PARTIAL ALIGNMENT OF REPRESENTATIONS VIA INTERVENTIONAL CONSISTENCY

Felix Leeb
Max Planck Institute for Intelligent Systems
fleeb@tue.mpg.de

Satoshi Hayakawa
Sony Group Corporation

Yuhta Takida
Sony AI

Yuki Mitsufuji
Sony AI, Sony Group Corporation

ABSTRACT

Multimodal representation learning aims to integrate diverse data modalities into a shared embedding space with a common approach to use contrastive learning. However, this approach is limited by the need for large amounts of paired data, sensitivity to data quality, and lack of scalability when introducing new modalities. We propose INTERVENTIONAL CONSISTENCY (ICON), a novel framework for learning structured representations that achieve partial alignment across modalities using unpaired annotated samples. The key is to align the annotation-specific information in the latent space by enforcing the consistency of controllable and recognizable semantic interventions across modalities. We demonstrate that our method is able to align representations sufficiently to achieve competitive results on a novel retrieval task we introduce called label-retrieval. Furthermore, when pre-training a model with ICON, and then fine-tuning it with a small amount of paired data using CLIP, we achieve comparable retrieval performance with 2-4x fewer samples, thereby alleviating the need for paired data to learn multi-modal representations. The code used in the paper is available at https://github.com/sony/alice.

1 INTRODUCTION

Intelligent systems must process and integrate information across diverse modalities such as images, text, audio, and video. A prevailing approach to this challenge is multimodal representation learning (Manzoor et al., 2023), which aims to create a shared embedding space where semantic information from different modalities can be aligned. This shared space facilitates tasks such as cross-modality retrieval, classification, and generation. One popular technique is contrastive learning (Chen et al., 2020), as exemplified by methods like CLIP (Radford et al., 2021), where embeddings for paired samples (e.g., image–caption pairs) are aligned directly using a contrastive objective.

However, contrastive learning methods exhibit a variety of significant limitations (Zhai et al., 2023). They require large amounts of high-quality paired data across all modalities of interest, which may be difficult to obtain, particularly beyond modalities like text and images (Zhu et al., 2024). Additionally, these methods are sensitive to the quality of pairings (e.g., poorly annotated captions (Betker et al., 2023)) and suffer from information asymmetry—some modalities, like images, often encapsulate far richer information than others, such as corresponding captions, thereby limiting the fidelity of embeddings to the weakest link (Fan et al., 2023). Finally, adding new modalities to a pre-trained model requires retraining from scratch, making it challenging to scale multimodal systems (Tejankar et al., 2022).

To address these issues, we explore a fundamentally different learning method to achieve useful structured representations: INTERVENTIONAL CONSISTENCY (ICON). Conceptually, ICON aims to align the semantic information in the representation space based on a shared definition of how that information can be selectively manipulated. While this only partially aligns the information across modalities up to the shared annotations in each modality, it provides a flexible and scalable approach to multimodal representation learning. Our contributions in this paper are as follows:

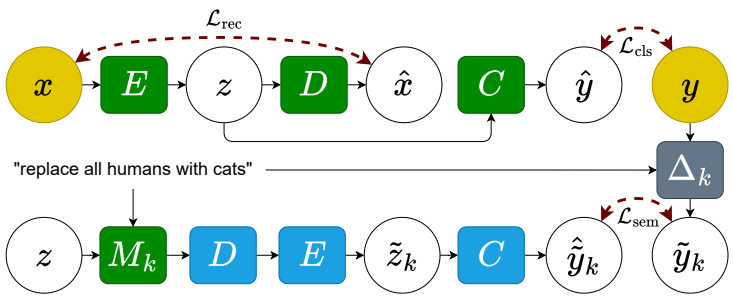

Figure 1: Overview of the four model components (green) and three training objectives (red arrows) of ICON: starting only from the observations x and labels y (shown in yellow), ICON trains the encoder E, decoder D, and classifier C using the reconstruction $\mathcal{L}_{\text{rec}}$, classification $\mathcal{L}_{\text{cls}}$. The semantic $\mathcal{L}_{\text{sem}}$ loss is used to train the intervention module M for which the other components are frozen (shown in blue), and the Δ_k (untrained) which defines specific semantic interventions on the labels.

1. We introduce a novel method for multimodal representation learning that achieves partial alignment across modalities using only label information in individual modalities.
2. We explore how the structure of representations of ICON enables controllable manipulation of the semantic information in the latent space for different intervention designs.
3. We evaluate ICON's effectiveness as a pretraining step before contrastive learning with paired samples to reduce the amount of paired data required.
4. We introduce a variant of the conventional cross-modality retrieval task called label-retrieval to evaluate extent to which label information is aligned across modalities.

