Connect, Collapse, Corrupt: Learning Cross-Modal Tasks with Uni-Modal Data

Yuhui Zhang* Stanford University yuhuiz@stanford.edu

Elaine Sui* Stanford University esui@stanford.edu

Serena Yeung-Levy Stanford University syyeung@stanford.edu

Abstract

In this work, we provide a theoretical explanation of the representation space's geometry resulting from multi-modal contrastive learning, and introduce a three-step method, C^3 (Connect, Collapse, Corrupt), to bridge the modality gap and enhance the interchangeability of embeddings from different modalities. Our C^3 method enables new opportunities to learn cross-modal tasks from uni-modal data, achieving state-of-the-art results on zero-shot image / audio / video captioning and text-to-image generation.

1. Introduction

Recent works reveal that the resulting geometry from multi-modal contrastive learning is nontrivial — corresponding image and text embeddings do not necessarily collapse to the same points in the space. Instead, there is a significant gap between embeddings from different modalities, potentially hindering the direct interchangeable use of image and text embeddings [14, 40]. The lack of comprehension of the resulting geometry from contrastive learning makes it challenging to design methods to leverage this representation space.

In this study, we conclude that the geometry of multi-modal contrastive representation space as (Figure 1):

$$\boldsymbol{e}_x - \boldsymbol{e}_y = \boldsymbol{c}_\perp + \boldsymbol{\epsilon},$$

where $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$ denote the embeddings of paired inputs from different modalities, $\boldsymbol{c}_\perp$ is a constant vector representing the modality gap, which is orthogonal to the embedding span of $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$, and ϵ is a random vector representing the alignment noise, which can be approximated by a Gaussian distribution.

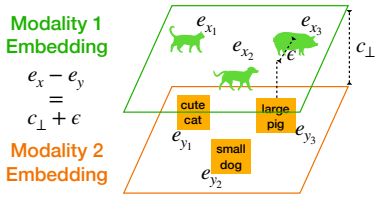


Figure 1. Geometry of the multi-modal contrastive representation space.

We provide a theoretical explanation of the above geometry. Specifically, the modality gap emerges during initialization, because certain dimensions of image and text embeddings remain approximately constant in the embedding space and the constants are distinct for images and texts due to separate initializations. During optimization, these constant dimensions lack a gradient that pushes them to align to the same value, leading to the preservation and orthogonality of the modality gap. Meanwhile, the alignment noise arises from the stable region produced by the contrastive loss, where points within a certain range result in a loss near zero, causing the optimization to halt.

Based on this understanding of the geometry, we propose a simple method, C^3, to improve the interchangeability of embeddings from different modalities, thereby enabling new opportunities to learn cross-modal tasks from uni-modal data. For instance, instead of training an image captioning system on image embeddings, one could train it on caption embeddings. During cross-modal inference, embeddings from the other modality are simply input into the model (Figure 2 and Appendix A).

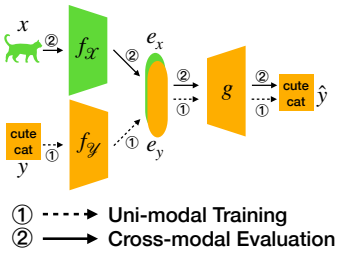


Figure 2. Interchangeable use of embeddings enables learning cross-modal tasks with uni-modal data.

We demonstrate the effectiveness of C^3 on four tasks: image, audio, video captioning and text-to-image generation, and achieve state-of-the-art performance on zero-shot evaluation settings when trained solely on uni-modal data.

2. Multi-Modal Contrastive Representation Space Geometry

We provide the following proposition that describes the geometry of the multi-modal contrastive representation space.

Proposition 1. (Multi-modal Contrastive Representation Space Geometry) *Given a paired image x and text y, the relationship between the ℓ_2-normalized image embedding e_x and text embedding e_y obtained from multi-modal contrastive learning can be described as:*

$$\boldsymbol{e}_x - \boldsymbol{e}_y = \boldsymbol{c}_\perp + \boldsymbol{\epsilon}$$

where $\boldsymbol{c}_\perp$ is a constant vector representing the modality gap and is orthogonal to the image and text embedding span, i.e., $\forall \boldsymbol{e}_{x_1}, \boldsymbol{e}_{x_2}, \boldsymbol{c}_\perp \cdot (\boldsymbol{e}_{x_1} - \boldsymbol{e}_{x_2}) = 0$; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is a random Gaussian vector representing the alignment noise.

In Appendix B, we first explain why the modality gap $\boldsymbol{c}_\perp$ exists before and after model optimization. Then, we introduce the alignment noise $\boldsymbol{\epsilon}$ and its relation to the temperature parameter in contrastive loss. These findings inform our general method of how to adapt uni-modal data for cross-modal learning.

3. Connect, Collapse, Corrupt

Proposition 1 from Section 2 reveals the geometry of the multi-modal contrastive representation space. Based on this, we propose three steps, connect, collapse, corrupt (C^3), to align the representation space.

Stage 1: Connect. This stage establishes connections between similar concepts across different modalities. We directly leverage publicly available models trained with multi-modal contrastive learning, such as CLIP and ImageBind [7, 22]. A modality gap and alignment noise exists in these models' representation spaces.

Stage 2: Collapse. When directly using embeddings of different modalities as input, there is a drastic degradation in performance due to the modality gap, which causes input distributions to the decoder to differ. To address this issue, we adopt a simple approach proposed by [40] that effectively removes the modality gap. Specifically, during training, in place of $\boldsymbol{e}_x$, we feed in $\boldsymbol{e}'_x = \boldsymbol{e}_x - \mathbb{E}_x[\boldsymbol{e}_x]$ to the decoder, and during inference with another modality, in place of $\boldsymbol{e}_y$, we feed in $\boldsymbol{e}'_y = \boldsymbol{e}_y - \mathbb{E}_y[\boldsymbol{e}_y]$. This approach collapses the modality gap, eliminating the input distribution mismatch between the two modalities:

$$\boldsymbol{e}'_x - \boldsymbol{e}'_y = (\boldsymbol{e}_x - \boldsymbol{e}_y) - (\mathbb{E}_x[\boldsymbol{e}_x] - \mathbb{E}_y[\boldsymbol{e}_y]) = \boldsymbol{\epsilon}$$

Stage 3: Corrupt. After removing the modality gap, there is still alignment noise which can be approximated by a $\mathcal{N}(0, \sigma^2 I)$. During unsupervised training, instead of directly decoding y from $\boldsymbol{e}'_y$, we add explicit Gaussian noise to the input and decode y from $\boldsymbol{e}''_y = \boldsymbol{e}'_y + \boldsymbol{\epsilon}$ following [19, 43]. By introducing this noise, the uni-modal and multi-modal training processes become similar, and the learned decoder is more robust and invariant to the small perturbation $\mathcal{N}(0, \sigma^2 I)$, leading to improved performance.

Method	Conn.	Coll.	Corr.	BLEU-1↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑	SPICE↑
Baselines									
ZeroCap (2022)	✗	✗	✗	49.8	7.0	15.4	31.8	34.5	-
MAGIC (2022)	✗	✗	✗	56.8	12.9	17.4	39.9	49.3	11.3
ESPER (2022)	✗	✗	✗	-	21.9	21.9	-	78.2	-
CLIPRe (2023)	✓	✗	✗	-	4.6	13.3	-	25.6	9.2
DeCap (2023)	✓	✗	✗	-	8.9	17.5	-	50.6	13.1
WS-ClipCap (2023)	✓	✗	✗	50.3	9.6	15.2	37.5	33.7	8.6
WS-ClipCap-Multi (2023)	✓	✗	✓	65.5	22.1	22.2	48.0	74.6	14.9
CapDec (2022)	✓	✗	✓	69.2	26.4	**25.1**	51.8	91.8	-
Ours									
C^1	✓	✗	✗	28.1	2.4	12.2	25.4	13.0	6.8
C^2_1	✓	✓	✗	44.4	6.1	15.5	33.6	25.2	9.2
C^2_2	✓	✗	✓	69.0	25.5	24.3	50.8	87.6	17.6
C^3	✓	✓	✓	**71.0**$_{\pm 0.1}$	**27.7**$_{\pm 0.1}$	25.0$_{\pm 0.0}$	**52.0**$_{\pm 0.0}$	**93.3**$_{\pm 0.3}$	**18.3**$_{\pm 0.1}$

Table 1. Image-free image-to-text captioning results. We achieve state-of-the-art zero-shot image captioning and our ablation shows the effectiveness of each component in our method.

Appendix Algorithm 1 summarizes the entire procedure of our proposed method, C^3, that enables learning cross-modal tasks with uni-modal data.

4. Results

In this section, we verify the effectiveness of C^3 on image captioning. For audio captioning, video captioning, text-to-image generation, and generalization other contrastive embedding spaces, please see Appendix D.

We use the ClipCap model [18] for image captioning, which pairs a frozen CLIP ViT-B/32 image encoder [22] with a GPT-2 decoder [21]. A lightweight MLP mapping network bridges the dimensional gap between CLIP (512-d) and GPT-2 (768-d) embeddings and produces a prefix for GPT-2 caption generation. We train and evaluate on the MS-COCO dataset [16] using the standard split [11]. We first train our model for text reconstruction using the MS-COCO captions only. Following C^3, we extract the text embedding from the frozen CLIP text encoder and apply the collapse operation (remove pre-computed mean) and corrupt operation (add Gaussian noise). After training, we evaluate our model in the cross-modal setting, replacing the CLIP text encoder with the CLIP image encoder and decoding captions from image embeddings. We refer to this evaluation setting as image-free zero-shot evaluation, as images are not seen during training.

As shown in Table 1, our proposed method, C^3, outperforms previous state-of-the-art methods in image-free zero-shot captioning. Details of the other methods can be found in Appendix Section G. Our ablation analysis demonstrates that both the collapse and corrupt components are crucial for improving cross-modal evaluation performance, as they eliminate the differences between embeddings from different modalities. Overall, C^3 improves the performance based on a geometric analysis of the multi-modal embedding space and represents a potential standard approach for future works that use a multi-modal contrastive embedding space.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision (ECCV), 2016. 15

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Workshop of Annual Meeting of the Association for Computational Linguistics (ACL Workshop), 2005. 15

[3] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. Conference on Neural Information Processing Systems (NeurIPS), 2021. 10, 12

[4] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. 11

[5] EleutherAI. Clasp: Contrastive language aminoacid sequence pretraining, 2021. 11

[6] Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In International Conference on Learning Representations (ICLR), 2022. 11

[7] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 2, 9, 11

[8] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In International Conference on Learning Representations (ICLR), 2022. 11

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Conference on Neural Information Processing Systems (NeurIPS), 2017. 9, 18

[10] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In International Conference on Learning Representations (ICLR), 2022. 5, 20

[11] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 2, 8, 15

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 9, 16

[13] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. DECAP: Decoding CLIP latents for zero-shot captioning. In International Conference on Learning Representations (ICLR), 2023. 2, 9, 10, 12

[14] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Conference on Neural Information Processing Systems (NeurIPS), 2022. 1, 5, 9, 11, 20

[15] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Workshop of Annual Meeting of the Association for Computational Linguistics (ACL Workshop), 2004. 8, 15

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision (ECCV), 2014. 2, 8, 9, 11, 14, 18

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations (ICLR), 2019. 15

[18] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734, 2021. 2, 8, 14, 16

[19] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected CLIP. In Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP Findings), 2022. 2, 8, 9, 10, 12

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Annual Meeting of the Association for Computational Linguistics (ACL), 2002. 8, 15

[21] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI, 2019. 2, 8, 14

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), 2021. 2, 5, 8, 9, 11, 14, 16

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International Conference on Machine Learning (ICML), 2021. 10, 12

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022. 9, 12

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. Conference on Neural Information Processing Systems (NeurIPS), 2016. 9, 18

[26] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Annual Meeting of the Association for Computational Linguistics (ACL), 2018. 18

[27] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In International Conference on Learning Representations (ICLR), 2022. 9

[28] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: plugging visual controls in text generation. arXiv preprint arXiv:2205.02655, 2022. 2, 10, 12

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 18

[30] Derek Tam, Colin Raffel, and Mohit Bansal. Simple weakly-supervised image captioning via CLIP's multimodal embeddings. In Workshop of AAAI Conference on Artificial Intelligence (AAAI Workshop), 2023. 2, 9, 10

[31] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 2, 10, 12

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 15

[33] Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. arXiv preprint arXiv:2302.07865, 2023. 11

[34] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 9

[35] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021. 11

[36] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 11

[37] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, JaeSung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, et al. Multimodal knowledge alignment with reinforcement learning. arXiv preprint arXiv:2205.12630, 2022. 2, 10, 12

[38] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432, 2021. 11

[39] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In Machine Learning for Healthcare (MLHC), 2022. 11

[40] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In International Conference on Learning Representations (ICLR), 2023. 1, 2, 8, 9, 11

[41] Yufan Zhou, Chunyuan Li, Changyou Chen, Jianfeng Gao, and Jinhui Xu. Lafite2: Few-shot text-to-image generation. arXiv preprint arXiv:2210.14124, 2022. 9, 10

[42] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted diffusion for text-to-image generation. arXiv preprint arXiv:2211.15388, 2022. 10, 12

[43] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. In Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 2, 8, 9, 10, 12, 16, 18

A. Learning Cross-Modal Tasks with Uni-Modal Data

In this section, we present the general notion of a cross-modal task and how we can leverage uni-modal data to learn such tasks.

