

Investigating Conceptual Blending of Diffusion Models for Improving Nonword-to-Image Generation

Anonymous Author(s)

A SUPPLEMENTARY MATERIALS

A.1 Dataset Creation

A.1.1 EvalNouns1000. *EvalNouns1000* is a list of 1,000 nouns created by our paper as evaluation data. This wordlist is used in our paper mainly for the following two purposes:

- Used to create 10,000 matching and 10,000 mismatching pairs in Section 3.
- Used to create 1,000 interpolated embeddings in Section 4.3.

As mentioned in our paper, these nouns are randomly taken from the MRC Psycholinguistic Database [2] with the restrictions of word imageability and frequency. This database provides an imageability score for each noun ranging from 100 to 700, where a high score indicates that the noun is highly imageable. Although it also provides word frequency scores, we do not use them but instead use the Python package *wordfreq* [24]. This is to ensure *EvalNouns1000* to be a subset of another dataset *TrainWords26143*, which will be described later. Selecting words with 500 or more imageability scores and 3.5 or more Zipf frequency values resulted in 1,183 words in total. *EvalNouns1000* is created by randomly sampling 1,000 words from these words.

A.1.2 TrainWords26143. *TrainWords26143* is a list of 26,143 words compiled by an existing study on nonword-to-image generation [9–11]. Our paper uses this wordlist mainly for the following three purposes:

- Used by the proposed embedding space conversion method to create anchors of k -nearest neighbor search and linear regression.
- These anchors are also used for calculating Spearman’s rank correlation metrics in Section 4.3.
- A minor-modified one is used to train a comparative Multi-Layer Perceptron (MLP).

As mentioned in our paper, the existing study created this wordlist using the Spell Checker Oriented Word Lists (SCOWL)¹ and the 26,143 words were selected based on word frequency and pronunciation availability. Specifically, a Python package *wordfreq* [24] was used to remove words having Zipf frequency less than 3.0. Also, the Carnegie Mellon University (CMU) dictionary² was looked up for checking the pronunciation availability.

The modified wordlist used to train the MLP consists of 26,455 words, which was created by adding 312 words filtered out during the pronunciation availability check.

A.1.3 Training Data of NonwordCLIP. To train a NonwordCLIP [9–11] in Section 4, we constructed a dataset in which each word appears almost an equal number of times. As mentioned in our paper, the dataset consists of 5,496 highly-imageable and -frequent

nouns and noun phrases created by combining the MRC Psycholinguistic Database [2], *wordfreq* [24], and an English lexical database WordNet [13].

First, from the MRC database, we collected highly-imageable nouns having an imageability score of 500 or more. Next, we used WordNet to augment the vocabulary based on Liu et al. [28]’s procedure, in which synonym and hyponym relationships on WordNet were used to extend the imageability dictionary. Specifically, for each noun in an imageability dictionary, their method propagated the same imageability score to the synonyms and hyponyms of the noun. Following this policy, for each word in our imageable noun list, we propagated its imageability score to its 1st, 2nd, and 3rd synonym nouns and all hyponym nouns. Natural Language ToolKit (NLTK) [27] was used to access the WordNet hierarchy and to judge whether each WordNet node is a noun or a noun phrase. After this augmentation, we used *wordfreq* to obtain nouns having 3.5 or more Zipf frequency values.

Lastly, we further augmented the dataset twice using the two prompts “<WORD>” and “a photo of a <WORD>”, resulting in training data of 10,992 samples.

A.2 Prompt Engineering for Calculating CLIP Score

In Section 3.2, Contrastive Language-Image Pretraining (CLIP) score [15] was calculated to detect the presence of a single concept in an image. To increase the precision of the scores, we adopted prompt engineering like the one adopted in the original paper [15] to solve an image classification task³. The original paper used 80 templates describing images containing a target concept, all of which ends with a period, such as “a bad photo of a <WORD>.”. Our paper increased the number of templates to 160 by creating a variant without the period in the ending position for each template, such as “a bad photo of a <WORD>”.

For each pair of a concept and an image, we calculated the final CLIP score by averaging the 160 CLIP similarity scores computed for each prompt.

A.3 Detailed Experimental Results

A.3.1 Results under Different ns . Tables 5, 6, and 7 show the ratios of conceptual blending evaluated in Section 3 under different ns . Our conclusions mentioned in the paper are consistent throughout all ns , while the ratio decreases as n increases because setting a larger n makes the detection criterion more strict.

