CLIP のモダリティギャップは本当に埋めるべきか？

佐藤祥太　木山朔　平澤寅庄　小町守
一橋大学
{shota, hajime, tosho, komachi}@scl.sds.hit-u.ac.jp

概要

CLIP [1] というマルチモーダル事前学習モデルには，モダリティギャップ [2] という現象の存在が報告されている．これは，画像とテキストの埋め込みが埋め込み空間内で分離してしまう現象を指す（図 1 左）．既存の研究は，モダリティギャップの存在を問題視し，このギャップを埋める手法を提案している．しかし，このギャップが埋まることで本当に zero-shot タスクの性能が向上するかは，未だ十分に検証されていない．本研究では，モダリティギャップと zero-shot タスク性能との関係について網羅的に分析する．また，モダリティギャップ全体に対する zero-shot タスクの性能に寄与する成分比がデータセットごとに異なることを示すことで，ギャップの絶対値がタスク性能を説明する指標として不十分であることを示す．

1　はじめに

マルチモーダルモデルに関する研究は長年に渡り取り組まれている [3, 4]．マルチモーダルモデルの中でも特に画像と言語を扱うモデルは，Vision-Language Model (VLM) [5, 6, 7, 4] と呼ばれ，近年特に盛んに研究されている領域の一つである [8]．VLM の代表的なモデルの一つに Contrastive Language–Image Pre-training (CLIP) [1] がある．CLIP は，画像とテキストのペアを用いて「正例を近づけて負例を引き離す」という対照学習を行うことで，画像と言語を同一の埋め込み空間へマッピングする．対照学習の性質 [9] からもわかるように，CLIP の埋め込み空間は, 同じものを指す画像とテキストの埋め込みが近くに分布しつつ，空間上には均一に分布しているのが理想的である [10]（図 1 右）．

しかし，実際の CLIP の埋め込み空間ではモダリティギャップ [2] という現象の存在が確認されている．モダリティギャップは，図 1 の左に示すように，モダリティごとに埋め込みがまとまっていて，

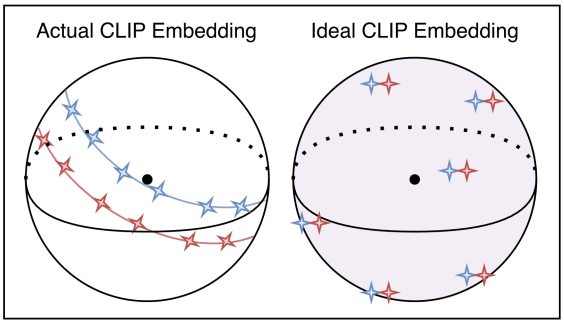

図 1　実際の CLIP の埋め込みと理想的な CLIP の埋め込みの比較．赤と青は別々のモダリティを表す．（左）実際の CLIP の埋め込みは同一モダリティが近くに配置されている [11, 12]．（右）理想的な CLIP の埋め込みは対応する画像とテキストが近くにあり，それらが散らばって配置されている [10]．

モダリティ間では埋め込みが分離している現象である．多くの研究が，このギャップの存在を問題視しており，これを縮める様々な手法を提案している [2, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]．ところが，これらの研究では「モダリティギャップは埋めるべきもの」という暗黙の前提が置かれており，モダリティギャップと zero-shot タスクの性能に関して，詳細な分析は未だ十分に行われていない．

そこで本研究では，既存のギャップ補正手法を整理し，複数のギャップ補正手法を用いてモダリティギャップとタスク性能の関係を網羅的に分析する．また，「埋めるべき」とされているモダリティギャップの指標としての限界を提示するため，zero-shot タスク性能に寄与する部分と寄与しない部分に分解する分析を行う．

本研究の貢献は，以下の通りである．

1. 提示する分類体系に基づいた 3 つの手法を用いて，全 15 種類の zero-shot 画像分類タスクによる広範な実験を行った．結果から，「モダリティギャップを縮めることは性能向上に必ずしも寄与しない」ことをより網羅的に示した．

2. zero-shot 画像分類タスクにおいて，モダリティギャップを zero-shot タスク性能に寄与する成分と寄与しない成分に分解した．その結果，データセットによってタスク性能に寄与する部分の割合が異なるため，ギャップベクトルの絶対値はタスク性能を説明する指標として不十分であることを提示した．

