# Appendix

## 1 Back-imagination and Back-speech

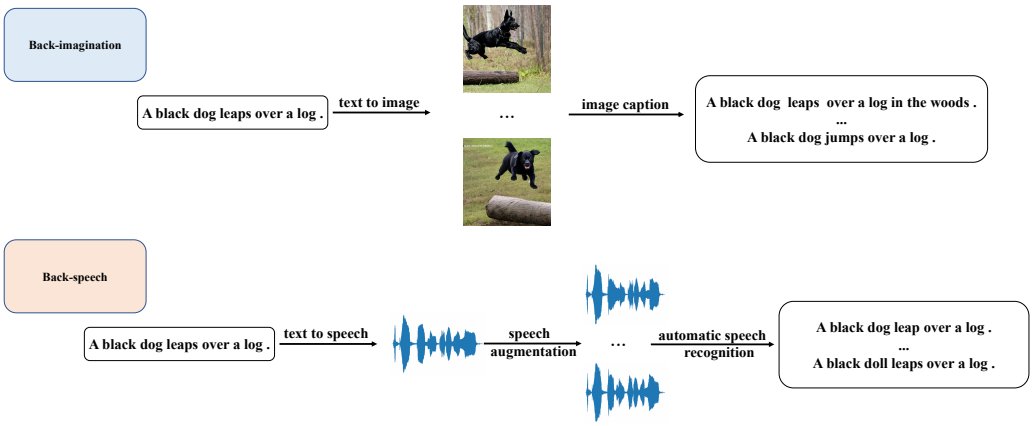


Figure 1: The illustrative examples for two proposed techniques: Back-imagination and Back-speech.

As shown in Figure 1, we present illustrative examples to facilitate a better understanding of two proposed techniques: Back-imagination and Back-speech.

## 2 Datasets

Tiny ImageNet [Le and Yang, 2015] serves as a compact version of the comprehensive ImageNet dataset. It comprises 100,000 images spanning 200 classes, with 500 images per class, and these images are downsized to 64×64 pixels. Each class is furnished with 500 training images, 50 validation images, and 50 test images.

The Stanford Sentiment Treebank-2 (SST-2) [Socher et al., 2013] is a sentiment classification dataset populated with movie reviews gathered from Rotten Tomatoes, paired with their corresponding binary labels. The dataset is partitioned into training, validation, and testing sets, comprising 67,349, 872, and 1,821 instances, respectively.

Given the scarcity of datasets for understanding natural language in visual scenes, we introduce a novel textual entailment dataset, named Textual Natural Contextual Classification (TNCC). This dataset is formulated on the foundation of Crisscrossed Captions [Parekh et al., 2020], an image captioning dataset supplied with human-rated semantic similarity ratings on a continuous scale from 0 to 5. We tailor the dataset to suit a binary classification task. Specifically, sentence pairs with annotation scores exceeding 4 are categorized as positive (entailment), whereas pairs with scores less than 1 are marked as negative (non-entailment). The TNCC dataset is partitioned into training, validation, and testing sets, containing 3,600, 1,200, and 1,560 instances, respectively. This dataset will be made available alongside our source codes.

## 3 Configurations

In this work, we employ a uniform experimental configuration for both textual entailment and sentiment classification tasks. We adopt BERT-BASE [Devlin et al., 2018], a model pretrained using Masked Language Modeling (MLM), as our primary experiment subject. For each individual downstream classification experiment, the classification model is initialized with the pretrained parameters from the BERT-BASE model. The classifier component, comprising of two fully connected layers that deduce class labels from the output embeddings generated by the transformer architectures, is randomly initialized. During the training phase, we leverage the Adam optimization algorithm with a learning rate set at $5e-5$, the first and second momentum terms, $\beta_1$ and $\beta_2$, are respectively set to 0.9 and 0.999. Additionally, we introduce an $L_2$ weight decay of 0.01 to the model. We select a batch

size of 2 for all trials. We save model checkpoints during training and ultimately employ the best checkpoint—determined based on performance on the validation dataset—for testing. The results are presented as classification accuracies on both datasets under investigation.

For the image classification task, we employ the ResNet18 [He et al., 2015] model, which is considered more suitable for small datasets. We initialize all learnable layer parameters randomly. During the training process, we employ the SGD optimizer with a learning rate of $0.1$, momentum of $0.9$, and a weight decay of $0.0001$.

## 4 Human Evaluation on Augmented Samples

In response to your suggestion, we conducted a human evaluation on the sampled augmented data. The results of the evaluation are as follows:

**For the images generated using the back-captioning method:**

- Label Invariance Score: 99.2%

**For the sentences generated using the back-imagination method:**

- Semantic Consistency Score: 98.8%

These high scores indicate that both methods performed exceptionally well in their respective evaluations. The results affirm that Back-Modality preserves the essential characteristics of the original data while introducing diversity, further validating our approach.

## 5 More Choices of Cross-Modal Generation Models

In our paper, for the Back-captioning with a 10-shot setting, we primarily used the OFA-large model, which yielded a top-1 accuracy of 20.07%. To assess the impact of different model sizes on the outcomes, we also conducted experiments with OFA-huge under the same conditions. The results showed a significant improvement, with the top-1 accuracy reaching 22.12%.

## 6 Cost of Obtaining the Augmented Samples

| Method | Additional Computational Time |
|---|---|
| RandErasing | 4 m 55 s |
| Puzzle Mix | 1 h 29 m 25 s |
| Alignmixup | 1 h 59 m 45 s |
| Back-captioning (our method) | 11 h 35 m |
| Auto augment | About 49 h |

Table 1: The additional computational overhead of various augmentation methods compared to the base model.