A.3.2 Results under Different ℓ s. Table 8 shows the transition of the rank correlation metric used in Section 4.3 with different ℓ s. The hyperparameter ℓ denotes how many nearest-neighbor embeddings in both the CLIP pooled and last-hidden-state embedding

¹<http://wordlist.aspell.net/> (Accessed April 9, 2024)

²<https://github.com/menelik3/cmudict-ipa/> (Accessed April 9, 2024)

³https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb (Accessed April 9, 2024)

Table 5: Ratios of respective cases when inputting interpolated embeddings between concepts A and B with different interpolation ratios measured under the setting $n = 1$.

Case	Interpolation Ratio of Concept A to Concept B									Total
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
Concept A	0.286	0.466	0.580	0.855	0.974	0.991	1.000	1.000	1.000	0.802
Concept B	1.000	1.000	1.000	1.000	0.957	0.870	0.648	0.495	0.496	0.829
BCD	0.286	0.466	0.562	0.744	0.819	0.778	0.600	0.465	0.487	0.584
MCD	0.286	0.466	0.580	0.855	0.931	0.861	0.648	0.495	0.496	0.631

Table 6: Ratios of respective cases when inputting interpolated embeddings between concepts A and B with different interpolation ratios measured under the setting $n = 2$ (Same as the results reported in our paper).

Case	Interpolation Ratio of Concept A to Concept B									Total
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
Concept A	0.152	0.311	0.411	0.709	0.948	0.981	1.000	1.000	1.000	0.732
Concept B	1.000	1.000	1.000	0.991	0.914	0.731	0.472	0.277	0.265	0.738
BCD	0.143	0.301	0.348	0.521	0.621	0.593	0.416	0.257	0.257	0.389
MCD	0.152	0.311	0.411	0.701	0.862	0.722	0.472	0.277	0.265	0.471

Table 7: Ratios of respective cases when inputting interpolated embeddings between concepts A and B with different interpolation ratios measured under the setting $n = 5$.

Case	Interpolation Ratio of Concept A to Concept B									Total
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
Concept A	0.010	0.107	0.134	0.291	0.716	0.898	0.984	1.000	1.000	0.578
Concept B	1.000	1.000	0.982	0.897	0.698	0.380	0.208	0.099	0.053	0.587
BCD	0.000	0.097	0.116	0.162	0.302	0.204	0.184	0.099	0.053	0.138
MCD	0.010	0.107	0.134	0.214	0.466	0.306	0.208	0.099	0.053	0.181

spaces are used to calculate the rank correlation. As a reference, we also measured the rank correlation metric between the nearest-neighbor ranking for the ground-truth interpolated embedding in the pooled embedding space and that for the ground-truth interpolated embedding in the last-hidden-state embedding space, averaged over all samples. This metric, shown as “Ground Truth” in the table, measures the alignment of the sample distributions in the two embedding spaces.

The results in the table indicate that the comparative MLP-based method yielded higher correlations than the proposed method under a large ℓ , and they are even higher than the metrics measured using the ground-truth interpolated embeddings presumably due to the curse of dimensionality.

A.4 Image Generation Examples

A.4.1 Image Generation from Interpolated Embeddings. Figures 8 and 9 showcase images generated using the pretrained Stable Diffusion from the embeddings interpolating between the embeddings of two concepts. The figures include results generated from embeddings computed by different text embedding space conversion methods evaluated in Section 4.3. Also, the images generated from the ground-truth last-hidden-state embeddings used in Section 3 are attached at the bottom of each figure.

Table 8: Spearman’s rank correlation under different ℓ s. For all metrics, a higher score indicates a higher consistency in neighborhood relationships before and after the embedding space conversion.

Method	RCorr $_{\ell=2}$	RCorr $_{\ell=5}$	RCorr $_{\ell=10}$	RCorr $_{\ell=100}$	RCorr $_{\ell=26143}$
MLP [9–11]	0.846	0.783	0.711	0.464	0.444
Ours ($k = 1$)	0.902	0.702	0.586	0.352	0.363
Ours ($k = 2$)	0.880	0.788	0.669	0.376	0.365
Ours ($k = 5$)	0.884	0.765	0.735	0.395	0.366
Ours ($k = 10$)	0.882	0.771	0.689	0.397	0.366
Ours ($k = 100$)	0.888	0.781	0.694	0.373	0.365
Ours ($k = 200$)	0.886	0.791	0.697	0.373	0.365
Ours ($k = 300$)	0.890	0.793	0.699	0.373	0.365
Ours ($k = 400$)	0.890	0.797	0.700	0.373	0.365
Ours ($k = 500$)	0.880	0.802	0.701	0.373	0.364
Ours ($k = 1,000$)	0.882	0.799	0.696	0.367	0.359
Ground Truth	0.868	0.809	0.703	0.373	0.365

As confirmed in the evaluations in our paper, the figures indicate that

- both the MLP-based comparative and proposed methods can depict blended concepts,
- the image generation quality of the proposed method is better than that of the MLP-based comparative method, and
- the proposed method generates images almost identical to the ones generated from the ground-truth interpolated embeddings.