2 関連研究

先行研究の多くは，モダリティギャップを低減するべきだとしていた．一方で，本研究と同様に，モダリティギャップの補正と zero-shot タスクの関係性について論じている研究も一部存在する．モダリティギャップが提唱された研究 [2] では，タスクの種類によってギャップ補正の影響が異なることを実験的に示しているが，このタスク依存性についての詳細な分析は今後の課題としている．Jiang et al. [22] では，画像–テキスト検索タスクを用いてモダリティギャップの縮小が下流タスクにとって必ずしも最適でなく，対応する画像–テキストの埋め込みを完全に一致させた場合，補正前に比べて下流タスクの性能が低下することが情報理論的に示された．本研究は，これらの知見と整合的でありつつ，既存の研究では未だ十分に検証されていない，モダリティギャップの補正が zero-shot 画像分類タスク性能に及ぼす影響をより網羅的に分析することを目指す．

3 ギャップ補正実験

本節では，CLIP のモダリティギャップの大きさと zero-shot タスク性能の関係について分析する．ギャップを直接補正したり，学習によってギャップを縮めるとしている手法に対して，CLIP におけるモダリティギャップをはじめとする埋め込み空間の指標と zero-shot タスク性能の関係性について分析する．

3.1 実験設定

手法　既存のモダリティギャップ低減手法にはさまざまなものがある．本研究では，これらの手法を補正方法に基づいて「後処理」「追加学習」「事前学習」の 3 つに分類した．「後処理」は，主に学習を用いずに CLIP の内部表現を変更する分類である．「追加学習」は，CLIP をベースモデルとして，何かしらのチューニングを施すことで，ギャップの縮小を図る分類である．「事前学習」は，モデルのアーキテクチャや学習関数を一から構築し，新しいモデルとして提案されている分類である．[1)] 実験で扱う手法については，3 つに分類した手法のそれぞれから，Linear [2] , CLIPRefine [16], AlignCLIP [19] を選定する．なお，本実験で用いるモデルは，いずれも CLIP ViT-B/32 をベースとする．これらの手法の選定は，補正処理を行う段階や補正方法の性質にバリエーションを持たせることで，ギャップ補正方法の違いに依存しない傾向を探索することを意図している．

Linear [2] の実装に関しては，原論文の embedding shift を参考にして，以下のようにギャップを補正し，タスク性能との関係を分析する．

$$\mathbf{x}_i^* = \mathbf{x}_i - \alpha \mathbf{g}, \qquad \mathbf{t}_i^* = \mathbf{t}_i + \alpha \mathbf{g}, \tag{1}$$

$$\mathbf{g} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \; - \; \frac{1}{N}\sum_{j=1}^{N} \mathbf{t}_j. \tag{2}$$

ここで，N はサンプル数，$\mathbf{x}_i, \mathbf{t}_i \in \mathbb{R}^d$ はモダリティギャップ補正前の画像／テキスト埋め込み，$\mathbf{x}_i^*, \mathbf{t}_i^* \in \mathbb{R}^d$ はモダリティギャップ補正後の画像／テキスト埋め込み，$\mathbf{g} \in \mathbb{R}^d$ は画像埋め込みの平均とテキスト埋め込みの平均の差，$\alpha \in \mathbb{R}$ はギャップ補正の強さを決めるスカラー値[2)] を表す．

CLIPRefine [16] は，両モダリティの特徴をランダム参照ベクトル（共有事前分布）へ整列させる RaFA と、凍結した事前学習 CLIP 教師＋正解ラベルを混ぜた hybrid soft label で蒸留する HyCD を組み合わせて学習する．

AlignCLIP [19] は，画像とテキストで Transformer と射影層を共有（SharedCLIP）し、対応テキストの意味距離に基づいて画像同士の負例分離を調整する IMSep 損失を加えて埋め込み整列を強め、モダリティギャップを低減する手法である．

ベンチマーク　実験には，以下の 15 種類の zero-shot 画像分類タスクを使用する．Caltech101, Caltech256, CIFAR-10, CIFAR-100, DTD, EuroSAT, FGVC-Aircraft, Flowers102, Food-101, GTSRB, Oxford-IIITPet, Places365, STL-10, SVHN, imagenet-1k. [3)] [4)] 入力はデータセット内の画像と a photo of a {クラス名}. というテキストであり，出力は入力画像と最もコサイン類似度の高いテキストである．