2 Interventional Consistency

We begin by describing how we learn a highly informative representation for a single modality with a very specific structure that can be exploited to partially align the embeddings across modalities. Given a set of samples $x \in \mathcal{X}$ in a single modality (e.g. images) and their corresponding semantic annotations/labels $y \in \mathcal{Y}$, ICON learns a latent representation $z \in \mathcal{Z}$ using an encoder $E : \mathcal{X} \to \mathcal{Z}$, a decoder $D : \mathcal{Z} \to \mathcal{X}$, a classifier $C : \mathcal{Z} \to \mathcal{Y}$, and an intervention module $M : \mathcal{Z} \to \mathcal{Z}$ using three training objectives. The encoder and decoder are trained to reconstruct the input samples $\mathcal{L}_{\text{rec}} = D(z) \leftrightarrow x$, while the encoder and classifier are trained to predict the labels from the latent samples $\mathcal{L}_{\text{cls}} = C(z) \leftrightarrow y$ where $\leftrightarrow$ represent appropriate metric.

Note that the reconstruction and classification objectives used to train the encoder, decoder, and classifier are based entirely on the observational data $\{x, y\}$. For the intervention module only, we implicitly define interventional distributions based on how semantic interventions $\Delta_k : \mathcal{Y} \to \mathcal{Y}$ (not trained) modify the annotations y to $\tilde{y}_k = \Delta_k(y)$. For example, if the annotations correspond to animals/objects present in the image, an intervention k could "replace all humans with cats" in an image by making the corresponding change to the object labels. Crucially, we do not need the corresponding counterfactual samples (i.e. the images where all humans are in fact replaced by cats), we only need to define the effects of the interventions in terms of changes to the annotations.

Now the intervention module M is trained to modify the latent samples z when conditioned on a specific desired intervention k to produce a modified latent sample $\tilde{z}_k = M_k(z)$ such that the classifier C predicts the modified annotations $\tilde{y}_k$. To ensure that the interventions are realistic, we further refine the training objective by taking advantage of the autoencoder being trained on the observational data distribution (Radhakrishnan et al., 2019). Consequently, we propose composing the decoder and encoder to project modified latent samples which may deviate from the latent distribution back onto the (observational) latent manifold (Leeb et al., 2022). This results in our *semantic* loss:

$$\mathcal{L}_{\text{sem}} = (C \circ E \circ D \circ M_k \circ E)(x) \leftrightarrow \Delta_k(y) \tag{1}$$

The overall optimization objective is $\mathcal{L} = \mathcal{L}_{\text{sem}}(M_k) + \mathcal{L}_{\text{rec}}(E, D) + \mathcal{L}_{\text{cls}}(E, C)$ where the arguments of each term indicate which components are updated using the gradients from that loss. This means the classifier, encoder, and decoder are frozen when computing the semantic loss.

Cross-Modal Alignment So far, we have described how ICON learns a representation for a single modality with a very specific structure. Let us now explain how we can exploit this to partially align the embeddings across other modalities. First, we use the samples of a chosen reference modality to learn a modality-specific encoder E^{ref}, decoder D^{ref}, classifier C^{ref}, and intervention module M^{ref} using $\mathcal{L}$ as described above.

Next, given a new modality m, we train a new encoder $E^{(m)}$, decoder $D^{(m)}$ using the objective $\mathcal{L}^{(m)} = \mathcal{L}_{\text{rec}}(E^{(m)}, D^{(m)}) + \mathcal{L}_{\text{cls}}(E^{(m)}) + \mathcal{L}_{\text{sem}}(E^{(m)})$ while using the frozen classifier C^{ref} and intervention module M^{ref} from the reference modality. This forces the representation of the new modality to align the information pertraining to the shared annotations in the reference modality to be recognizable and consistent with the classifier C^{ref} and the intervention module M^{ref}.

This process can be repeated for any number of subsequent modalities, each time learning a new modality-specific representation, with universally shared recognizable interventions. Note that there is a subtle caveat that similar annotations must be available in each modality individually for the classification and semantic losses to have an effect in downstream modalities. The more similar the annotations are across modalities, the more aligned the representations will be.

3 EXPERIMENTS

In this paper, we evaluate ICON on the COCO dataset (Lin et al., 2015), a common choice for multi-modal representation learning due to the high-quality paired samples of images and captions as well as extensive labels of objects present in the scene. All results reported here are on the official validation set of COCO (25K samples, referred to hereafter as the "test set"), while we use 10% of the images and corresponding captions from the full training set (592K samples) as a held out validation set for hyperparameter tuning.