A.1. Cross-Modal Task Formulation

Cross-modal tasks aim to learn a model that maps inputs from one modality $\mathcal{X}$ (e.g., images) to another modality $\mathcal{Y}$ (e.g., texts). Given a paired multi-modal dataset $\mathcal{D} = \{(x, y) \in \mathcal{X} \times \mathcal{Y}\}$ (e.g., an image-caption dataset), the task can be achieved by minimizing the empirical risk $\mathcal{L}_d$ between the predicted target $\hat{y} = g(f_{\mathcal{X}}(x))$ and the true target y over the dataset $\mathcal{D}$, where $f_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}^d$ is an encoder that maps inputs from $\mathcal{X}$ to a d-dimensional representation space, and $g : \mathbb{R}^d \to \mathcal{Y}$ is a decoder that maps outputs from the encoder to $\mathcal{Y}$:

$$\min_{g, f_{\mathcal{X}}} \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \mathcal{L}_d(\hat{y}, y), \quad \hat{y} = g(f_{\mathcal{X}}(x))$$

$\mathcal{L}_d$ measures the discrepancy between the predicted and true targets, which can be mean squared error (MSE) for images and cross-entropy loss for texts.

A.2. Enabling Cross-Modal Tasks with Uni-Modal Data

The need for a multi-modal paired dataset $\mathcal{D}$ to learn cross-modal tasks is suboptimal, as collecting such datasets can be expensive and time-consuming. However, if we have a encoder $f_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}^d$ that maps inputs from $\mathcal{Y}$ to the same representation space as the encoder $f_{\mathcal{X}}$, i.e., $\forall x, y \in \mathcal{D}, f_{\mathcal{X}}(x) = f_{\mathcal{Y}}(y)$, we can train the cross-modal task using a uni-modal dataset $\mathcal{D}' = \{y \in \mathcal{Y}\}$:

$$\min_{g} \frac{1}{|\mathcal{D}'|} \sum_{y\in\mathcal{D}'} \mathcal{L}_d(\hat{y}, y), \quad \hat{y} = g(f_{\mathcal{Y}}(y))$$

Note that $f_{\mathcal{Y}}$ should be frozen during training to maintain its embedding alignment with $f_{\mathcal{X}}$. When evaluating in a cross-modal setting, we can replace $f_{\mathcal{X}}$ with $f_{\mathcal{Y}}$, thus enabling cross-modal tasks with only uni-modal training data.

A.3. Establishing a Shared Representation Space

Recent advances in multi-modal contrastive learning has enabled for encoders that map similar inputs from different modalities to a shared representation space. Specifically, given a large multi-modal dataset[1], n paired inputs are randomly sampled during each iteration and the following objective is optimized [22]:

$$\min_{f_{\mathcal{X}}, f_{\mathcal{Y}}} \mathcal{L} = -\frac{1}{2n} \sum_{i=1}^{n} \Big(\log \frac{\exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j}/\tau)} + \log \frac{\exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_j} \cdot \boldsymbol{e}_{y_i}/\tau)} \Big)$$

where $\boldsymbol{e}_x = f_{\mathcal{X}}(x)$, $\boldsymbol{e}_y = f_{\mathcal{Y}}(y)$, and τ is the temperature hyperparameter. This loss function encourages high similarity between the embeddings of the paired image-texts relative to the similarities between unpaired ones. Intuitively, after optimizing the loss, paired image and text embeddings should collapse to the same point. However, empirical results have shown a significant gap between paired embeddings [14], which prevents the direct interchangeable use of image and text embeddings. In the next section, we provide a detailed analysis of the geometry of the multi-modal contrastive representation space.

B. Multi-Modal Contrastive Representation Space Geometry

We begin by providing the following proposition that describes the geometry of the multi-modal contrastive representation space. These findings inform our general method of how to adapt uni-modal data for cross-modal learning.

Proposition 2. (Multi-modal Contrastive Representation Space Geometry)
Given a paired image x and text y, the relationship between the ℓ_2-normalized image embedding e_x and text embedding e_y obtained from multi-modal contrastive learning can be described as:

$$\boldsymbol{e}_x - \boldsymbol{e}_y = \boldsymbol{c}_\perp + \boldsymbol{\epsilon}$$

where $\boldsymbol{c}_\perp$ is a constant vector representing the modality gap and is orthogonal to the image and text embedding span, i.e., $\forall \boldsymbol{e}_{x_1}, \boldsymbol{e}_{x_2}, \boldsymbol{c}_\perp \cdot (\boldsymbol{e}_{x_1} - \boldsymbol{e}_{x_2}) = 0$; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is a random Gaussian vector representing the alignment noise.

In the following subsections, we will first explain why the modality gap $\boldsymbol{c}_\perp$ exists before and after model optimization. Then, we will introduce the alignment noise $\boldsymbol{\epsilon}$ and its relation to the temperature parameter in contrastive loss.

B.1. Modality Gap

The presence of a modality gap and its orthogonality to the image and text embedding span are due to the joint effect of initialization and optimization.

Initialization. The modality gap exists when randomly initializing multi-modal models, which can be explained by the dimensional collapse [10] of the representation space defined below.

[1] Acquiring domain-specific paired multi-modal datasets can be challenging. However, several works have gathered large-scale noisy image-caption pairs from the web and made pre-trained encoders $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$ available for direct use.

Definition 1. (Dimensional Collapse of the Representation Space)
Given a d-dimensional representation space $\mathbb{R}^d$, we define its effective dimension d_e as:

$$d_e = \arg\min_{d'} \frac{\sum_{i=1}^{d'} \sigma_i}{\sum_{i=1}^{d} \sigma_i} \geq \gamma$$

where the σ_i's are the singular values of the representation covariance matrix in decreasing order, and γ thresholds the minimum variance explained by the d_e dimensions. Dimensional collapse occurs when $d_e \ll d$.

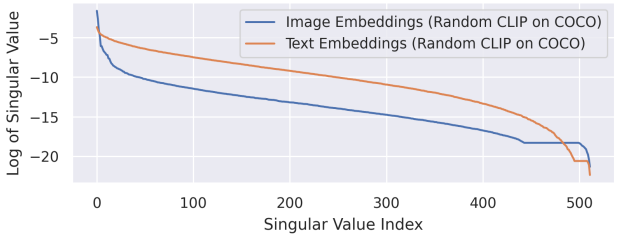


Figure 3. Dimensional collapse of the CLIP representation space. Singular values obtained from SVD reveal that the effective dimension of the image and text representation space is much smaller than the total number of dimensions.

To demonstrate the dimensional collapse phenomenon, we took a randomly initialized CLIP with a $d = 512$ representation space and fed MS-COCO images and captions as input. We obtained the corresponding image features and text features and performed SVD on the image feature and text feature covariance matrices. The distribution of the singular values is shown in Figure 3, where we can clearly see that the effective dimension of both the image and text features is small. Specifically, when setting $\gamma = 0.99$, the effective dimension of image embeddings is $d_{e,x} = 25$ and that of text embeddings is $d_{e,y} = 230$. Therefore, the effective dimension of the shared representation space $d_e \leq d_{e,x} + d_{e,y} = 255$. The equality holds only when all the effective dimensions of the image and text are orthogonal.

Dimensional collapse indicates that only a small number of dimensions contribute significantly to the variance of the representation space, while the remaining dimensions can be viewed as constant. Suppose the shared representation space has a maximum effective dimension $d_e = 255$. This indicates that the image and text embeddings will remain constant in the $d_c = d - d_e = 257$ ineffective dimensions. As a result, a modality gap exists at the beginning of model optimization, as these d_c ineffective dimensions will be inherently different for images and texts, given random initialization.

To verify this, we synthesize $n = 1,000$ image and text embeddings in $d = 512$ space. For each image embedding, we initialize the first $d_{e,x}$ dimensions, and for each text embedding, the $d_{e,x}$-th to $(d_{e,x} + d_{e,y})$-th dimensions, by randomly sampling a standard Gaussian distribution $\mathcal{N}(0, 1)$. All the other dimensions are set to a constant value drawn from a $\mathcal{N}(0, 1)$, where the constant for the image and text embeddings are different. Figure 4 (left) illustrates this setup by showing the variance of each dimension. This setup mimics our findings of SVD analysis on CLIP, where we set image embeddings to have $d_{e,x}$ effective dimensions and text embeddings to have $d_{e,y}$ effective dimensions, and assume these effective dimensions are fully orthogonal. We normalize the embeddings to unit length before performing our analysis.

We observe a clear modality gap at the beginning: the ℓ_2-distance between the mean of the image embeddings and the mean of the text embeddings is 1.21. When we only consider the last d_c ineffective dimensions, the average distance is 0.99. More discussion is in Appendix O.

Optimization. The analysis above reveals that the modality gap exists at initialization, and we further analyze why optimizing for the multi-modal contrastive loss fails to close the gap. The following lemma reveals that there is no gradient in the modality gap direction, therefore the gap and its orthogonality will be preserved.

Lemma 1. (Gradients in Contrastive Optimization) *(Proof in Appendix H)*
With the mild assumption of equal presence of n images and texts with $p(x_i) = p(y_i) = 1/n$, optimizing the multi-modal contrastive loss $\mathcal{L} = -\frac{1}{2n}\sum_{i=1}^{n}\left(\log \frac{\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n}\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_j}/\tau)} + \log \frac{\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n}\exp(\boldsymbol{e}_{x_j}\cdot\boldsymbol{e}_{y_i}/\tau)}\right)$ yields the following gradients:

$$\nabla_{\boldsymbol{e}_{x_i}}\mathcal{L} = \lambda\sum_{j=1}^{n}\alpha_{y_j}(\boldsymbol{e}_{y_j} - \boldsymbol{e}_{y_i}),$$

$$\nabla_{\boldsymbol{e}_{y_i}}\mathcal{L} = \lambda\sum_{j=1}^{n}\alpha_{x_j}(\boldsymbol{e}_{x_j} - \boldsymbol{e}_{x_i})$$

where $\lambda = 1/(2n\tau)$, $\alpha_{x_j} = p(x_j|y_i) + p(y_i|x_j)$, $\alpha_{y_j} = p(y_j|x_i) + p(x_i|y_j)$, $p(x_i|y_j) = \frac{\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_j}/\tau)}{\sum_{k=1}^{n}\exp(\boldsymbol{e}_{x_k}\cdot\boldsymbol{e}_{y_j}/\tau)}$, $p(y_i|x_j) = \frac{\exp(\boldsymbol{e}_{y_i}\cdot\boldsymbol{e}_{x_j}/\tau)}{\sum_{k=1}^{n}\exp(\boldsymbol{e}_{y_k}\cdot\boldsymbol{e}_{x_j}/\tau)}$, and τ is temperature.