A.4.2 Nonword-to-Image Generation. Figure 10 showcases more nonword-to-image generation results generated by different methods used in the evaluation in Section 4.4. As mentioned in Section 4.4.2, these nonwords are taken from Sabbatino et al. [19]’s work, in which they annotated evoked emotion labels to each of them.

REFERENCES

- [27] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc., Sebastopol, CA, USA.
- [28] Ting Liu, Kit Cho, G. Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz, Ignacio Cases, and Ching-sheng Lin. 2014. Automatic expansion of the MRC psycholinguistic database imageability ratings. In *Proc. 9th Int. Conf. Lang. Resour. Eval.* European Language Resources Association (ELRA), Reykjavik, Iceland, 2800–2805.

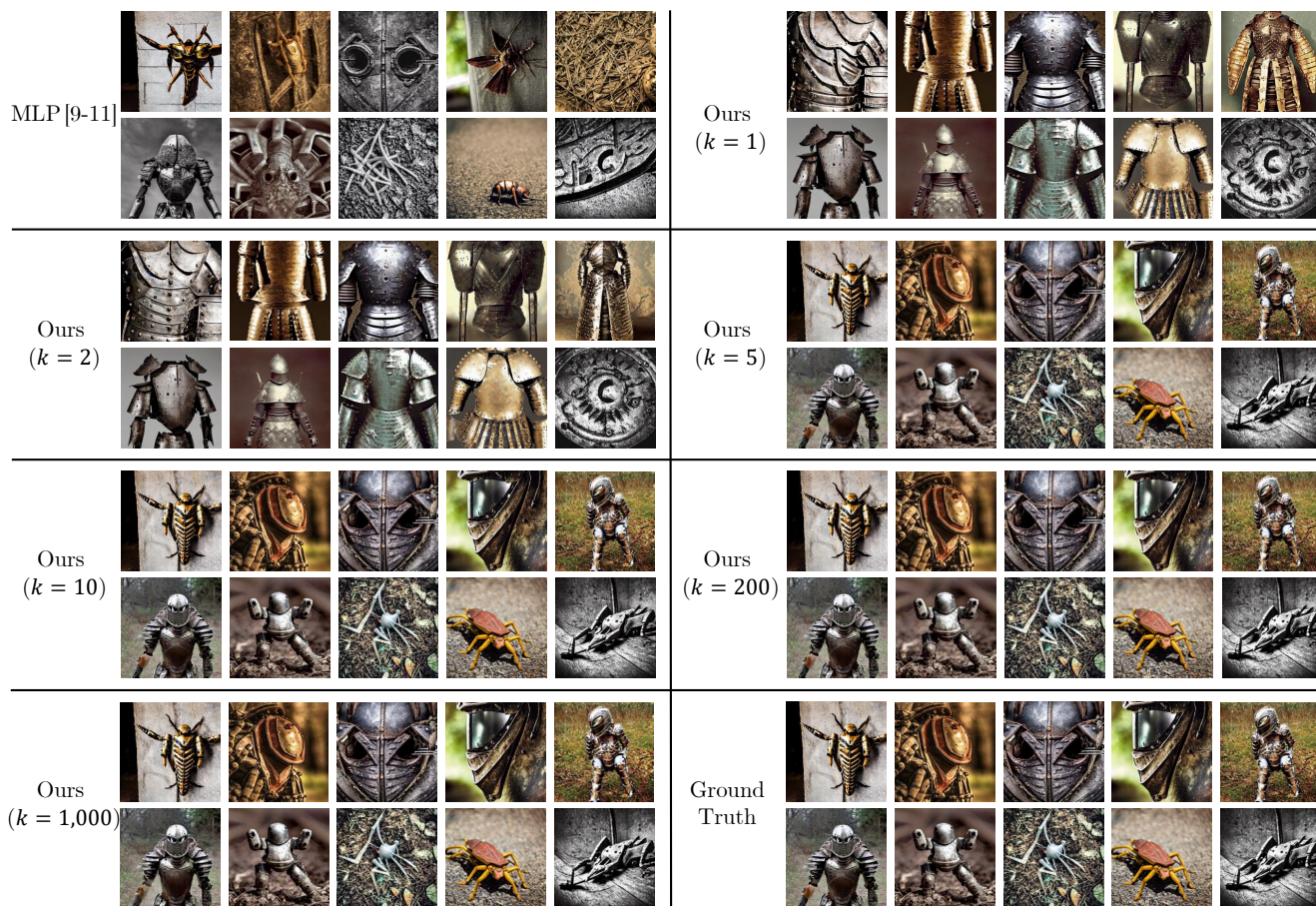


Figure 8: Image generation results generated from interpolated embeddings between Concept A = “*armour*” and Concept B = “*spider*” with an interpolation ratio = 0.6 using different methods.

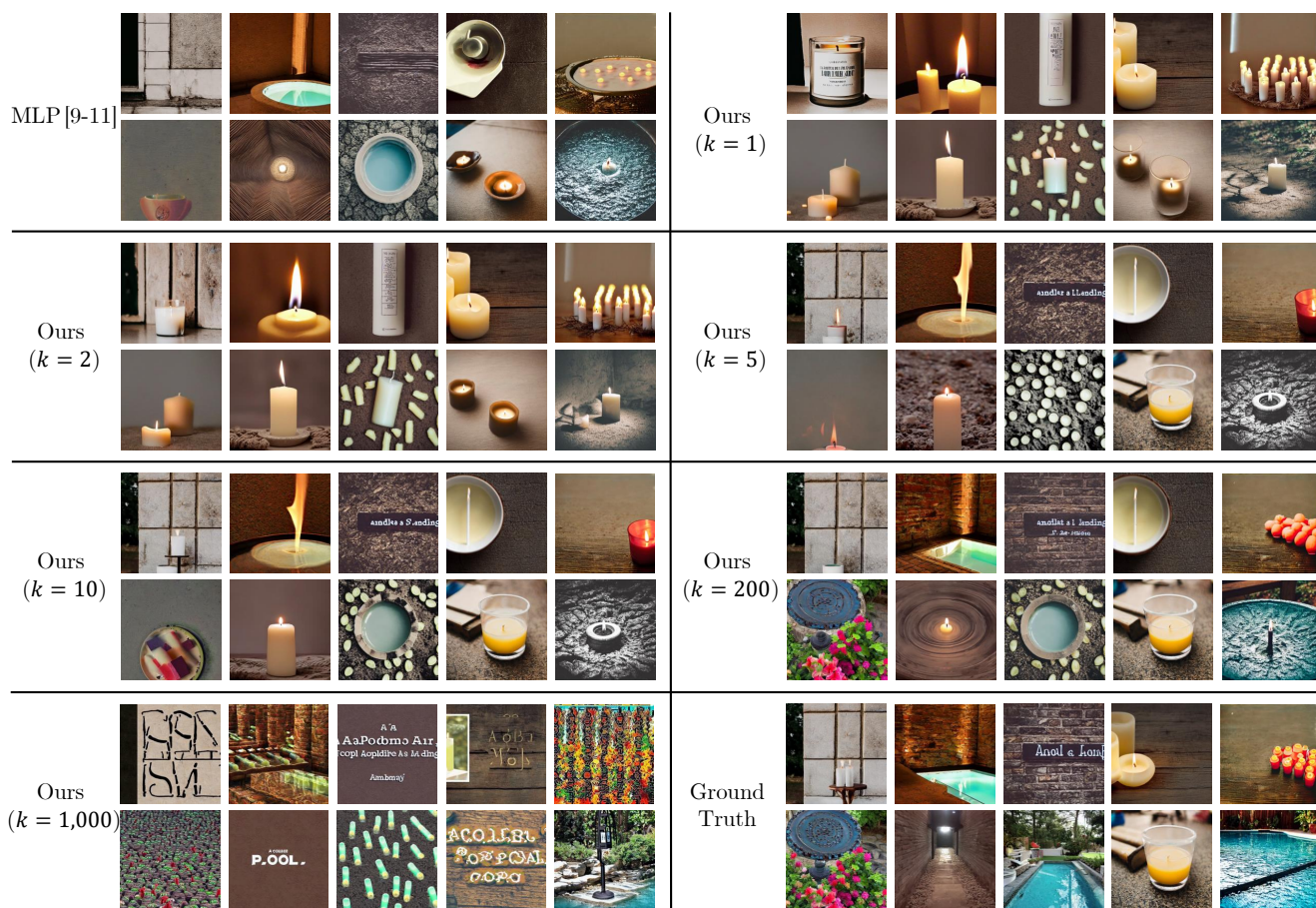


Figure 9: Image generation results generated from interpolated embeddings between Concept A = “pool” and Concept B = “candle” with an interpolation ratio = 0.5 using different methods.