1) 分類の結果は付録の表 1 にまとめている．
2) $\alpha > 0$ のときはギャップを縮小し，$\alpha < 0$ のときはギャップを拡大する．
3) https://docs.pytorch.org/vision/main/datasets.html
4) https://huggingface.co/datasets/ILSVRC/imagenet-1k

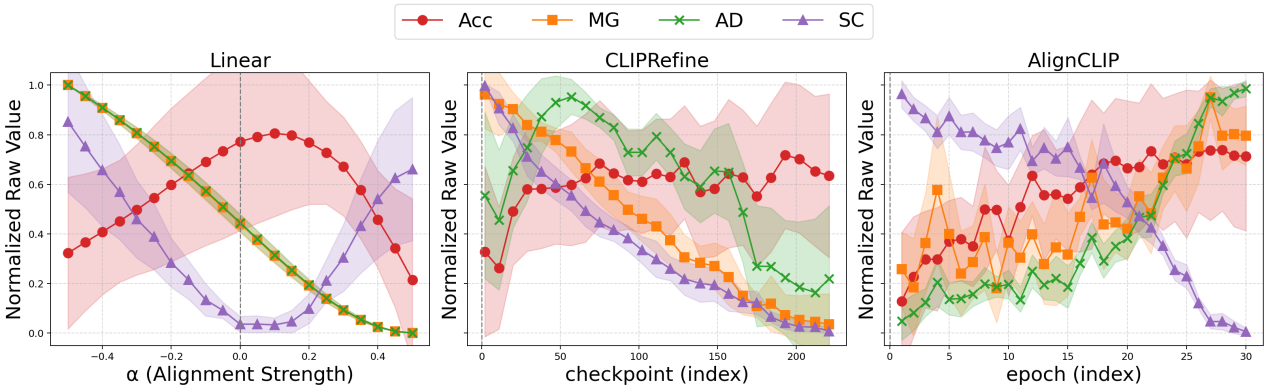

図 2　（左）Linear（後処理）による zero-shot 画像分類タスクの正解率と各評価指標の結果．（中央）CLIPRefine（追加学習）による zero-shot 画像分類タスクの正解率と各評価指標の結果．（右）AlignCLIP（事前学習）による zero-shot 画像分類タスクの正解率と各評価指標の結果．各値は，1 を最高，0 を最低として正規化されており，図中の実線が全 15 データセットでの平均値を示す．また，帯は標準偏差を示す．

評価指標　本研究では，Yamaguchi et al. [16] に倣い，埋め込み空間の性質を表す評価指標として以下の指標を用いる．

$$\text{ModalityGap} := \left\| \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \;-\; \frac{1}{N}\sum_{j=1}^{N} \mathbf{t}_j \right\|_2^2,$$

$$\text{AlignmentDistance} := \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{t}_i\|_2^2,$$

$$\text{SpatialConcentration} := \frac{1}{N_{SC}^2} \sum_{e_1,\, e_2 \in E} \exp\left(-2\,\|e_1 - e_2\|_2^2\right),$$

$$E = \{\mathbf{x}_i\}_{i=1}^{N_{\text{image}}} \cup \{\mathbf{t}_i\}_{i=1}^{N_{\text{text}}}, \qquad N_{SC} = N_{\text{image}} + N_{\text{text}}$$

Modality Gap (MG) [2] は，両モダリティの平均ベクトルの差を測っている．この値が小さいほど，モダリティギャップが小さいことを示す．**Alignment Distance (AD)** [16, 9] は，対応する画像–テキストのペアがどのくらい近くにあるかを測っている．この値が小さいほど，対応する画像とテキストが埋め込み空間上で近くに配置されることを意味している．**Spatial Concentration (SC)** [16, 9] は，空間での埋め込みの散らばり具合を定量化する指標である．この値が小さいほど，空間上に埋め込みが一様に分布していることを示す．[5), 6)]

3.2　実験結果

この節では，先述の実験設定に基づいた CLIP のモダリティギャップと zero-shot タスク性能の関係性についての実験結果を示す．ギャップが小さいにもかかわらず性能が改善しない場合（図 2 左）やギャップの改善に伴って性能も向上していく場合（図 2 中央），ギャップが大きくても性能が向上し続ける場合（図 2 右）が観測された．したがって，モダリティギャップおよびその関連指標を小さくすることは，必ずしも下流の zero-shot 画像分類性能の向上を保証するものではない．以下にそれぞれの手法の詳細を述べる．