Lemma 1 highlights that the gradients of image embeddings during contrastive optimization are fully determined by the text embedding span, and vice versa. Due to the dimensional collapse of the image and text embedding span, the contrastive optimization process fails to propagate gradients in the direction of the ineffective dimensions, resulting in gap preservation and orthogonality to the image and

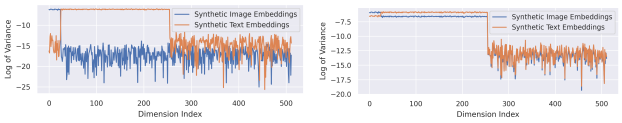


Figure 4. Variance of each dimension before (left) and after (right) multi-modal contrastive optimization. Our analysis reveals that gradients will only be propagated to effective dimensions and no gradient will be propagated to ineffective dimensions. Therefore, the effective dimensions are aligned while ineffective dimensions remain constant after optimization.

text embedding span after optimization. Lemma 1 also implies that the effective dimensionality of the joint representation space remains unchanged after optimization.

To empirically verify this, we optimize the multi-modal contrastive loss $\mathcal{L}$ on the $n = 1,000$ previously synthesized image and text embeddings. We optimize for 200K steps with a learning rate 0.1, and CLIP's initial temperature of $\tau = 0.07$. We compute the variance of each dimension pre- and post-contrastive optimization and plot the results in Figure 4 (right). From the figure, we can see that the first $d_e = 255$ effective dimensions are aligned after optimization while the last $d_c = 257$ ineffective ones remain unchanged. This verifies that no gradient is propagated to the ineffective dimensions, as the variance for these dimensions remains zero after optimization. The modality gap becomes slightly smaller (0.82 compared to 0.99 before optimization) mainly due to ℓ_2-regularization, where changes in the effective dimensions affect changes in the ineffective ones.

In summary, due to dimensional collapse at model initialization, there are ineffective dimensions where image and text embeddings can be viewed as different constants, resulting in a modality gap at initialization. During optimization, these ineffective dimensions have no gradient update, and thus, the modality gap and its orthogonality are preserved after optimization.

B.2. Alignment Noise

In this section, we explain alignment noise after multi-modal contrastive learning. This noise results from the stable region of contrastive loss demonstrated in the following lemma, where we can consider a single term in the loss given the symmetry of the added terms.

Lemma 2. (Stable Region Controlled by Temperature) *(Proof in Appendix H)*
We consider a single term in the multi-modal contrastive loss $\mathcal{L}_i = -\log \frac{\exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j}/\tau)}$. *We define the margin* $r = \boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i} - \max_{j \neq i} \boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j}$ *as the measure of the similarity difference between the matched pair and the hardest negative pair. When* r *exceeds a threshold given below,* $\mathcal{L}_i$ *falls below a small pre-set value* δ*, where we assume optimization ends:*

$$r \geq \tau \log \frac{o(\tau)}{\exp(\delta) - 1},$$

where $o(\tau)$ *is a monotonically increasing function of temperature* τ *that satisfies* $1 < o(\tau) < n$*. Therefore, the required margin* r *is monotonically increasing with* τ*.*

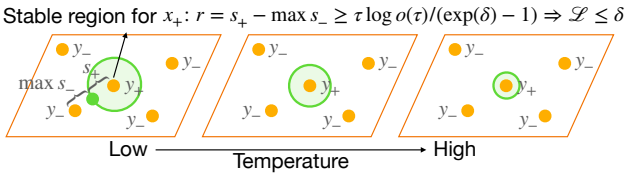


Figure 5. Stable region (green area) of contrastive learning controlled by temperature. Within the stable region, the loss falls below a small preset value, indicating that optimization has ended. The region increases as the temperature decreases.

Lemma 2 suggests that there is a stable region of the contrastive loss. This region can be viewed as a function of temperature where region size increases as the temperature decreases, with the required margin becoming smaller. Within this stable region, the loss falls below the small value δ, indicating that optimization has ended. In the extreme case where $\tau \to 0_+$, if $r \geq 0$, the loss will be less than δ ($\delta \to 0_+$ for this case). This means that given all the y_j, x_i can end up within a region instead of a fixed point, resulting in the same zero loss. Figure 5 illustrates the stable region defined by the margin with regard to the temperature. Therefore, there may be a mismatch between the matched pairs in the representation space, resulting in alignment noise.

B.3. Summary

Statistic	Mean	Std
$\|\boldsymbol{d}^{(i)}\|_2$	0.83	0.01
$\cos(\boldsymbol{d}^{(i)}, \boldsymbol{d}^{(j)})$	0.99	0.00
$\cos(\boldsymbol{d}^{(i)}, \boldsymbol{r}_{j,k}^{(i)})]$	0.00	0.06
$\mathbb{E}[\boldsymbol{\epsilon}_j^{(i)}]_k$	0.00	0.00
$\cos(\boldsymbol{\epsilon}_j^{(i)}, \boldsymbol{\epsilon}_k^{(i)})$	0.00	0.10

Table 2. Statistics that reveals representation space geometry.

In Section B.1 and B.2, we explain the modality gap $\boldsymbol{c}_\perp$ and alignment noise $\boldsymbol{\epsilon}$ in Proposition 2, respectively. Combining them together, we explain the geometric relation of paired embeddings $\boldsymbol{e}_x - \boldsymbol{e}_y = \boldsymbol{c}_\perp + \boldsymbol{\epsilon}$.

We verify this geometric relation using CLIP on the MS-COCO image-caption dataset. We randomly group

each 100 images into group i, and define individual gap $\boldsymbol{d}_j^{(i)} = \boldsymbol{e}_{x_j}^{(i)} - \boldsymbol{e}_{y_j}^{(i)}$, group gap $\boldsymbol{d}^{(i)} = \mathbb{E}_j[\boldsymbol{d}_j^{(i)}]$, image difference $\boldsymbol{r}_{j,k}^{(i)} = \boldsymbol{e}_{x_j}^{(i)} - \boldsymbol{e}_{x_k}^{(i)}$, alignment noise $\boldsymbol{\epsilon}_j^{(i)} = \boldsymbol{d}_j^{(i)} - \boldsymbol{d}^{(i)}$, where j, k are image or text indices. These statistics are computed in Table 2, where the first three statistics show that the modality gap $\boldsymbol{c}_\perp$ approximates a constant vector orthogonal to the image and text embedding span, and the last two statistics show that the alignment noise $\boldsymbol{\epsilon}$ can be viewed as Gaussian noise. We provide a detailed explanation of how to interpret these statistics in Appendix I.

This geometric analysis in this section serves as the foundation of the approach we introduce in the next section, where we develop a simple method to align the shared representation space and enable learning cross-modal tasks with uni-modal data.

C. Connect, Collapse, Corrupt

Proposition 2 from Section B reveals the geometry of the multi-modal contrastive representation space. Based on this, we propose three steps, connect, collapse, corrupt (C^3), to align the representation space, making it possible for embeddings from different modalities to be interchangeably consumed by the decoder and thus enabling learning cross-modal tasks from uni-modal data.

Stage 1: Connect. This stage establishes connections between similar concepts across different modalities. We leverage recent advances in multi-modal contrastive learning [22] and use encoders trained with this strategy to build cross-modal models. However, a modality gap and alignment noise exists after multi-modal contrastive learning, as shown in Proposition 2.

Stage 2: Collapse. When directly using embeddings of different modalities as input, there is a drastic degradation in performance due to the modality gap, which causes input distributions to the decoder to differ. To address this issue, we adopt a simple approach proposed by [40] that effectively removes the modality gap. Specifically, during training, in place of $\boldsymbol{e}_x$, we feed in $\boldsymbol{e}'_x = \boldsymbol{e}_x - \mathbb{E}_x[\boldsymbol{e}_x]$ to the decoder, and during inference with another modality, in place of $\boldsymbol{e}_y$, we feed in $\boldsymbol{e}'_y = \boldsymbol{e}_y - \mathbb{E}_y[\boldsymbol{e}_y]$. This approach collapses the modality gap, eliminating the input distribution mismatch between the two modalities:

$$\boldsymbol{e}'_x - \boldsymbol{e}'_y = (\boldsymbol{e}_x - \boldsymbol{e}_y) - (\mathbb{E}_x[\boldsymbol{e}_x] - \mathbb{E}_y[\boldsymbol{e}_y]) = \boldsymbol{\epsilon}$$

Stage 3: Corrupt. After removing the modality gap, there is still alignment noise which can be approximated by a $\mathcal{N}(0, \sigma^2 I)$. During unsupervised training, instead of directly decoding y from $\boldsymbol{e}'_y$, we add explicit Gaussian noise to the input and decode y from $\boldsymbol{e}''_y = \boldsymbol{e}'_y + \boldsymbol{\epsilon}$ following [19, 43]. By introducing this noise, the uni-modal and multi-modal training processes become similar, and the learned decoder is more robust and invariant to the small perturbation $\mathcal{N}(0, \sigma^2 I)$, leading to improved performance.

Appendix Algorithm 1 summarizes the entire procedure of our proposed method, C^3, that enables learning cross-modal tasks with uni-modal data.

D. Results

In this section, we verify the effectiveness of our proposed method, C^3, on four tasks: image captioning, audio captioning, video captioning, and text-to-image generation. We show that our method achieves state-of-the-art performances, generalizes to different modalities and contrastive embedding spaces, and is especially useful when multi-modal data are limited.

D.1. Image Captioning

We use the ClipCap model [18], pairing a frozen CLIP ViT-B/32 image encoder [22] with a GPT-2 decoder [21]. A lightweight MLP mapping network bridges the dimensional gap between CLIP (512-d) and GPT-2 (768-d) embeddings and also produces a prefix for GPT-2 caption generation. We train and evaluate on the MS-COCO dataset [16] using the standard split [11], comprising 113K training images and 5K each for validation and testing, with each image having 5 captions. We utilize metrics such as BLEU [20] and ROUGE [15] to evaluate lexical and semantic similarity between generated and human captions.

We first train our model for text reconstruction using the MS-COCO captions only. Following C^3, we extract the text embedding from the frozen CLIP text encoder and apply the collapse operation (remove pre-computed mean) and corrupt operation (add Gaussian noise). After training, we evaluate our model in the cross-modal setting, replacing the text encoder with the CLIP image encoder and decoding captions from image embeddings. We refer to this evaluation setting as image-free zero-shot evaluation, as images are not seen during training. Additionally, we fine-tune the pre-trained model on different amounts of image-caption pairs and evaluate its performance. We refer to this evaluation setting as semi-supervised evaluation. More details can be found in Appendix K. We show image-free zero-shot captioning results in Table 3 and semi-supervised captioning results in Figure 6.

C^3 achieves state-of-the-art image-free zero-shot captioning results. As shown in Table 3, our proposed method, C^3, outperforms previous state-of-the-art methods in image-free captioning. Details of the other methods can be found in Appendix Section G. Our ablation anal-

ysis demonstrates that both the collapse and corrupt components are crucial for improving cross-modal evaluation performance, as they eliminate the differences between embeddings from different modalities. Notably, the most competitive baseline, CapDec [19], can be viewed as an ablated version of C^3, but without an analysis of why the corruption works. Our proposed method, however, provides a clear explanation based on a geometric analysis of the multi-modal embedding space, and we further improve performance by introducing a collapse step. Overall, C^3 represents a potential standard approach for future works that use a multi-modal contrastive embedding space.

C^3 is particularly useful in low-data regimes. In scenarios where multi-modal data is limited, our method remains highly effective. To demonstrate this, we fine-tuned our pre-trained model on 1%, 5%, 25%, and 100% of the MS-COCO training image-text pairs, and compared our performance with a fully supervised baseline (ClipCap) and ablated models (see Figure 6). The results clearly show that C^3 outperforms the fully supervised baseline across all metrics, with the most significant improvements seen in low-data regimes where multi-modal paired data are limited. Thus, our method represents a promising solution for achieving cross-modal tasks in such scenarios.