We use the the class-level annotations as the semantic information for the intervention module. This means we define semantic interventions as replacing the presence of one class with another (e.g. replacing all humans with cats in the scene), where the classes are selected uniformly from the all classes present and absent in the original sample respectively. While additional care could be taken to define more complex interventions, such as replacing individual instances of a class, or a more sophisticated selection mechanism, we leave this for future work.

Intervention Design A crucial aspect of ICON is not just what semantic interventions are defined with the annotations but how precisely the intervention module M applies them to the latent samples. To investigate how different intervention designs affect the structure of the learned representation, we explore these three implementations for the intervention module M:

- **Global**: Here the presence of each class i has one global learned vector b_i associated to it, so the intervention is independent of the latent sample. For example, if class i is replaced with class j, the latent sample is modified by adding the difference between the corresponding vectors $M_{i\to j}(z) = z + (b_j - b_i)$.
- **Affine**: This design allows for each class i to have a learned affine transformation $\{\mathbf{A}_i, b_i\}$ associated to it, while class replacements are defined as in the **global** setting: $M_{i\to j}(z) = z + (\mathbf{A}_j - \mathbf{A}_i)z + (b_j - b_i)$.
- **Non-Linear**: The intervention module is a neural network f with one hidden layer that takes the latent sample z and the pre-trained BERT (Devlin et al., 2019) embedding of a fixed text description of the desired intervention as input $M_{i\to j}(z) = z + f(z, \text{BERT}(\text{"replace all } \mathbf{i} \text{ with } \mathbf{j}\text{"}))$. This design is more flexible and can learn complex transformations which can align with the learned latent manifold as well as potentially generalize to unseen interventions by leveraging the BERT embeddings.

Model Architecture and Training Since our focus here is on the alignment of the representations across modalities, rather than the fidelity of the representations for each modality individually, we use an (unaligned) pre-trained `[cls]` token of the Vision Transformer (ViT) (Dosovitskiy et al., 2021) for the images and BERT (Devlin et al., 2019) for the captions so $\mathcal{X} \in \mathbb{R}^{768}$. The encoder, decoder, and classifier are MLPs with approximately 4-5M parameters each, while the **non-linear** intervention module has about 2M parameters (see subsection A.1 for details). The latent space is

a hypersphere in a 512-dimensional space $\mathcal{Z} \in S^{511}$, and since the annotations are the class-level labels $\mathcal{Y} \in \{0,1\}^{80}$, the metric for the classification and semantic loss is the binary cross-entropy, while the reconstruction loss is the mean squared error.

We train our ICON models in two stages: first, we train the model on the ViT features of the COCO images for 100k steps, and then we train a new encoder and decoder on the BERT features of the COCO captions for 100k steps.

For the fine-tuning experiments, we evaluate to what extent the partial alignment achieved by ICON can help alleviate the need for large amounts of paired data to fully align the embeddings using contrastive learning as in CLIP (Radford et al., 2021). After training ICON, we fine-tune the encoders using CLIP's contrastive loss on the COCO dataset for 20k steps with a learning rate 10^{-5} (100x smaller than otherwise).

Baselines To put the structured representation learned by ICON in proper context, we include a baseline we call "Post-training" inspired by concept algebra (Wang et al., 2024). Here we train the encoder of each modality individually, and then find the optimal affine transformation to modify the class-specific information in the latent samples as described in Wang et al. (2024) for each class using the validation set, resulting in an analogue of the **affine** intervention design, but where the intervention module is optimized post-training the other components.

	Unpaired Data	Paired Data
Contrastive	Concept Algebra	CLIP
Consistency	ICON	CyCLIP

In the alignment experiments, we use a CLIP model pretrained by Radford et al. (2021) and a CLIP model we train on COCO using the same architecture as our encoders trained from scratch. Lastly, we include the analogous CyCLIP (Goel et al., 2022) as a baseline to compare the effectiveness of using consistency-based auxiliary losses for the contrastive alignment. Conceptually, ICON distinguishes itself from these baselines by (1) not using any paired data, and (2) using a consistency-based loss to align the representations across modalities (see the mini-table to the right).

4 Results and Analysis

Table 1: Comparison of different intervention module designs reporting the average AUROC and balanced accuracy for the classification task on the observational distribution vs random interventions for images. Note that the post-training baseline performs near random (50%) on the classes that the interventions should not change, suggesting training a non-linear intervention module is necessary to avoid losing semantic information.