Linear　図 2（左）は，Linear の α を動かしながら，zero-shot 画像分類タスクを解いたときの図である．α の増加に伴って，MG, AD の値はともに低下している．一方で，SC に関しては，$\alpha = 0.1$ で最良の値を取っている．タスクの性能に関しては，$\alpha = 0.1$ でピークを迎えており，その後は低下している．これらの結果から，MG と AD が改善しているにもかかわらず，下流性能が悪化するという反例が観測される．むしろ，この図からはタスク性能は SC とより関連していることが示唆される．

CLIPRefine　図 2（中央）は，CLIPRefine の学習を 100 分割して，各チェックポイントで分類タスクを解いたときの図である．学習の進行に伴って，MG と SC の値は低下していることが確認できる．AD に関しても 50 checkpoint くらいまでは悪化しているものの，その後は低下しており，学習終了時には元の値よりも低下していることが確認できる．タスクの性能に関しては，25 checkpoint 以降緩やかな右肩上がりとなっている．したがって，ギャップ関連指標の改善がそのまま性能向上として現れるケースが確認できる．

AlignCLIP　図 2（右）は，AlignCLIP の学習過程の全 30 epoch で分類タスクを解いたときの図である．15 epoch を過ぎたあたりから MG と AD が上昇

5)　本研究では画像・テキスト合わせた全埋め込みからランダムに 2,000 点を選択して値を算出している (N=2000)
6)　原論文では Uniformity として定義されている．

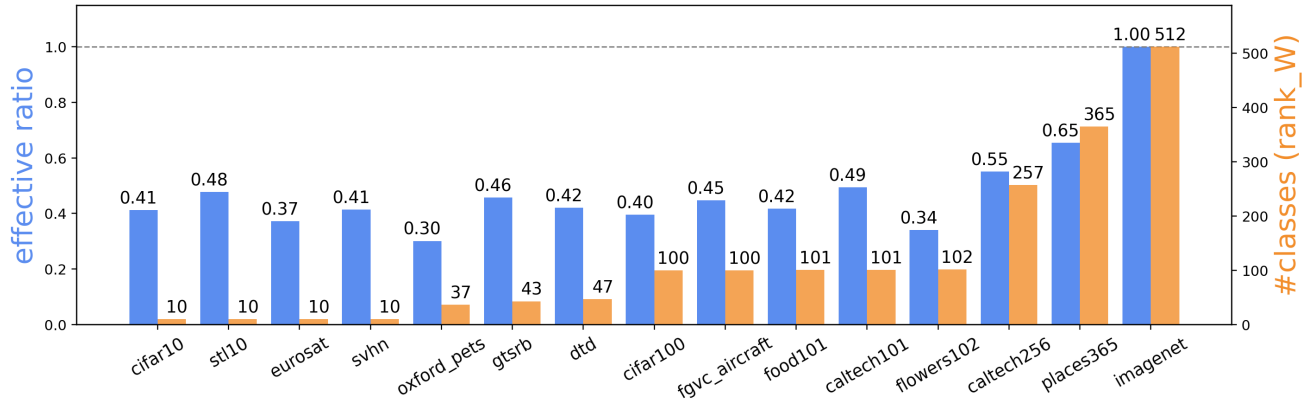

図 3　各データセットのタスクに寄与する成分の比率（青）と分類の重み行列のランク（橙）．ImageNet に関しては，次元数 (d=512) よりもクラス数の方が多いため，比率・クラス数ともに最大値となっている．なお，スピアマン係数は 0.53 であり，データセットのクラス数と effective 成分には中程度の正の相関が見られる．

しており，SC のみ低下が確認できる．しかしながら，タスクの性能は右肩上がりになっていることがわかる．つまり，ギャップが大きくても性能が向上し続けるという逆方向の例も観測される．

4　ギャップ分解分析

本節では，モダリティギャップそのものの構造を明らかにするため，Linear の手法に限定し，分類器の重み行列に基づいて式 (2) のギャップベクトルを「タスクの性能に寄与する部分」と「タスクの性能に寄与しない部分」に分解する．この分解に基づいて，zero-shot タスクの性能に関係する指標としてのモダリティギャップの限界について議論する．