Qualitative analysis of collapse and corrupt. Both the collapse and corrupt components of our method show consistent improvements, but it is not immediately clear how they do so. To address this, we provide qualitative results in Appendix K. We can see that after collapsing, the generated captions are much more natural and fluent, as it removes the most significant distributional difference between image and text embeddings. After corrupting, the model generates more accurate and faithful text descriptions of the image. We hypothesize that adding noise makes the decoder robust to small variations in the embedding space. Therefore when evaluating in cross-modal settings, the alignment noise will not affect the prediction and thus reduces hallucination in the generated caption.

D.2. Text-to-Image Generation

We further apply our method, C^3, to text-to-image generation, the reverse of image captioning. We utilize LAFITE [43], which integrates a frozen CLIP ViT-B/32 [22] as the text encoder and a modified StyleGAN2's generator [12] as the trainable decoder. We use the MS-COCO dataset [16] with LAFITE's official split, including 82K training and 40K validation images, with 5 captions each. We use the standard metrics such as FID (Fréchet Inception Distance) [9] and IS (Inception Score) [25] to assess the realism of generated images. Similar to the image captioning setting, we first train the model for image reconstruction using images only, and then evaluate the model in the cross-modal setting by generating images from text. We show the language-free zero-shot image generation results in Table 4. Similar to image captioning, we find that our method C^3 **consistently outperforms the baselines** in terms of FID and IS. Our ablation study further reveals that each component of C^3 is useful for improving the performance. More detailed setups and qualitative comparisons can be found in Appendix L.

D.3. Generalization to Other Modalities and Embedding Spaces

To verify the generalization of our method to other modalities, datasets, and embedding spaces, we further conduct experiments on zero-shot captioning from image, audio and video using ImageBind [7] embeddings. We use the same settings as image captioning with CLIP. Results are shown in Table 5. We find that C^3 **consistently improves baselines in all the settings**, and **using ImageBind embeddings achieves further improvements in image captioning compared to CLIP embeddings**.

E. Related Works (Full Version in Appendix G)

Multi-modal contrastive learning and resulting geometry. Multi-modal contrastive learning aims to bridge representations from different modalities, drawing similar concepts closer and distancing dissimilar ones [22]. CLIP [22], ImageBind [7], and similar models have leveraged extensive multi-modal data to construct such representation spaces, which have been demonstrated to effectively support a range of uni-modal and multi-modal applications [24, 27, 34]. However, the resulting geometry in the shared representation space, particularly the "modality gap", where embeddings from different modalities are clearly separate in the shared representation space, remains under-explored [14, 40]. In our work, we unify the observations from Liang et al. [14] and Zhang et al. [40], and contribute a formal formulation and theoretical explanation of the unique geometry resulting from multi-modal contrastive learning.

Learning cross-modal tasks with uni-modal data. Given the expense of multi-modal data collection, there is a growing interest in learning cross-modal tasks using uni-modal data. Based on the assumption that contrastive optimization makes representations from different modalities interchangeable, recent works have leveraged these representation spaces and shown great success in building image captioning models with text data only [13, 19, 30] and text-to-image generation models with image data only [41–43]. Despite these advancements, these methods have

Method	Conn.	Coll.	Corr.	BLEU-1↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑	SPICE↑
Baselines									
ZeroCap (2022)	✗	✗	✗	49.8	7.0	15.4	31.8	34.5	-
MAGIC (2022)	✗	✗	✗	56.8	12.9	17.4	39.9	49.3	11.3
ESPER (2022)	✗	✗	✗	-	21.9	21.9	-	78.2	-
CLIPRe (2023)	✓	✗	✗	-	4.6	13.3	-	25.6	9.2
DeCap (2023)	✓	✗	✗	-	8.9	17.5	-	50.6	13.1
WS-ClipCap (2023)	✓	✗	✗	50.3	9.6	15.2	37.5	33.7	8.6
WS-ClipCap-Multi (2023)	✓	✗	✓	65.5	22.1	22.2	48.0	74.6	14.9
CapDec (2022)	✓	✗	✓	69.2	26.4	**25.1**	51.8	91.8	-
Ours									
C^1	✓	✗	✗	28.1	2.4	12.2	25.4	13.0	6.8
C_1^2	✓	✓	✗	44.4	6.1	15.5	33.6	25.2	9.2
C_2^2	✓	✗	✓	69.0	25.5	24.3	50.8	87.6	17.6
C^3	✓	✓	✓	**71.0**$_{\pm 0.1}$	**27.7**$_{\pm 0.1}$	25.0$_{\pm 0.0}$	**52.0**$_{\pm 0.0}$	**93.3**$_{\pm 0.3}$	**18.3**$_{\pm 0.1}$

Table 3. Image-free image-to-text captioning results. We achieve state-of-the-art zero-shot image captioning and our ablation shows the effectiveness of each component in our method.

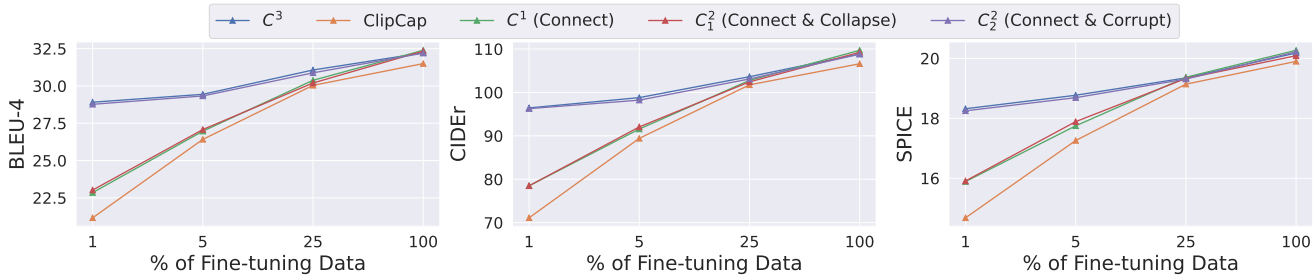


Figure 6. Image-to-text captioning results in the low data regime. When paired multi-modal data are limited, our approach that leverages uni-modal data for pre-training leads to substantial improvements compared to the purely supervised method (ClipCap).

Method	Conn.	Coll.	Corr.	FID↓	IS↑
Baselines					
DALL-E (2021)	✗	✗	✗	27.5	17.9
CogView (2021)	✗	✗	✗	27.1	18.2
LAFITE$_G$ (2022)	✓	✗	✓	20.9	24.9
Ours					
C^1	✓	✗	✗	29.8	22.4
C_1^2	✓	✓	✗	21.7	24.4
C_2^2	✓	✗	✓	19.8	25.5
C^3	✓	✓	✓	**19.6**	**26.0**

Table 4. Language-free text-to-image generation results. Our method C^3 consistently outperforms the baselines.

noted the intriguing "modality gap" phenomenon and proposed different empirical methods to address this gap, such as [30]'s paraphrased decoding, [13, 41]'s memory retrieval, [19, 43]'s noise addition, and [42]'s prior network. In our work, we first provide a theoretical analysis of the multi-modal representation space geometry. Based on the geometry, we propose a simple method that addresses the "modality gap" in a principled manner and ultimately improves performance on cross-modal tasks and outperforms these strategies.

F. Conclusion

In this work, we provide a theoretical explanation of the unique geometry that arises from multi-modal contrastive learning. Building upon this, we present a straightforward technique, C^3, which enhances the interchangeability of embeddings between modalities, enabling the creation of cross-modal applications using only uni-modal data. We demonstrate the effectiveness of our approach on image, audio, video captioning and text-to-image generation, achieving state-of-the-art performance on zero-shot evaluation settings when trained solely on uni-modal data.

Overview of Appendix

In this appendix, we supplement related works and additional details of theory and experiments.

	Image Captioning (MS-COCO 2014)			Audio Captioning (Clotho 2020)			Video Captioning (MSR-VTT 2016)		
	BLEU-1↑	METEOR↑	ROUGE-L↑	BLEU-1↑	METEOR↑	ROUGE-L↑	BLEU-1↑	METEOR↑	ROUGE-L↑
C^1	33.5	12.9	25.8	21.8	17.3	18.1	16.7	12.6	15.2
C_1^2	53.8	17.4	38.6	26.7	20.0	21.4	25.1	17.8	23.1
C_2^2	64.4	22.2	45.8	27.6	20.0	20.6	25.2	18.1	23.8
C^3	**74.0**	**26.6**	**54.0**	**29.5**	**20.1**	**23.0**	**31.4**	**20.0**	**26.9**

Table 5. Generalization of C^3 to other modalities, datasets, and contrastive embedding spaces.

- In Appendix G, we provide detailed related works to contextualize our work in existing works of multi-modal contrastive learning and learning cross-modal tasks with uni-modal data.
- In Appendix H, we provide proofs of the two lemmas used in the main paper that reveal important properties of multi-modal contrastive learning.
- In Appendix I, we employ statistical methods to validate the proposed geometric structure on large pre-trained contrastive models.
- In Appendix J, we summarize the C^3 method into an algorithm.
- In Appendix K, we provide additional experimental details and qualitative results of image captioning.
- In Appendix L, we provide additional experimental details and qualitative results of text-to-image generation.
- In Appendix M, we explain the importance of aligning embeddings from different modalities.
- In Appendix N, we offer further discussions about the effectiveness of collapse vs corrupt.
- In Appendix O, we offer further insights into dimensional collapse and connect it to the cone effect identified by Liang et al. [14].

G. Related Works

Multi-modal contrastive learning and resulting geometry. Multi-modal contrastive learning aims to create a shared representation for different modalities by attracting similar while repelling dissimilar concepts from different modalities during the optimization process [5, 7, 22, 35, 38, 39]. Recent works such as CLIP [22] have leveraged large-scale image-text data during pre-training, resulting in models that can build strong uni-modal and cross-modal applications. Connecting different modalities can be advantageous given the complementarity of different modalities. For example, connecting vision and language enables zero-shot categorization of visual objects [22, 38], explanation of model prediction errors or internal representations [6, 8], diagnosis and rectification of vision models using language by composing different concepts [33, 40], and learning cross-modal tasks with uni-modal data.

However, the geometry resulting from multi-modal contrastive learning has received limited study. Recent work [14] found that there is a clear distinction between embeddings from different modalities in the shared representation space, which is referred to as the modality gap. Liang et al. [14] correctly attributed the gap to the joint effect of model initialization and optimization. However, they did not study the geometric property of the gap and their theory cannot explain the geometry as well. Their theory can only show there is a distributional difference between image and text embeddings. Subsequently, Zhang et al. [40] studied the geometric properties of the modality gap and found that the gap can be empirically well approximated by a constant vector orthogonal to the image or text embedding subspace. However, this work did not provide an explanation for how this unique geometry arises.

In our work, we unify the observations from Liang et al. [14] and Zhang et al. [40], and contribute a formal formulation and theoretical explanation of the unique geometry resulting from multi-modal contrastive learning. Specifically, we show two important factors of the geometry of the multi-modal representation space: modality gap and alignment noise, where the modality gap is a constant vector orthogonal to the image and text embedding span, and the alignment noise can be approximated by a Gaussian distribution. The modality gap arises due to the interplay between the dimensional collapse in model initialization and the resulting collapsed gradient during optimization, whereas the alignment noise is due to the stable region of contrastive loss. These explanations provide a deeper understanding of multi-modal contrastive learning and resulting geometry. Moreover, this non-trivial geometry has important implications for building applications on top of a multi-modal representation space.

Learning cross-modal tasks with uni-modal data. Multi-modal data are less abundant and more expensive to collect than uni-modal data, making it ideal to learn cross-modal tasks with uni-modal data. The recent rise of multi-modal contrastive learning provides this possibility, as similar concepts from different modalities establish close connections during the optimization process. Many recent works have leveraged CLIP, which aligns images and text to achieve image-to-text captioning with text-only data and text-to-image generation with image-only data.

These methods assume and explore how to interchangeably use image embeddings and text embeddings resulting from multi-modal contrastive learning. They achieve better performance than prior methods that did not leverage a multi-modal representation space, outperforming ZeroCap [31], MAGIC [28], and ESPER [37] on captioning, and DALL-E [23] and CogView [3] on image generation. However, due to the peculiar geometry that arises from multi-modal contrastive learning, paired image embeddings and text embeddings are not collapsed to the same point, making it non-trivial to substitute one for the other. Therefore, researchers are proposing different methods to tackle this problem.

