.63
	polite→impolite	20.57	23.27
Authorship	shakespeare→modern	65.87	73.40
	modern→shakespeare	42.43	45.14
Sentiment	positive→negative	73.12	77.93
	negative→positive	50.28	54.73

Table 6: Comparison of different layer selection strategies between Dola and our approach.

E Style Layers vs. Dola Layers

In Section 3.3, our method selects the style layers, specifically the last few layers of the LLMs, to decoding from contrasting against the final layer. In contrast, Dola selects the early layers to decode by contrasting the final layer. To verify the superiority of our selected style layers, we conduct a comparison experiment, the results of which are shown in Table 6. We can clearly observe the superiority of selecting the last few layers for contrastive decoding in the TST task.

F Different content preservation metrics

In Table 2, we find that our method is not optimal in content preservation. To verify whether this phenomenon occurs with other content preservation metric, we conduct a comparison experiment and present the results in Table 11. We observe the same conclusion as in Table 2, namely, our method is inferior to the baseline method in terms of meaning preservation. For a detailed analysis, please refer to Section 5.

G Different decoding strategy

In Section 3.3, we present a decoding strategy for contrasting style layers. To verify the advantages of this decoding strategy, we compare it with two additional decoding methods: nucleus sampling and contrastive search. As shown in Table 12, our decoding method outperforms the others. This is primarily because contrastive search focuses on the isotropy of token representations during decoding, which means that the semantically similar words have less variation in the representation space and their probabilities should be increased. However, this does not align with the goal of the TST task, which aims to expose more target-style words. Source-style words and target-side style words are actually similar in representation. For example, in the emotion task, "like" and "hate" are semantically different but similar in the embedding space because both represent an emotion, making it difficult to distinguish between these words using isotropy at the representation level.

In addition, nucleus sampling (NP) is a decoding method by setting a threshold p and then restricting the sampling to the set of most probable tokens with cumulative probability less than p. NP is not suited for TST because after deactivating the style neurons at the source side, the probability distribution of the words is changed. The probability of all words in the target style becomes higher, resulting in candidate words predominantly being in the target style. This can cause issues with fluency, as words in the target style are not always meant to be revealed in every context.

Benchmark	**Dataset**	**Tasks**	**Size**		
			train	vald	test
Politeness	Politness (Madaan et al., 2020)	impolite ↔ polite	100k	2000	2000
Toxicity	ParaDetox (Logacheva et al., 2022)	toxic ↔ neutral	18k	2000	2000
Formality	GYAFC (Rao and Tetreault, 2018)	informal ↔ formal	52k	500	500
Authorship	Shakespeare (Xu et al., 2012)	shakespeare ↔ modern	27k	500	500
Politics	Political (Voigt et al., 2018)	democratic ↔ republican	100k	1000	1000
Sentiment	Yelp (Shen et al., 2017)	positive ↔ negative	100k	1000	1000

Table 7: Data statistics on six benchmarks containing the size of train/valid/test set and transfer task we evaluated.

Benchmark	**Source**
Politeness	`https://huggingface.co/Genius1237/xlm-roberta-large-tydip`
Toxicity	`https://huggingface.co/s-nlp/roberta_toxicity_classifier`
Formality	`https://huggingface.co/s-nlp/xlmr_formality_classifier`
Authorship	`https://huggingface.co/notaphoenix/shakespeare_classifier_model`
Politics	`https://huggingface.co/m-newhauser/distilbert-political-tweets`
Sentiment	`https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english`

Table 8: Classifiers used to evaluate the accuracy of style transfer.

Instruction: Please transfer the following informal style sentence into a formal style sentence and maintain the meaning of the sentence.
Input: the movie The In-Laws not exactly a holiday movie but funny and good!
Output: **The** movie "The In-Laws" **is** not exactly a holiday movie, but **it is** funny and good!