Intervention Design	**Observations**			**Interventions**	
	Avg. AUROC	**Avg. Acc**	**Min Acc**	**Changed Acc**	**Unchanged Acc**
Post-training	0.96	90	**80**	89	53
Global	0.96	**91**	79	**99**	69
Affine	0.96	**91**	79	**99**	87
Non-linear	0.96	**91**	78	74	**94**

Table 1 shows the classification performance of the different intervention module designs on the COCO test set. Specifically, we report the average AUROC and balanced accuracy across classes on the observational (data) distribution. For the interventional distribution, we separate the balanced accuracy for the classes that the interventions should affect versus the classes that should remain unchanged to characterize how selective the interventions are. Note that although the post-training baseline based on Wang et al. (2024) performs well on the classes that the interventions should change, it performs near random (50%) on the classes that should remain unchanged. Similarly the simpler intervention designs also struggle to maintain the semantic information that should not be changed by the intervention. This suggests that training a non-linear intervention module is necessary to avoid losing semantic information when applying interventions to the latent samples. However, the **non-linear** intervention design performs somewhat worse on the changed classes, suggesting it is challenging to selectively manipulate the semantic information in the latent space. See the qualitatively similar results for the text modality in subsection A.2

4.1 Cross-Modal Alignment

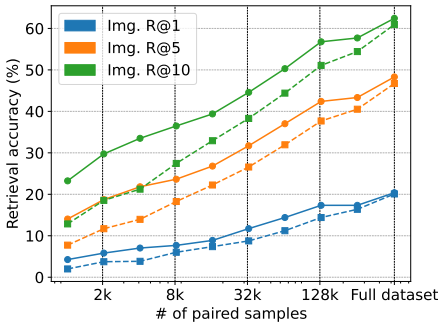

Figure 2: Text-to-Image retrieval performance (%) on the COCO test set when pre-training a CLIP model with ICon (line) vs training from scratch (dashed) given limited paired samples. Note that pre-training with ICon consistently performs comparable when CLIP uses 2-4x as much data.

Method	Paired Samples	Label Retrieval	Image Retrieval @1	Image Retrieval @5	Image Retrieval @10
Pre-trained CLIP	400M	**67**	**42**	**68**	**78**
CLIP	532K	49	20	47	61
CyCLIP	532K	49	20	48	62
ICon + CLIP	131K	49	17	42	57
ICon only	0	19	1	5	8
No training	0	4	0	1	1

Table 2: Text-to-Image retrieval performance (%) for different methods. Note that not only does ICon enable CLIP to achieve comparable performance with almost 5x less paired data, but even without any paired samples, ICon achieves a non-trivial retrieval performance, particularly on the label retrieval task. While the performance is still lower than when using pre-trained CLIP, note that CLIP was trained on a proprietary dataset with 400M paired samples.

As seen in Figure 2, we find that pre-training CLIP with ICon consistently achieves comparable retrieval performance with 2-4x less paired data (text retrieval in Figure 3). Meanwhile, when using all the training samples (532k) to train a CLIP or CyCLIP model from scratch, the performance is not significantly better than when using ICon with only 131k paired samples, as seen in Table 2. This suggests that the partial alignment of the representations learned by ICon is highly informative and useful for alignment.

From Table 2, an additional slightly surprising result is that even without any paired samples, ICon achieves significant performance on our label retrieval task. In the label retrieval task, rather than trying to retrieval the image corresponding to a given caption (in 1, 5, or 10 guesses), we try to retrieve the caption any image sample which has exactly the same labels as the query caption. While naively this might seem easier than sample retrieval, it focuses the search on the semantic information in the labels, which can often be largely implicit due to short or ambiguous captions.

Overall, using the pre-trained CLIP model from Radford et al. (2021), still significantly outperforms CLIP trained on COCO only, as well as ICon. However, pre-trained CLIP is only available for an image + text embedding, so if any other modalities are of interest, it is unlikely that a pre-trained model or large paired dataset (on the order of 400M samples) will be available. Here ICon provides a highly flexible and scalable approach to multimodal representation learning that can be applied to a wide range of modalities with minimal paired data.

5 Conclusion

We present a novel highly versatile method for multimodal representation learning. Rather than requiring relatively expensive paired data across all modalities, ICon only requires similar annotations or label information in each modality individually. This requirement is significantly less stringent than the paired data opening the door to a wider range of different modalities where, for example, samples may individually have tags associated with them, but not corresponding samples in other modalities (as may be common in, for example, social media tags).