Table1 illustrates the additional computational overhead of various augmentation methods compared to the base model on RTX A6000. The primary cost of our method is related to generating images with the diffusion model, and the primary overhead of auto augment is associated with learning augmentation policies. In terms of text augmentation, on the textual entailment task, our method, back-imagination, took 4 hours 13 m 45 s, while the back-translation method took 5 h 38 m 12 s. On the sentiment analysis task, back-speech took 35 m 27 s, whereas the back-translation method took 5 h 22 m 4 s.

## 7 Discussion About Real-World Applicability

While current multi-modal and cross-modal models have achieved impressive results and continue to rapidly advance, it is worth noting that not all domains currently have easy access to open-source

cross-modal models. This limitation can, to some extent, restrict the effectiveness of our method in real-world applications. However, recent research has increasingly focused on the adaptability of diffusion-based cross-modal models in domains with limited data. This research encompasses areas such as few-shot [Giannone et al., 2022], one-shot [Wu et al., 2023], zero-shot [Li et al., 2023], domain adaptation [Kim et al., 2023], and unsupervised domain adaptation [Benigmim et al., 2023]. These research directions will further expand the real-world applicability boundaries of our method.

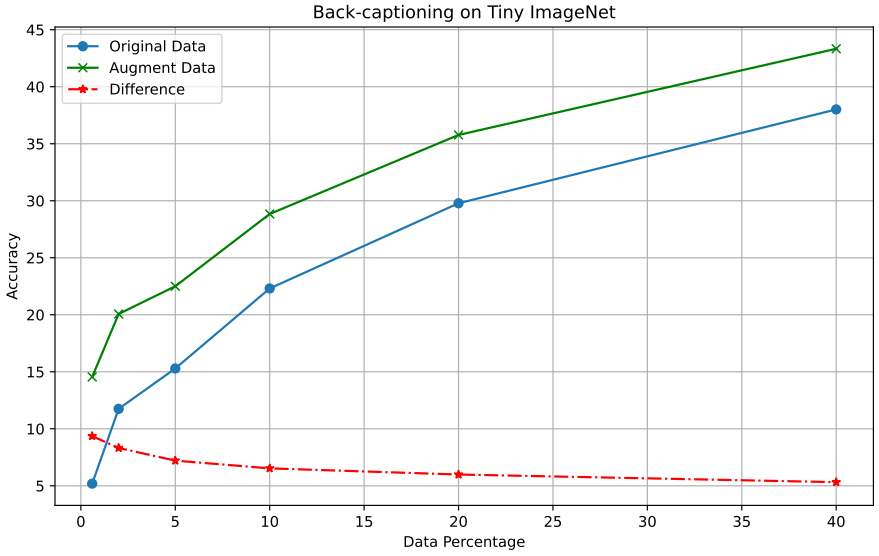


Figure 2: Effect of increasing original data.

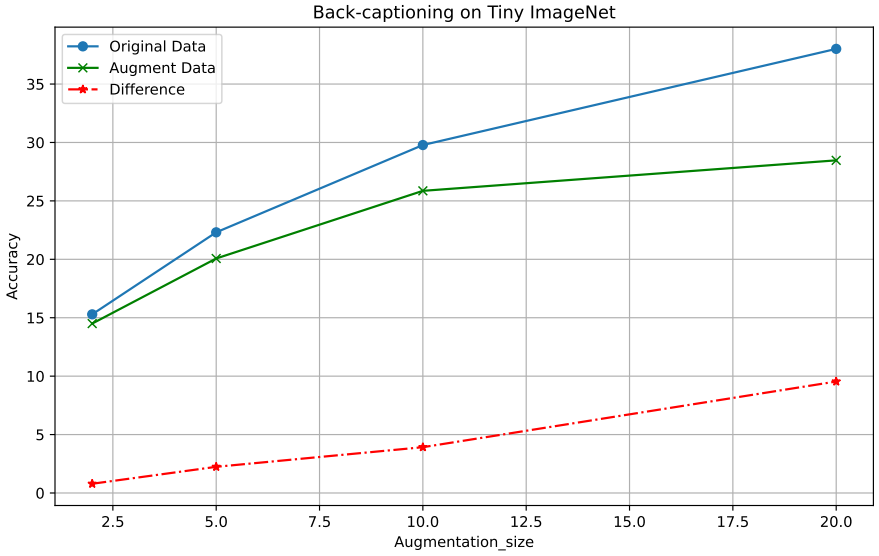


Figure 3: Effect of increasing augmentation multiple.

## 8 Proportion and Size

We conducted an analysis of the proportion of augmented data and the augmentation size by sampling a subset of the data. Figure 2 illustrates the variation in the model's performance when increasing the volume of original data while keeping the augmentation multiple at 5. From the curve in this figure, it is evident that as the volume of original training data increases, augmented data continues to provide benefits. However, these benefits exhibit diminishing returns as more original data is added. Figure 3 addresses the aspect of extending data generation. It demonstrates that as the augmentation size increases, the model's performance improves. However, after a certain point, the gains tend to plateau. We believe the primary reason for the phenomena observed in these two figures is that the diversity introduced by augmented data leads to performance gains. However, this diversity may not match the affinity of the original data, which can result in diminishing returns.

## References


Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, 2013. URL https://api.semanticscholar.org/CorpusID:990233.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. *arXiv preprint arXiv:2004.15020*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2022.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *arXiv preprint arXiv:2304.02742*, 2023.

Gwanghyun Kim, Ji Ha Jang, and Se Young Chun. Podia-3d: Domain adaptation of 3d generative model across large domain gap using pose-preserved text-to-image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22603–22612, 2023.

Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. One-shot unsupervised domain adaptation with personalized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 698–708, 2023.