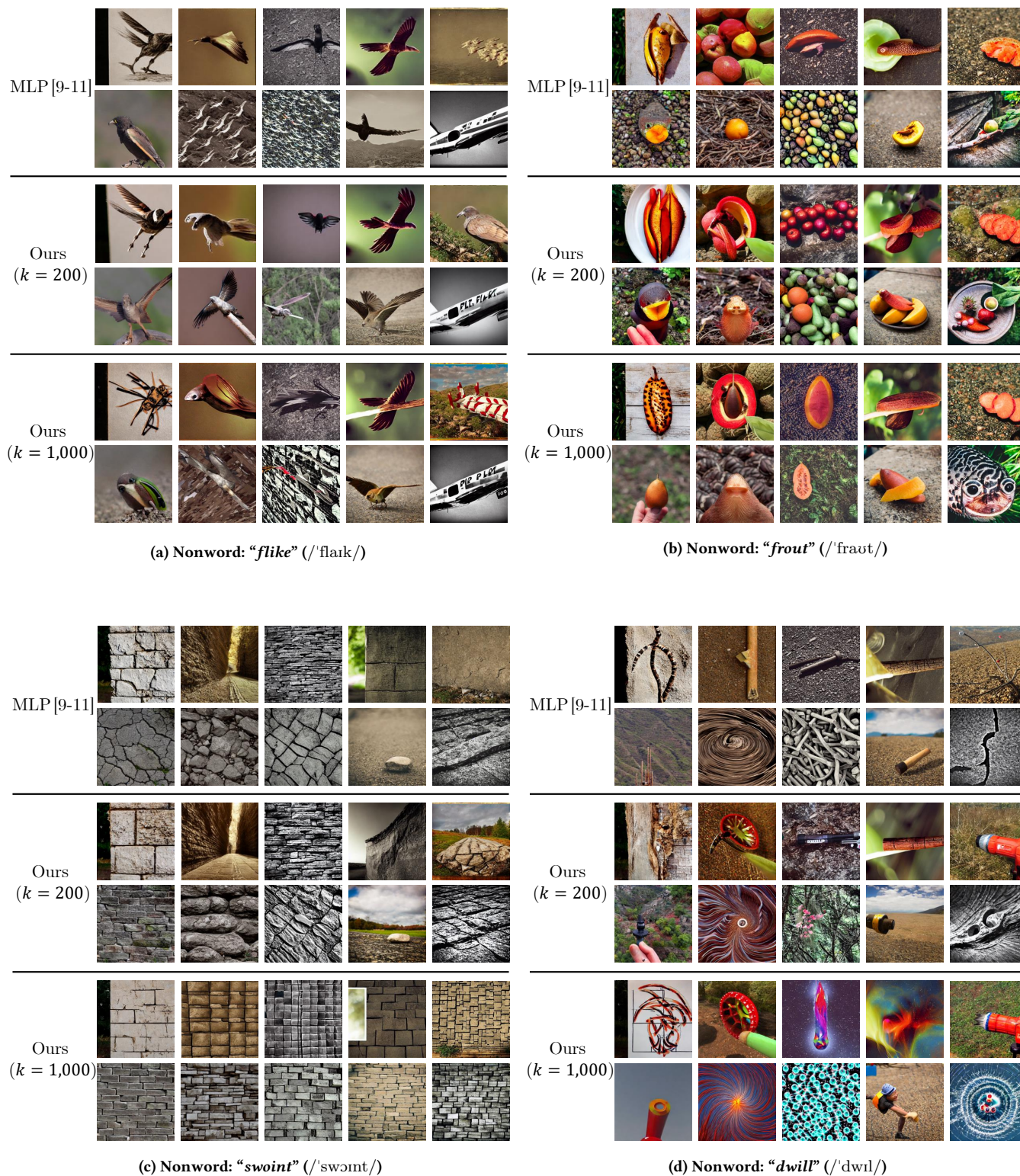


Figure 10: More nonword-to-image generation results generated using different methods.