For image-to-text captioning, WS-ClipCap [13] was the first work to leverage CLIP's text embeddings to decode text during training, finding that training to decode a paraphrased version of the text from the corresponding CLIP's text embedding leads to significantly better performance (WS-ClipCap-Multi). That is, for datasets where a single image is paired with multiple captions, it is better to decode a caption by feeding the text embedding obtained from one of the other captions corresponding to the same image. Despite proposing the method, WS-ClipCap failed to explain why it works and that on a high level, this paraphrased decoding can be viewed as adding noise. Decap [13] maintains a memory of image embeddings and converts image embeddings to text embeddings by using a weighted average of the most similar image embeddings. They then feed the decoder with the converted embedding to decode the same text. While their method outperforms the baseline CLIPRe, which directly retrieves the most similar captions based on image embeddings, it underperforms WS-ClipCap-Multi, despite proposing a more complex method to tackle the modality gap. CapDec [19] found that directly adding Gaussian noise to the embedding and then feeding it into the decoder leads to comparably better performance. This method corresponds exactly to the corrupt stage in our method. CapDec attributes its success to the hypothesis that adding Gaussian noise closes the modality gap, however, this intuition is inaccurate.

For text-to-image generation, LAFITE [43] was the first work to use CLIP with uni-modal images only. Their approach can be seen as the inverse version of CapDec, where they add Gaussian noise to the image embedding and decode the same image. They also incorrectly suggest that adding Gaussian noise closes the modality gap. In contrast, DALLE-2 [24] uses a complex and heavily trained prior network to convert CLIP text embeddings to CLIP image embeddings but did not provide clear evidence of why the prior network is necessary and its effectiveness. Corgi [42] modifies the prior network by restricting its starting point, resulting in improved performance.

Despite the plethora of complex tricks proposed, there is no consensus on the best approach to interchangeably use image and text embeddings, and none of them reveal fundamental reasons or perform systematic ablations. Our work unifies previous approaches by first analyzing the multi-modal representation space and how it arises. Based on the geometric relation, we provide a straightforward method to tackle embedding interchangeability and demonstrate that our method achieves state-of-the-art results on cross-modal tasks when only training on uni-modal data.

H. Theory Proof

In this section, we provide proofs of the two lemmas used in the main paper that reveal important properties of multi-modal contrastive learning.

H.1. Proof of Lemma 1

Lemma 3. (Gradients in Contrastive Optimization) *With the mild assumption of equal presence of n images and texts with $p(x_i) = p(y_i) = 1/n$, optimizing the multi-modal contrastive loss $\mathcal{L} = -\frac{1}{2n}\sum_{i=1}^{n}\left(\log \frac{\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n}\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_j}/\tau)} + \log \frac{\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n}\exp(\boldsymbol{e}_{x_j}\cdot\boldsymbol{e}_{y_i}/\tau)}\right)$ yields the following gradients:*

$$\nabla_{\boldsymbol{e}_{x_i}}\mathcal{L} = \lambda\sum_{j=1}^{n}\alpha_{y_j}(\boldsymbol{e}_{y_j} - \boldsymbol{e}_{y_i})$$

$$\nabla_{\boldsymbol{e}_{y_i}}\mathcal{L} = \lambda\sum_{j=1}^{n}\alpha_{x_j}(\boldsymbol{e}_{x_j} - \boldsymbol{e}_{x_i})$$

where $\lambda = 1/(2n\tau)$, $\alpha_{x_j} = p(x_j|y_i) + p(y_i|x_j)$, $\alpha_{y_j} = p(y_j|x_i) + p(x_i|y_j)$, $p(x_i|y_j) = \frac{\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_j}/\tau)}{\sum_{k=1}^{n}\exp(\boldsymbol{e}_{x_k}\cdot\boldsymbol{e}_{y_j}/\tau)}$, $p(y_i|x_j) = \frac{\exp(\boldsymbol{e}_{y_i}\cdot\boldsymbol{e}_{x_j}/\tau)}{\sum_{k=1}^{n}\exp(\boldsymbol{e}_{y_k}\cdot\boldsymbol{e}_{x_j}/\tau)}$, and τ is temperature.

Proof of Lemma 1. We first prove $\forall k, \sum_{i=1}^{n} p(x_k|y_i) = 1$ using Bayes' theorem:

$$\sum_{i=1}^{n} p(x_k|y_i) = \sum_{i=1}^{n}\frac{p(y_i|x_k)p(x_k)}{p(y_i)} = \sum_{i=1}^{n}\frac{p(y_i|x_k)(1/n)}{(1/n)} = 1$$

Then, we can prove the lemma using chain rule (Figure **??**):

Similarly, we can get $\nabla_{\boldsymbol{e}_{y_k}}\mathcal{L} = \lambda\sum_{i=1}^{n}\alpha_{x_i}(\boldsymbol{e}_{x_i} - \boldsymbol{e}_{x_k})$, which finishes the proof. □

H.2. Proof of Lemma 2

Lemma 4. (Stable Region Controlled by Temperature) *We consider a single term in the multi-modal contrastive loss $\mathcal{L}_i = -\log \frac{\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n}\exp(\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_j}/\tau)}$. We define the margin $r = \boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_i} - \max_{j\neq i}\boldsymbol{e}_{x_i}\cdot\boldsymbol{e}_{y_j}$ as the measure of the similarity difference between the matched pair and the hardest negative pair. When r exceeds a threshold given below, $\mathcal{L}_i$*

$$
\begin{aligned}
&\nabla_{\boldsymbol{e}_{x_k}} \mathcal{L} \\
&= \nabla_{\boldsymbol{e}_{x_k}} \big[-\frac{1}{2n} \sum_{i=1}^{n} \big(\log \frac{\exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j}/\tau)} + \log \frac{\exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_j} \cdot \boldsymbol{e}_{y_i}/\tau)} \big) \big] \\
&= -\frac{1}{2n} \nabla_{\boldsymbol{e}_{x_k}} \big[\sum_{i=1}^{n} 2\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau - \sum_{i=1}^{n} \log(\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j}/\tau)) - \sum_{i=1}^{n} \log(\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_j} \cdot \boldsymbol{e}_{y_i}/\tau)) \big] \\
&= -\frac{1}{2n} \big[2\boldsymbol{e}_{y_k}/\tau - \nabla_{\boldsymbol{e}_{x_k}} \log(\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_k} \cdot \boldsymbol{e}_{y_j}/\tau)) - \sum_{i=1}^{n} \nabla_{\boldsymbol{e}_{x_k}} \log(\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_j} \cdot \boldsymbol{e}_{y_i}/\tau)) \big] \\
&= -\frac{1}{2n} \big[2\boldsymbol{e}_{y_k}/\tau - \sum_{i=1}^{n} \frac{\exp(\boldsymbol{e}_{x_k} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_k} \cdot \boldsymbol{e}_{y_j}/\tau)} \boldsymbol{e}_{y_i}/\tau - \sum_{i=1}^{n} \frac{\exp(\boldsymbol{e}_{x_k} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_j} \cdot \boldsymbol{e}_{y_i}/\tau)} \boldsymbol{e}_{y_i}/\tau \big] \\
&= -\frac{1}{2n\tau} \big[2\boldsymbol{e}_{y_k} - \sum_{i=1}^{n} p(y_i|x_k) \boldsymbol{e}_{y_i} - \sum_{i=1}^{n} p(x_k|y_i) \boldsymbol{e}_{y_i} \big] \\
&= \frac{1}{2n\tau} \big[\sum_{i=1}^{n} p(y_i|x_k)(\boldsymbol{e}_{y_i} - \boldsymbol{e}_{y_k}) + \sum_{i=1}^{n} p(x_k|y_i)(\boldsymbol{e}_{y_i} - \boldsymbol{e}_{y_k}) - (1 - \sum_{i=1}^{n} p(x_k|y_i)) \boldsymbol{e}_{y_k} \big] \\
&= \frac{1}{2n\tau} \big[\sum_{i=1}^{n} (p(y_i|x_k) + p(x_k|y_i))(\boldsymbol{e}_{y_i} - \boldsymbol{e}_{y_k}) \big] \\
&= \lambda \sum_{i=1}^{n} \alpha_{y_i} (\boldsymbol{e}_{y_i} - \boldsymbol{e}_{y_k})
\end{aligned}
$$

falls below a small pre-set value δ, where we assume optimization ends:

$$
r \geq \tau \log \frac{o(\tau)}{\exp(\delta) - 1},
$$

where $o(\tau)$ is a monotonically increasing function of temperature τ that satisfies $1 < o(\tau) < n$. Therefore, the required margin r is monotonically increasing with τ.

Proof of Lemma 2. We first prove $\sum_i \exp(t_i/\tau) \leq o(\tau) \exp(\max_i t_i/\tau)$, where $o(\tau)$ is a monotonically increasing function of τ that satisfies $1 < o(\tau) < n$. Let us denote $m = \arg\max_i t_i$ (no tie), we have:

$$
\sum_i \exp(t_i/\tau) = \exp(t_m/\tau)(1 + \sum_{i \neq m} \exp((t_i - t_m)/\tau))
$$

Let us denote $o'(\tau) = 1 + \sum_{i \neq m} \exp((t_i - t_m)/\tau)$, since $\forall i \neq m, t_i - t_m < 0$, we have $0 < \exp((t_i - t_m)/\tau) < 1$, so $0 < \sum_{i \neq m} \exp((t_i - t_m)/\tau) < n - 1$, therefore $1 < o'(\tau) < n$. Moreover, $\exp((t_i - t_m)/\tau)$ monotonically increases with τ, therefore $o'(\tau)$ is a monotonically increasing function of τ. We can denote $o(\tau) = \lceil o'(\tau) \rceil$.

Based on this, we have:

$$
\begin{aligned}
&\mathcal{L}_i \\
&= -\log \frac{\exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j}/\tau)} \\
&= -\log \frac{1}{1 + \sum_{j \neq i} \exp((\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j} - \boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i})/\tau)} \\
&= \log(1 + \sum_{j \neq i} \exp((\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j} - \boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i})/\tau)) \\
&\leq \log(1 + o(\tau) \max_{j \neq i} \exp((\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j} - \boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i})/\tau)) \\
&= \log(1 + o(\tau) \exp(-r/\tau))
\end{aligned}
$$

Suppose $\mathcal{L}_i \leq \log(1 + o(\tau) \exp(-r/\tau)) \leq \delta$, we have $r \geq \tau \log \frac{o(\tau)}{\exp(\delta) - 1}$, which finishes the proof. □

I. Empirical Verification of Geometry

In Section B, we established a theoretical framework for the multi-modal contrastive representation space geometry. We prove that after contrastive learning, we have $\boldsymbol{e}_x - \boldsymbol{e}_y = \boldsymbol{c}_\perp + \boldsymbol{\epsilon}$, where $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$ are ℓ_2-normalized embeddings of a paired image x and text y, $\boldsymbol{c}_\perp$ is a constant vector representing the modality gap and is orthogonal to the image and text

embedding span, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is a random Gaussian vector representing the alignment noise. Here we employ statistical methods to validate the proposed geometric structure on large pre-trained contrastive models. Figure 7 visualizes all the definitions introduced in this section.

To analyze the modality gap $\boldsymbol{c}_\perp$, we need to first find a way to isolate this vector, because it is entangled with alignment noise $\boldsymbol{\epsilon}$ in $\boldsymbol{e}_x - \boldsymbol{e}_y$. Since $\boldsymbol{\epsilon}$ is Gaussian, averaging multiple instances of $\boldsymbol{e}_x - \boldsymbol{e}_y$ should neutralize this noise, and thus isolating $\boldsymbol{c}_\perp$. Therefore, we randomly group every 100 image-text pairs into a group i. We define embedding difference for pair j in group i as $\boldsymbol{d}_j^{(i)} = \boldsymbol{e}_{x_j}^{(i)} - \boldsymbol{e}_{y_j}^{(i)}$. This difference includes both the modality gap and alignment noise. By computing the expected value, $\boldsymbol{d}^{(i)} = \mathbb{E}_j[\boldsymbol{d}_j^{(i)}]$, we effectively eliminate the noise component and leave the modality gap.