	The	movie	"	The	In	-L	aws	"	**is**	not	exactly	a	holiday	movie	,	but	**it**	**is**	funny	and	good	!
0	2.35	2.93	3.18	3.70	3.80	3.53	5.44	3.76	3.59	3.92	3.70	4.02	4.02	4.29	3.86	4.03	3.95	4.28	3.89	4.37	4.12	3.04
1	2.35	2.93	3.18	3.68	3.79	3.52	5.42	3.74	3.58	3.91	3.70	4.02	4.01	4.28	3.86	4.02	3.93	4.27	3.87	4.36	4.11	3.04
2	2.35	2.93	3.18	3.68	3.79	3.52	5.43	3.74	3.58	3.91	3.70	4.02	4.01	4.28	3.86	4.01	3.92	4.26	3.87	4.36	4.10	3.04
3	2.35	2.93	3.18	3.68	3.79	3.52	5.43	3.74	3.58	3.91	3.68	4.02	4.01	4.27	3.87	4.01	3.92	4.26	3.87	4.36	4.10	3.04
4	2.35	2.93	3.18	3.68	3.79	3.52	5.44	3.74	3.59	3.92	3.68	4.01	4.02	4.27	3.87	4.01	3.92	4.26	3.87	4.36	4.10	3.04
5	2.35	2.93	3.18	3.67	3.79	3.52	5.44	3.75	3.59	3.90	3.68	4.02	4.02	4.28	3.87	3.99	3.91	4.26	3.87	4.37	4.10	3.05
6	2.35	2.92	3.18	3.67	3.79	3.52	5.42	3.74	3.60	3.91	3.68	4.01	4.01	4.27	3.87	3.99	3.91	4.26	3.88	4.35	4.10	3.04
7	2.35	2.92	3.18	3.67	3.78	3.50	5.43	3.75	3.58	3.90	3.67	3.99	4.01	4.28	3.87	3.99	3.90	4.24	3.87	4.33	4.10	3.04
8	2.36	2.92	3.18	3.67	3.79	3.50	5.42	3.76	3.58	3.90	3.67	3.98	4.01	4.27	3.86	3.99	3.91	4.24	3.89	4.35	4.11	3.06
9	2.36	2.92	3.18	3.67	3.77	3.50	5.42	3.75	3.58	3.90	3.68	3.99	4.01	4.26	3.86	4.01	3.91	4.24	3.87	4.33	4.11	3.05
10	2.35	2.91	3.17	3.66	3.77	3.49	5.45	3.74	3.58	3.90	3.68	3.98	4.01	4.26	3.86	3.99	3.90	4.23	3.89	4.34	4.11	3.05
11	2.35	2.91	3.16	3.65	3.76	3.48	5.44	3.75	3.58	3.89	3.68	3.97	4.00	4.27	3.86	3.99	3.89	4.23	3.89	4.35	4.12	3.05
12	2.36	2.93	3.16	3.65	3.76	3.49	5.44	3.74	3.58	3.90	3.67	3.97	4.00	4.26	3.85	3.99	3.89	4.22	3.89	4.36	4.12	3.05
13	2.35	2.93	3.17	3.66	3.76	3.50	5.44	3.73	3.58	3.89	3.68	3.97	4.02	4.27	3.89	3.98	3.89	4.23	3.91	4.35	4.12	3.06
14	2.34	2.91	3.15	3.64	3.76	3.50	5.46	3.71	3.58	3.87	3.67	3.98	4.01	4.27	3.86	3.96	3.87	4.21	3.90	4.33	4.10	3.05
15	2.34	2.90	3.14	3.62	3.78	3.50	5.44	3.71	3.55	3.86	3.66	3.97	4.01	4.26	3.84	3.93	3.87	4.20	3.91	4.31	4.11	3.04
16	2.34	2.87	3.11	3.62	3.74	3.49	5.39	3.72	3.52	3.81	3.61	3.92	3.96	4.23	3.81	3.87	3.84	4.18	3.87	4.24	4.05	3.03
17	2.32	2.87	3.08	3.61	3.71	3.46	5.39	3.71	3.53	3.79	3.60	3.89	3.93	4.21	3.79	3.74	3.81	4.17	3.85	4.22	4.02	2.99
18	2.31	2.84	3.02	3.56	3.64	3.45	5.39	3.67	3.46	3.68	3.54	3.83	3.86	4.16	3.75	3.71	3.77	4.12	3.80	4.13	3.96	2.97
19	2.30	2.80	3.00	3.53	3.61	3.42	5.38	3.62	3.41	3.62	3.47	3.78	3.84	4.13	3.71	3.67	3.68	4.07	3.77	4.09	3.92	2.93
20	2.26	2.77	2.96	3.50	3.55	3.39	5.36	3.60	3.37	3.63	3.43	3.73	3.80	4.09	3.68	3.58	3.62	4.01	3.76	4.04	3.89	2.91
21	2.23	2.74	2.92	3.46	3.50	3.39	5.33	3.60	3.30	3.50	3.39	3.65	3.78	4.04	3.62	3.44	3.58	3.95	3.72	3.93	3.84	2.87
22	2.19	2.68	2.87	3.40	3.45	3.35	5.31	3.49	3.25	3.36	3.25	3.58	3.50	3.96	3.56	3.35	3.40	3.86	3.62	3.87	3.74	2.84
23	2.14	2.57	2.80	3.33	3.35	3.33	5.27	3.44	3.15	3.28	3.11	3.47	3.34	3.88	3.49	3.25	3.28	3.73	3.54	3.77	3.61	2.81
24	2.10	2.43	2.73	3.27	3.25	3.30	5.26	3.39	3.06	3.14	2.96	3.36	3.22	3.72	3.42	3.08	3.14	3.61	3.36	3.71	3.53	2.75
25	2.07	2.37	2.60	3.22	3.16	3.25	5.24	3.33	2.96	3.02	2.77	3.22	2.71	3.65	3.34	3.03	3.00	3.54	3.20	3.60	3.38	2.70
26	2.06	2.29	2.56	3.18	3.14	3.17	5.19	3.31	2.88	2.93	2.65	3.14	2.59	3.41	3.30	2.91	2.93	3.45	3.09	3.52	3.28	2.66
27	1.98	2.15	2.46	3.13	3.09	3.15	5.16	3.20	2.72	2.80	2.57	2.98	2.46	3.24	3.11	2.77	2.80	3.26	2.96	3.39	3.17	2.56
28	1.94	2.07	2.36	3.09	2.96	3.05	5.17	3.08	2.53	2.72	2.60	2.84	2.68	3.15	2.98	2.50	2.69	3.07	2.87	3.22	3.04	2.46
29	1.85	1.95	2.13	2.86	2.81	2.79	5.09	2.91	2.30	2.54	2.32	2.52	2.49	3.05	2.72	2.25	2.48	2.84	2.78	2.92	2.84	2.26
30	1.84	1.93	1.99	2.80	2.52	2.41	4.87	2.87	2.19	2.34	2.17	2.32	2.35	3.04	2.55	2.09	2.31	2.74	2.62	2.77	2.74	2.21
31	1.57	1.69	1.62	2.30	2.24	2.23	4.51	2.31	1.94	2.10	2.05	2.19	2.27	2.74	2.08	1.92	2.05	2.46	2.30	2.10	2.56	1.86

Table 9: JSD (scaled by 10^5) between the final layer and all previous layer in LLaMA-3. Each row represents the distance between all previous layers and the final layer, while each column corresponds to the token generated at each decoding step. Example taken from the TST task to transfer from informal style to formal style. The 0-th layer is the embedding layer.