By exploiting the structure of the learned representations, we can partially align the embeddings across modalities using a shared set of recognizable interventions. We achieve on par retrieval performance with less than a quarter as many samples when pre-training CLIP with ICon compared to training from scratch. So far we have focused our experiments on images and text with COCO, as the high quality data can readily be applied to more restrictive baselines like CLIP. However, one of the key advantages of ICon is that it can be applied to a wide range of different modalities, and that new modalities can be added incrementally without retraining everything from scratch. Therefore, we intend to explore the effectiveness of ICon on a wider range of modalities next.

ACKNOWLEDGMENTS

The authors would like to thank Chieh-Hsin Lai, Qiyu Wu, Mao Zhuoyuan Mao, and Hiromi Wakaki for their feedback and suggestions.

REFERENCES

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving Image Generation with Better Captions. 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP Training with Language Rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, December 2023.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic Contrastive Language-Image Pretraining. In *Advances in Neural Information Processing Systems*, October 2022.

Felix Leeb, Stefan Bauer, Michel Besserve, and Bernhard Schölkopf. Exploring the Latent Space of Autoencoders with Interventional Assays, June 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(3):74:1–74:34, October 2023. ISSN 1551-6857. doi: 10.1145/3617833.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021.

Adityanarayanan Radhakrishnan, Karren Yang, Mikhail Belkin, and Caroline Uhler. Memorization in Overparameterized Autoencoders, September 2019.

Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A Fistful of Words: Learning Transferable Visual Models from Bag-of-Words Supervision, January 2022.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept Algebra for (Score-Based) Text-Controlled Generative Models, February 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training, September 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment, January 2024.

A APPENDIX

A.1 TRAINING DETAILS

Hyperparameter	**Value**
Training steps	100k (20k for finetuning)
Batch Size	1024
Learning Rate	0.001 (10^{-5} for finetuning)
Optimizer	Adam
AMSGrad	True
Beta1	0.9
Beta2	0.999
Epsilon (Adam)	1×10^{-8}
Latent Dimension	512
Encoder Hidden Layers	[1024, 1024]
Decoder Hidden Layers	[1024, 1024]
Classifier Hidden Layers	[1024, 1024]
(Non-linear) Intervention Module Hidden Layers	[1024]
Activation function	ELU
Dropout (Classifier only)	0.1

Table 3: Summary of key hyperparameters used for training.

A.2 TEXT CLASSIFICATION RESULTS

Table 4: Comparison of different intervention module designs reporting the average AUROC and balanced accuracy for the classification task on the observational distribution vs random interventions for the text modality (captions). Qualitatively the performance is quite similar as for the image modality, except that the observational performance is slightly lower, which is most likely due to many of the captions only implicitly referring to certain objects in the scene.

Intervention Design	**Observations**			**Interventions**	
	Avg. AUROC	**Avg. Acc**	**Min Acc**	**Changed Acc**	**Unchanged Acc**
Post-training	0.94	88	71	**88**	51
Global	0.95	**89**	74	83	68
Affine	0.95	**89**	74	87	87
Non-linear	0.95	**89**	**75**	82	**95**

A.3 TEXT RETRIEVAL RESULTS

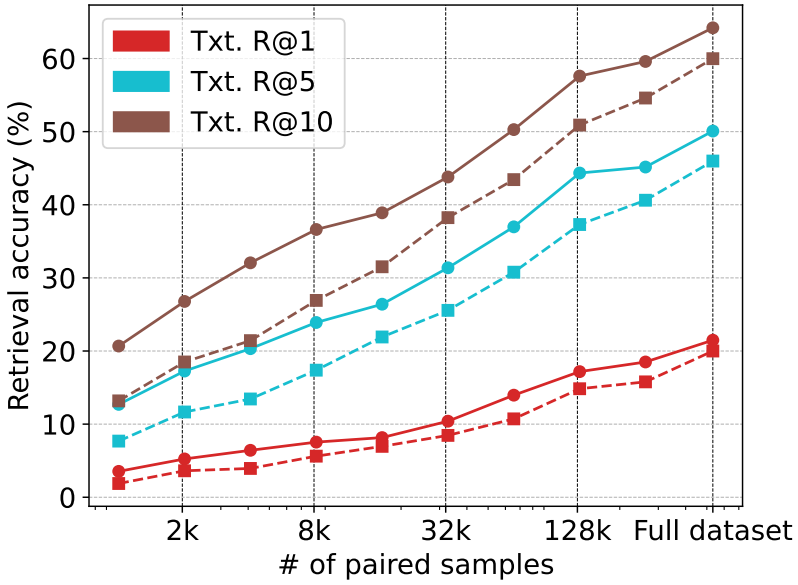

Figure 3: Image-to-Text Retrieval results, which are, unsuprisingly, qualitatively very similar to the Text-to-Image retrieval results seen in Figure 2.