To demonstrate that the modality gap is a constant vector, we analyze its magnitude and direction across different groups. The distribution of $\|\boldsymbol{d}^{(i)}\|_2$ (Figure 8 (Gap Length)) shows that the gap has a near constant length of 0.83, while the 0.99 mean of $\cos(\boldsymbol{d}^{(i)}, \boldsymbol{d}^{(j)})$ (Figure 8 (Gap Direction)) confirms that the gap has the same direction. These findings collectively validate that the modality gap is a constant vector.

Next, we establish the orthogonality of the modality gap to the embedding spans. By demonstrating its orthogonality to the image embedding span, we implicitly confirm its orthogonality to the text embedding span, given that the modality gap is a constant vector. We define the embedding difference of image j and k as $\boldsymbol{r}_{j,k}^{(i)} = \boldsymbol{e}_{x_j}^{(i)} - \boldsymbol{e}_{x_k}^{(i)}$. We observe the distribution of $\cos(\boldsymbol{d}^{(i)}, \boldsymbol{r}_{j,k}^{(i)})$ shown in Figure 8 (Gap Orthogonality) has zero mean with a small standard deviation 0.06, which demonstrates the gap's orthogonality to embedding spans.

Finally, we address the alignment noise. We define $\boldsymbol{\epsilon}_j^{(i)} = \boldsymbol{d}_j^{(i)} - \boldsymbol{d}^{(i)}$, which eliminates the gap and only leaves the noise. We first demonstrate its zero-mean nature $\mathbb{E}_{i,j}[\boldsymbol{\epsilon}_j^{(i)}] = \mathbf{0}$ through the distribution of each dimension's mean. We see the mean of each dimension of the noise vectors is bounded between -1e-8 and 1e-8 (Figure 8 (Noise Mean)). Furthermore, we show in Figure 8 (Noise Direction) that the distribution of $\cos(\boldsymbol{\epsilon}_j^{(i)}, \boldsymbol{\epsilon}_k^{(i)})$ has zero mean and small standard deviation 0.10, indicating that noise has random directions as every two noises are likely to be orthogonal. These findings affirm that the alignment noise can be approximated by a Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

In summary, our empirical analyses support the proposed geometric model of the multi-modal contrastive representation space, with a constant modality gap vector orthogonal to embedding spans and an alignment noise characterizable as zero-mean Gaussian.

In Figure 9, we show that our analyses above fully apply to other modalities, datasets, and contrastive embedding spaces, such as image-caption (MS-COCO), audio-caption (Clotho), and video-caption (MSR-VTT) on ImageBind embeddings.

J. C^3 Algorithm

In this section, we summarize the C^3 algorithm as follows:

Algorithm 1 C^3 algorithm

Require: unpaired uni-modal dataset $\mathcal{X}, \mathcal{Y}$, fixed encoders $f_\mathcal{X} : \mathcal{X} \mapsto \mathbb{R}^d$, $f_\mathcal{Y} : \mathcal{Y} \mapsto \mathbb{R}^d$ obtained from multi-modal contrastive learning, trainable decoder $g : \mathbb{R}^d \mapsto \mathcal{Y}$, noise level σ

$\bar{\boldsymbol{e}}_x \leftarrow \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} f_\mathcal{X}(x)$
$\bar{\boldsymbol{e}}_y \leftarrow \sum_{y \in \mathcal{Y}} \frac{1}{|\mathcal{Y}|} f_\mathcal{Y}(y)$

function TRAIN(y, $f_\mathcal{Y}$, g)
  $\boldsymbol{e}_y \leftarrow f_\mathcal{Y}(y)$
  $\boldsymbol{\epsilon} \leftarrow \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$
  $\hat{y} \leftarrow g(\boldsymbol{e}_y - \bar{\boldsymbol{e}}_y + \boldsymbol{\epsilon})$
  return $\mathcal{L}(\hat{y}, y)$
end function

function TEST(x, $f_\mathcal{X}$, g)
  $\boldsymbol{e}_x \leftarrow f_\mathcal{X}(x)$
  $\hat{y} \leftarrow g(\boldsymbol{e}_x - \bar{\boldsymbol{e}}_x)$
  return $\hat{y}$
end function

K. Image Captioning

In this section, we provide additional experimental details and qualitative results of image captioning.

K.1. Experimental Setup

Model. We employ the ClipCap model architecture [18], which utilizes the CLIP ViT-B/32 [22] as the image encoder $f_\mathcal{X}$, and a mapping network with a pre-trained GPT-2 [21] as the decoder g. The mapping network is designed to handle the difference in embedding dimensions between CLIP (512-d) and GPT-2 (768-d) and to generate a "prefix" as input to GPT-2. It is implemented as a lightweight MLP with a single hidden layer, which transforms the CLIP embedding into prefix embeddings for GPT-2 to generate captions. During training, we fix the CLIP encoder to maintain the connection between image and text embeddings and only update the decoder, which includes the mapping network and GPT-2.

Data. To train and evaluate our model, we use the MS-COCO image-caption dataset [16]. We adopt the widely-

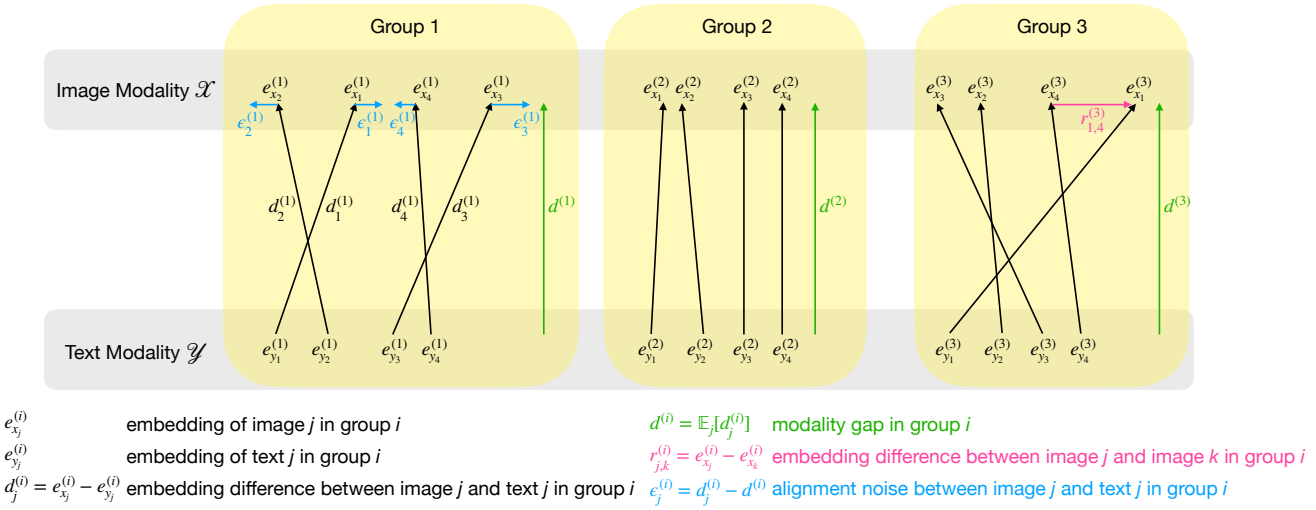


Figure 7. Visualization of the multi-modal contrastive representation space and various definitions introduced in Appendix I.

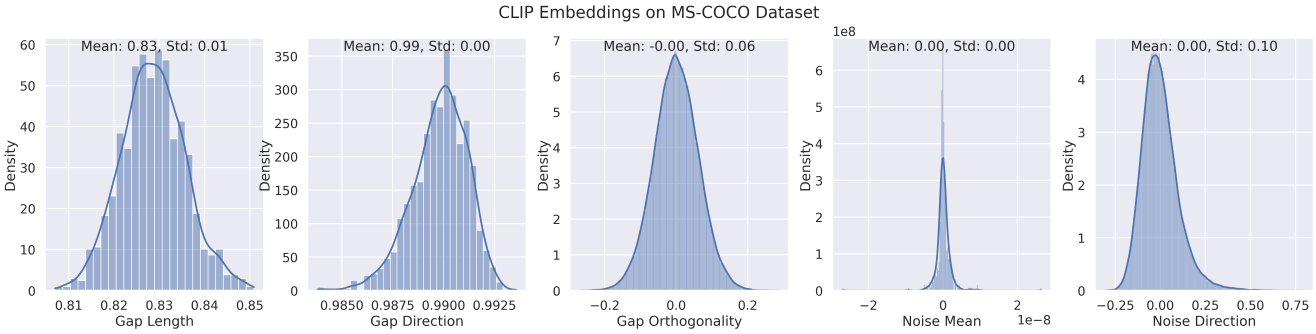


Figure 8. Empirical verification of the multi-modal contrastive representation space geometry. The modality gap approximates a constant vector, indicated by the gap length and direction distributions. The modality gap is orthogonal to the span of embeddings from two modalities, indicated by the gap orthogonality distributions. The alignment noise can be approximated by zero-mean Gaussian, indicated by the noise mean and direction distributions.

used data split [11], which consists of a training set of approximately 113K images, and validation and test sets of 5K images each, where each image has 5 ground truth captions.

Evaluation. We evaluate our model using various commonly-used image captioning metrics, including BLEU [20], METEOR [2], ROUGE [15], CIDEr [32], and SPICE [1]. These metrics measure the lexical and semantic similarities between the generated captions and the ground-truth captions.

Setup. We train our model for text reconstruction using the MS-COCO captions only. Following the C^3 algorithm, we extract the text embedding from the CLIP text encoder $f_{\mathcal{Y}}$ and apply the collapse operation (removing the pre-computed mean) and corrupt operation (adding Gaussian noise). After pre-training, we evaluate our model in the cross-modal setting. We replace the encoder with the CLIP image encoder and decode captions from image embeddings. Since no image is seen during pre-training, we refer to this evaluation setting as image-free zero-shot evaluation. Additionally, we fine-tune our model on different amounts of image-caption pairs and evaluate its performance. We refer to this evaluation setting as semi-supervised evaluation. During both the pre-training and fine-tuning stages, we train the model for 10 epochs with a batch size of 40, a learning rate of 2e-5, and AdamW [17] optimizer with a linear warmup of 5K steps. We use early stopping on the validation set and report the test set performance.

K.2. Qualitative Examples

We provide qualitative results for image captioning in Appendix Figure 10, which helps us better understand the improvements of each component of C^3. We observe that:

- C^1 generates captions that are **highly repetitive and/or**

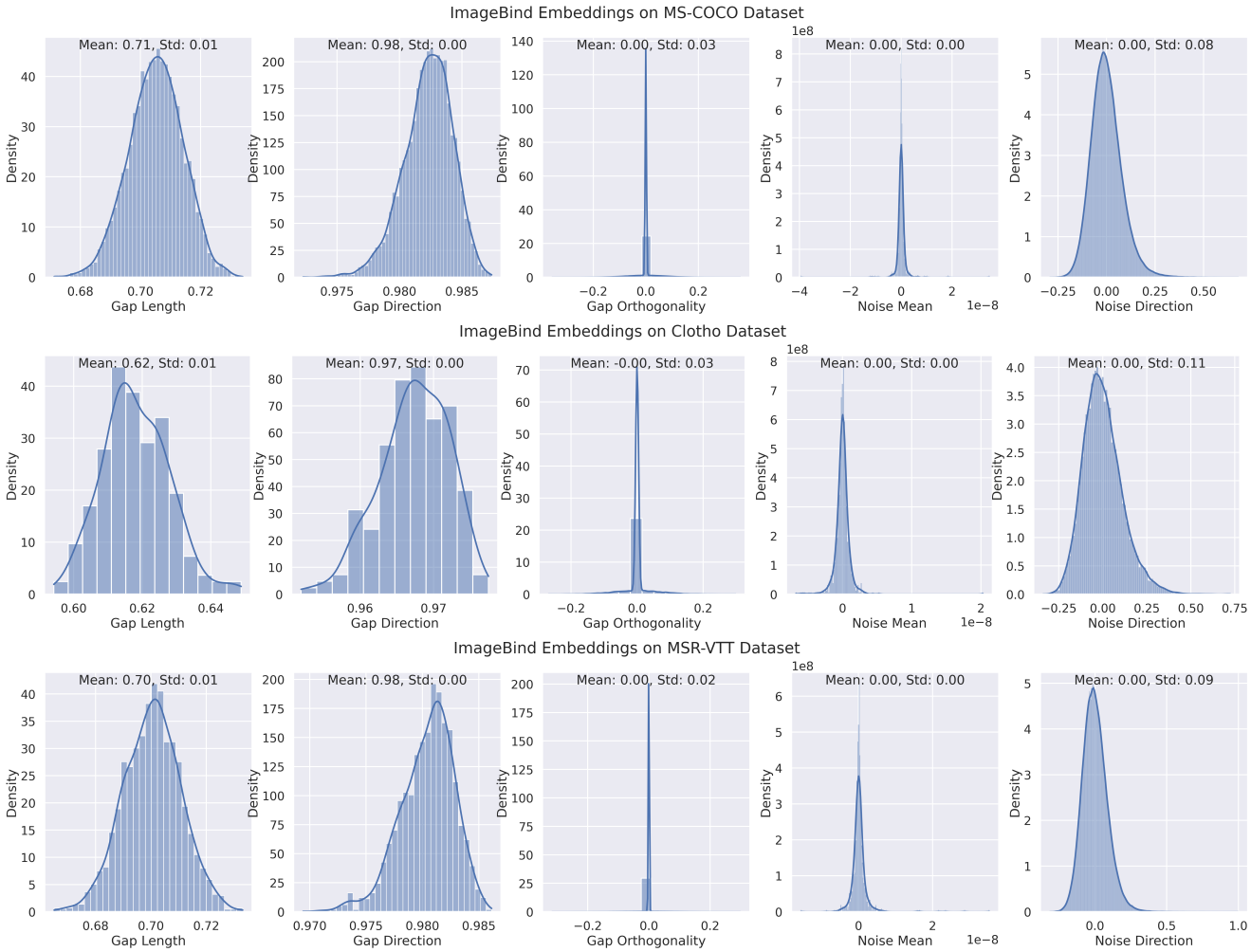


Figure 9. Generalization of Figure 8 to other modalities and multi-modal contrastive embedding spaces, including image-caption (MS-COCO), audio-caption (Clotho), and video-caption (MSR-VTT) on ImageBind embeddings.

nonsensical.

- C_1^2 and C_2^2 generates captions that are **more fluent, but contain some hallucinations**.
- C^3 generates captions that are **more correct and concise with no extraneous details**.

K.3. Quantitative Results

We include Appendix Table 6, which is the table version to produce Figure 5 in the main text. In this table, we compare the fully supervised ClipCap [18] to an ablation of components in our method C^3, and demonstrate the effectiveness of learning cross-modal tasks with uni-modal data. Results are averaged over three random seeds for 1-25% fine-tuning to reduce the effect of randomness.

L. Text-to-Image Generation

In this section, we provide additional experimental details and qualitative results of text-to-image generation.

L.1. Experimental Setup

Model. We use the LAFITE [43] model for image generation, which uses the CLIP ViT-B/32 [22] as the text encoder $f_{\mathcal{X}}$, and an adapted version of the unconditional generator from StyleGAN2 [12] as the decoder g. LAFITE's generator is adversarially trained alongside a discriminator with an additional contrastive objective to align the generator's representation space to that of CLIP's. As with image captioning, during training, we fix the CLIP encoder to maintain the connection between image and text embeddings, and only update the decoder, which in this case, involves updating both the generator and discriminator.

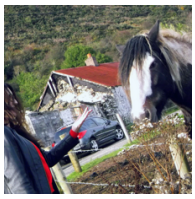
C^1: *Panglicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicliclicli....*

C_1^2: *Two sport skiers in a row look into a lake*

C_2^2: *A group of people in a canoe in the water.*

C^3: *A group of people on a boat in the water.*

Ground truth: a number of people riding on a small boat



C^1: *A van with a red unicorn horn is driving by a van that has a photo of the photographer's grandmother's van.*

C_1^2: *A very strange horse with a red wig is showing a van in a field of zombies.*

C_2^2: *A horse is standing in a field with a man.*

C^3: *A horse is standing in a field next to a house.*

Ground truth: A woman is feeding a horse at a ranch.

C^1: *A road sign that reads \"Bus Stop, Los Angeles, California, USA\".*

C_1^2: *A road sign that is closed while a man is walking and working.*

C_2^2: *A street sign is shown with a red arrow.*

C^3: *A street sign is shown on a road.*

Ground truth: A sign saying "END ROAD WORK" is in front of a truck on a road.

C^1: *A bus stops at the entrance to a European Union bus.*

C_1^2: *A train is parked alongside a passenger's luggage.*

C_2^2: *A man in a suit and tie is standing next to a train.*

C^3: *A man is waiting at a train station while a train passes him.*

Ground truth: A man is standing in front of a passing train

C^1: *A little league player catching a ball during the game of baseball at the Baltimore Orioles.*

C_1^2:*A kid that is hitting a ball in a baseball game that is on the field.*

C_2^2: *A boy is getting ready to swing at a baseball.*

C^3: *A boy is getting ready to hit a baseball while a catcher and umpire watch.*

Ground truth: a baseball player playing catcher with an umpire behind them

C^1: *Dog sniffing the back of the black and white dog in the cobblestone.*

C_1^2: *Dog behind a man's ear in the street.*

C_2^2: *A street sign is shown with a red arrow.*

C^3: *A dog looking at a black cat on the street.*

Ground truth: A black cat is scared by a large dog.

C^1: *A red and white bus is en route to the International Space Station.*

C_1^2: *A PERSON IS DRIVING THE SEA WAVE*

C_2^2: *A group of people in a boat with a large white boat in the background.*

C^3: *A group of people on a boat in the water.*

Ground truth: A boat in the water filled with people

C^1: *A person and their luggage, resting in the courtyard of the University of Southern California, observing the Eiffel Tower.*

C_1^2: *A person on a bench in the midst of the street to take a walk with their library.*

C_2^2: *A man sitting on a bench with a book.*

C^3: *A man and woman sitting on a park bench.*

Ground truth: A man and woman sitting on a park bench under a huge tree

Figure 10. Qualitative examples of image-to-text captioning on the MS-COCO dataset. C^1 generates captions that are highly repetitive and/or nonsensical. C_1^2 and C_2^2 generates captions that are more fluent, but contain some hallucinations. C^3 generates captions that are more correct and concise with no extraneous details.

Method	Connect	Collapse	Corrupted	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
1% Fine-tuning									
ClipCap	-	-	-	64.5	21.2	21.2	47.6	71.1	14.7
C^1	✓	✗	✗	66.5	22.9	22.6	48.7	78.5	15.9
C_1^2	✓	✓	✗	66.8	23.0	22.6	48.7	78.5	15.9
C_2^2	✓	✗	✓	71.7	28.8	25.3	52.7	96.3	18.3
C^3 (Ours)	✓	✓	✓	71.9	28.9	25.3	52.7	96.4	18.3
5% Fine-tuning									
ClipCap	-	-	-	70.1	26.4	24.0	51.3	89.4	17.3
C^1	✓	✗	✗	70.6	27.0	24.5	51.6	91.6	17.8
C_1^2	✓	✓	✗	70.7	27.1	24.6	51.7	92.1	17.9
C_2^2	✓	✗	✓	72.1	29.3	25.6	53.1	98.2	18.7
C^3 (Ours)	✓	✓	✓	72.3	29.4	25.7	53.1	98.8	18.8
25% Fine-tuning									
ClipCap	-	-	-	73.0	30.0	26.0	53.7	101.7	19.1
C^1	✓	✗	✗	73.2	30.4	26.1	53.9	102.8	19.4
C_1^2	✓	✓	✗	73.1	30.2	26.1	53.8	102.4	19.3
C_2^2	✓	✗	✓	73.2	30.9	26.3	54.2	103.1	19.3
C^3 (Ours)	✓	✓	✓	73.4	31.1	26.3	54.3	103.6	19.4
100% Fine-tuning									
ClipCap	-	-	-	74.0	31.5	26.8	54.7	106.6	19.9
C^1	✓	✗	✗	74.6	32.4	27.2	55.2	109.7	20.3
C_1^2	✓	✓	✗	74.6	32.3	27.1	55.2	109.2	20.1
C_2^2	✓	✗	✓	74.1	32.2	27.1	55.2	108.8	20.2
C^3 (Ours)	✓	✓	✓	74.0	32.2	27.1	55.2	108.9	20.2

Table 6. Image-to-text captioning results in the low data regime. When paired multi-modal data are limited, our approach that leverages uni-modal data for pre-training leads to substantial improvements compared to the purely supervised method (ClipCap). This is the table version used to reproduce Figure 5 in the main paper. Results are averaged over three random seeds for 1-25% fine-tuning to reduce the effect of randomness.

Data. Same as image-to-text captioning, we train and evaluate the model on the MS-COCO dataset [16]. We use the same pre-processed data and data split provided in the LAFITE official code repository, comprised of 82K training images and 40K validation images with 5 captions per image.

Evaluation. We evaluate our model using the widely-used image generation metric Frechet Inception Distance (FID) [9]. This metric measure the realism of generated images by computing feature similarity between those of generated images and those of ground-truth images, where the features are derived from the pre-trained Inception-v3 model [29]. We also report Inception Score (IS) [25], which is similar to FID. Similarly to LAFITE, FID/IS scores are computed based on 50K generated images, using captions that are randomly sampled from the validation set.

Setup. We train the model for at most 750,000 steps with a batch size of 16, a learning rate of 2.5e-3, and Adam optimizer. We initialize our model with the official pre-trained weights from LAFITE [43] that was trained on CC3M [26]. We report the validation set performance with the lowest FID score.

L.2. Qualitative Examples

We provide qualitative results for image generation in Figure 11, which helps us better understand the improvements of each component of C^3. We observe that:

- C^1 generates the **basic scene** that pertains to the text description, but are less photo-realistic in terms of contrast and color.
- C_1^2 and C_2^2 add more **fine-grained detail** despite still generating some artifacts.
- C^3 generates **sharper images** that are more **detailed with fewer artifacts**.

M. Why Align Embedding from Different Modalities?

If embeddings from different modalities are aligned, we can train a model on one modality and then infer on another modality, enabling us to build cross-modal applications with only uni-modal data. This is an emerging field

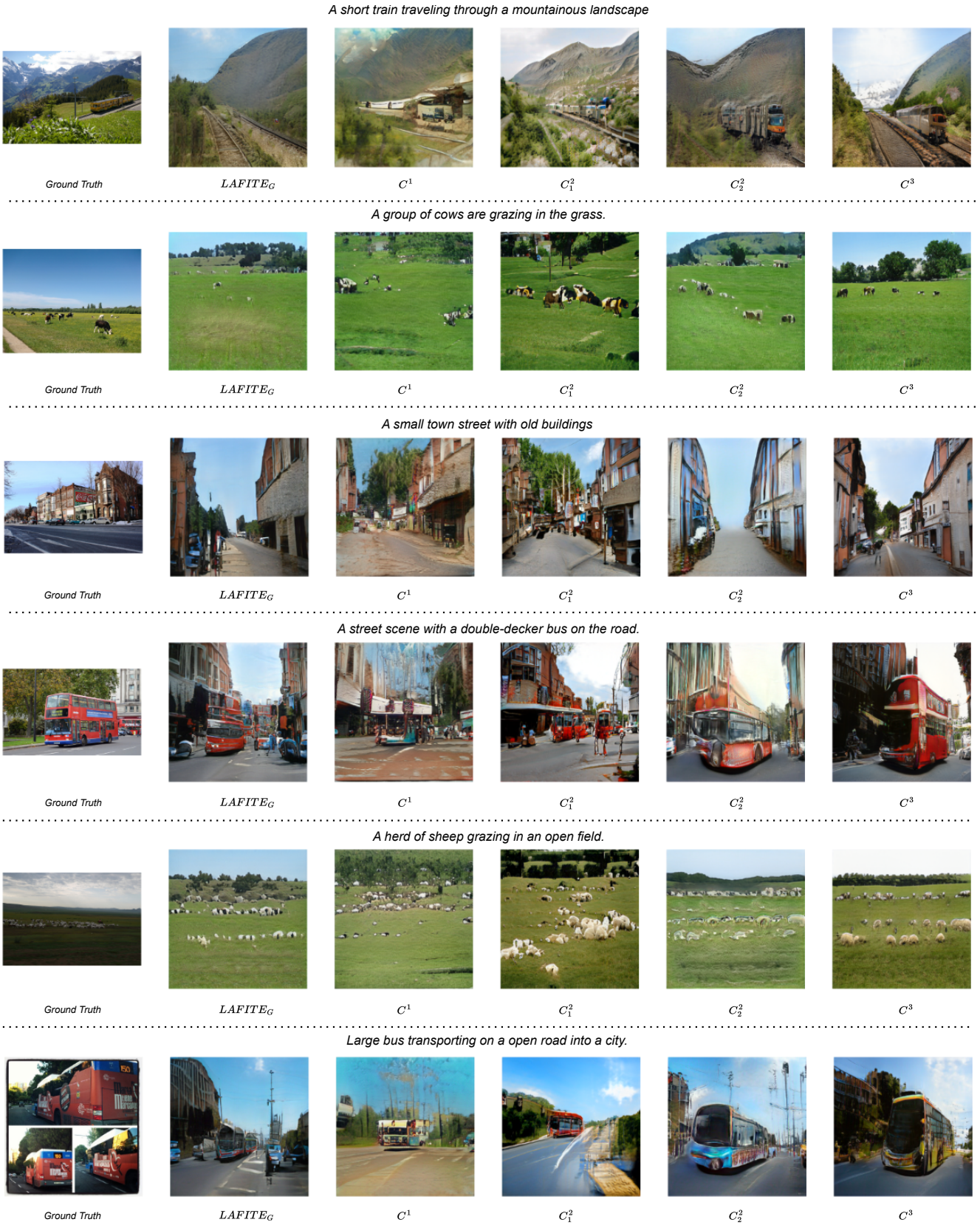


Figure 11. Qualitative examples of text-to-image generation on the MS-COCO dataset. C^1 generates the basic scene that pertains to the text description, but are less photo-realistic in terms of contrast and color. C_1^2 and C_2^2 add more fine-grained detail despite still generating some artifacts. C^3 generates sharper images that are more detailed with fewer artifacts.

that lacks principled approaches to be easily applied without requiring more empirical tuning.

We added an experiment to explain this further. We train a text generator (image captioner) over CLIP's text embeddings x. During inference, we manually shift all the x to $x + c$ to simulate the modality gap (a constant vector orthogonal to original spans). We report the captioning performance in terms of gap distance $\|c\|$ in Table 7. We observe substantial performance drops when the gap grows, showing the need to align embeddings.

Gap Distance $\|c\|$	ROUGE-1	ROUGE-L	METEOR
0.0	85.5	81.2	83.1
0.2	76.3	71.5	73.8
0.4	55.8	50.6	52.5
0.6	40.6	35.9	36.6
0.8	30.5	26.7	26.5
1.0	24.1	21.0	20.3
1.5	16.6	14.4	13.8
2.0	13.8	12.1	11.6

Table 7. Image captioning performance when trained on embedding x and tested on $x + c$. This shows the necessity to align embeddings when training a model on one modality and then inferring on another modality.

N. Effectiveness of Collapse vs Corrupt

In Table 3, we observe that adding noise (i.e., corrupt, C_2^2) is more effective than closing the modality gap (i.e., collapse, C_2^1). We hypothesize that the greater effectiveness of C_2^2 is because C_2^2 has two effects: 1) injecting noise in the span to mitigate alignment noise; 2) injecting noise in the modality gap direction to mitigate the model's sensitivity to this gap.

We have added an experiment to verify this hypothesis. When adding noise sampled from Gaussian distributions, we remove its component in the modality gap direction. Specifically, given $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, we compute its projection on the gap direction as $\epsilon_g = \frac{\epsilon \cdot g}{\|g\|} \frac{g}{\|g\|}$, where $\frac{g}{\|g\|}$ is the modality gap direction, then we remove this projection to get a new noise $\epsilon' = \epsilon - \epsilon_g$. We add this new noise during training and name this experiment as C_2^2 (span noise only).

From Table 8, we see that adding noise only in the span (i.e., C_2^2 (span noise only)) makes its performance much worse than adding noise to all the directions (i.e., C_2^2), and its performance is similar to removing the modality gap (i.e., C_1^2). Therefore, adding noise (i.e., corrupt) actually leads to a similar improvement to removing the modality gap (i.e., collapse). The reason for the greater effectiveness of corrupt than collapse is that injecting Gaussian noise not only adds noise in the span but also to the modality gap direction.

Given the substantial size of the modality gap, adding noise is not enough to fully diminish the gap. Therefore, adding noise and removing the gap (i.e., C^3) still enhance the overall performance.

O. Dimensional Collapse

Echoing our main paper, the phenomenon of dimensional collapse [10] in randomly initialized image and text encoders creates a modality gap prior to optimization that also persists after optimization. In this section, we offer further insights into dimensional collapse and demonstrate that 1) it is an inherent characteristic of deep neural networks and 2) deeper networks experience more pronounced dimensional collapse. Additionally, we establish a connection between dimensional collapse and the cone effect identified by Liang et al. [14].

O.1. Real Networks

In Figure 12, we find that the vision encoder (vision transformer) and text encoder (transformer) of the CLIP have the dimensional collapse phenomenon at initialization. Specifically, the effective dimension of the embeddings generated by the vision and text encoder is both much smaller than the full dimension, which induces a modality gap. Since the gradients of multi-modal contrastive learning will only be propagated in the effective dimensions, the modality gap will be preserved after optimization.

O.2. Simulation

To investigate the dimensional collapse phenomenon, we initialize a simple Multi-Layer Perceptron (MLP) with n blocks, where each block contains a linear layer and Rectified Linear Unit (ReLU) activation. We set the dimensionality of the input space and hidden states to 512, and initialize the weights with Xavier uniform distribution and biases to zero. We initialize $N = 1000$ input embeddings, where each dimension is sampled from a standard normal distribution $\mathcal{N}(0, 1)$. We feed these embeddings into the MLP and extract the features after every 5 layers. We perform Singular Value Decomposition (SVD) on the feature covariance matrix to quantify the degree of dimensional collapse.

As shown in Figure 13, our experiments reveal two key insights:

1. Dimensional collapse is an inherent characteristic of deep neural networks and even a simple MLP exhibits this behavior.
2. Deeper networks are more prone to dimensional collapse, resulting in a smaller effective dimension of the feature space.

O.3. Connection to Cone Effect

The dimensional collapse phenomenon can provide an explanation for the cone effect observed by Liang et al. [14].

Method	Conn.	Coll.	Corr.	BLEU-1↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑	SPICE↑
C^1	✓	✗	✗	28.1	2.4	12.2	25.4	13.0	6.8
C_1^2	✓	✓	✗	44.4	6.1	15.5	33.6	25.2	9.2
C_2^2 (span noise only)	✓	✗	✓	41.2	6.2	14.9	33.6	22.8	8.3
C_2^2	✓	✗	✓	69.0	25.5	24.3	50.8	87.6	17.6
C^3	✓	✓	✓	**71.0**	**27.7**	25.0	**52.0**	**93.3**	**18.3**

Table 8. Collapse is actually as effective as corrupt. The reason for the greater effectiveness of corrupt than collapse is that injecting Gaussian noise not only adds noise in the span but also to the modality gap direction.

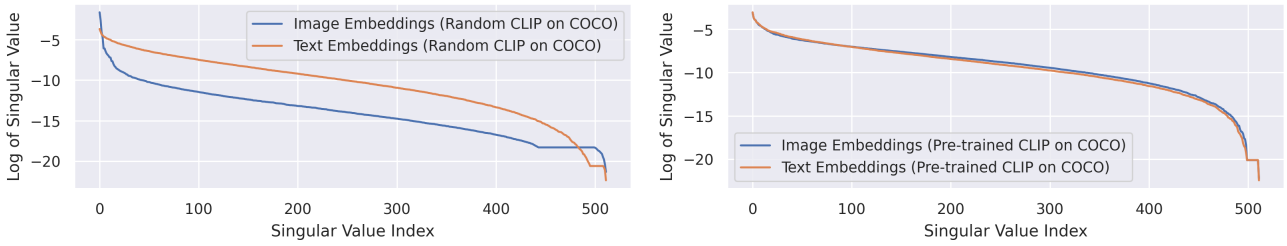


Figure 12. Dimensional collapse of the randomly initialized (left) and pre-trained (right) CLIP representation space. Singular values obtained from SVD reveal that the effective dimension of the image and text representation space is much smaller than the total number of dimensions.

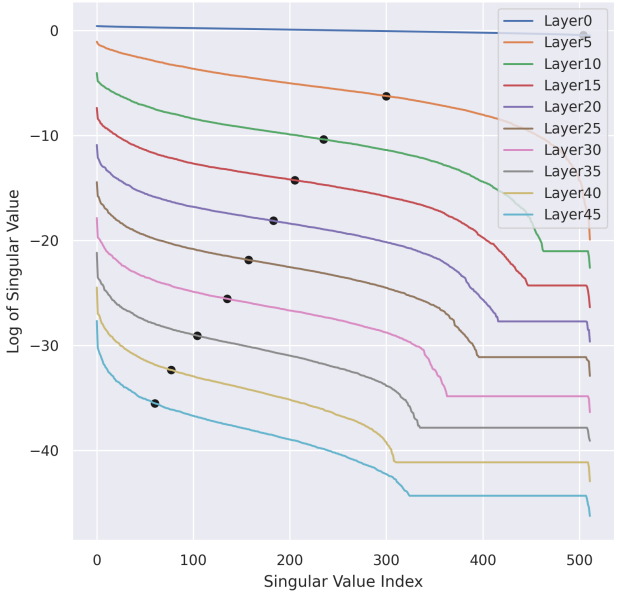


Figure 13. Simulation of dimensional collapse on a MLP (n*(ReLU(Linear))) network. The x-coordinates of the black dots indicate the effective dimensions for the embeddings from each of the layers, respectively, which quantifies the extent of collapse.

They found that the cosine similarities between any two embeddings outputted by a deep neural network were significantly higher than zero. Due to the dimensional collapse, all embeddings share similar values in the ineffective dimensions, leading to high cosine similarities. The stronger the dimensional collapse, the greater the number of ineffective dimensions, and hence, a stronger cone effect. This explanation is illustrated in Figure 14, where collapsing the z-axis of a 3D representation space induces a cone, leading to significantly higher cosine similarities between any two embeddings.

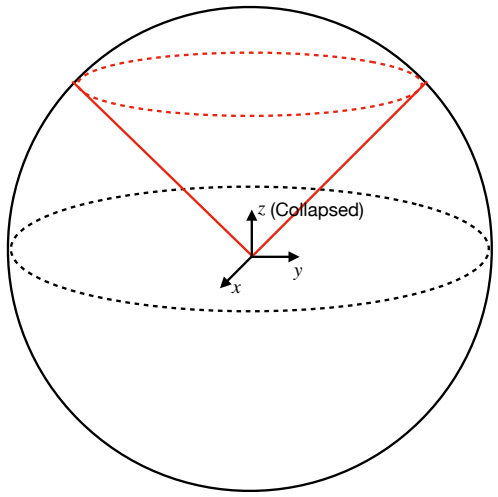


Figure 14. Dimensional collapse explains the cone effect of deep neural networks. When the z-axis of a 3D representation space is collapsed, it results in a cone shape, where the cosine similarities between any two embeddings are significantly higher than zero.