4.1　分析設定

手法　各 zero-shot 画像分類タスクにおけるテキスト a photo of {クラス名}の埋め込みの行列 W に対して，特異値分解を適用する．$W = U\Sigma V^\top$ としたときの行列 V の列ベクトルを用いて，式 (2) に示すギャップベクトル $\mathbf{g}$ を射影し，これを $\mathbf{g}_{\text{effective}}$ とする．また，$\mathbf{g}$ のうちこれと直交する成分を $\mathbf{g}_{\text{ineffective}}$ とする．ここで，$\mathbf{g}_{\text{effective}}$ がタスクの性能に寄与する部分であり，$\mathbf{g}_{\text{ineffective}}$ がタスクの性能に寄与しない部分である．ギャップベクトル全体に対する各成分の割合は以下のように表される．[7)]

$$\frac{\|\mathbf{g}_{\text{effective}}\|^2}{\|\mathbf{g}\|^2} \qquad \frac{\|\mathbf{g}_{\text{ineffective}}\|^2}{\|\mathbf{g}\|^2} \tag{3}$$

評価　3.1 節で述べた 15 種類の zero-shot 画像分類タスクについて，式 (3) に示すタスク性能に寄与する成分の大きさを比較する．

7)　導出の詳細については付録 B を参照されたい．

4.2　分析結果

ギャップ分解に関して，データセット毎のタスク有効成分の割合の分析結果を以下に示す．図 3 は，式 (3) に示したタスク有効成分の寄与割合を集計した図である．タスクの性能に寄与する成分の割合とデータセットのクラス数の間には，スピアマン係数で 0.53 と中程度の正の相関が確認できた．このことから，モダリティギャップがタスク性能に及ぼす影響の大きさは，クラス数などのデータセット特性に依存して変化する可能性が示唆される．この結果から，モダリティギャップと zero-shot 画像分類タスクの関係について，zero-shot タスクの性能向上を目指す文脈では，単にモダリティ間の重心距離（幾何学的アライメント）を近づけること自体を目的化するのではなく，識別境界に対する各成分の寄与を考慮した機能的なアライメントの視点に基づく議論が必要である，ということが考えられる．

5　おわりに

本研究では，複数の手法・データセットにおいて，モダリティ間のギャップを補正することは必ずしも zero-shot タスクに良い影響を及ぼすとは限らないことを示した．また，モダリティギャップをタスクの性能に寄与する部分と寄与しない部分に分解し，モダリティギャップの絶対値が zero-shot タスク性能を説明する指標として不十分であることを示した．今後は，zero-shot タスクの性能改善という観点から，CLIP の望ましい表現空間についてより詳細に明らかにしたい．

謝辞

本研究の一部は JSPS 科研費 25K03178 および 24H00079 の助成を受けたものである．

参考文献

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[2] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 17612–17625. Curran Associates, Inc., 2022.

[3] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. **ACM Comput. Surv.**, Vol. 56, No. 10, June 2024.

[4] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. **National Science Review**, Vol. 11, No. 12, November 2024.

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In **International conference on machine learning**, pp. 19730–19742. PMLR, 2023.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 34892–34916. Curran Associates, Inc., 2023.

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 26296–26306, June 2024.

[8] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. In **2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**, pp. 1578–1597, 2025.

[9] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh, editors, **Proceedings of the 37th International Conference on Machine Learning**, Vol. 119 of **Proceedings of Machine Learning Research**, pp. 9929–9939. PMLR, 13–18 Jul 2020.

[10] Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In **ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls**, 2023.

[11] Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. **arXiv preprint arXiv:2405.18570**, 2024.

[12] Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. In Beidi Chen, Shijia Liu, Mert Pilanci, Weijie Su, Jeremias Sulam, Yuxiang Wang, and Zhihui Zhu, editors, **Conference on Parsimony and Learning**, Vol. 280 of **Proceedings of Machine Learning Research**, pp. 1365–1387. PMLR, 24–27 Mar 2025.

[13] Na Min An, Eunki Kim, James Thorne, and Hyunjung Shim. I0T: Embedding standardization method towards zero modality gap. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 27182–27199, Vienna, Austria, July 2025. Association for Computational Linguistics.

[14] Lucas Maystre, Alvaro Ortega Gonzalez, Charles Park, Rares Dolga, Tudor Berariu, Yu Zhao, and Kamil Ciosek. When embedding models meet: Procrustes bounds and applications. **arXiv preprint arXiv:2510.13406**, 2025.

[15] François Role, Sébastien Meyer, and Victor Amblard. Fill the gap: Quantifying and reducing the modality gap in image-text representation learning. **arXiv preprint arXiv:2505.03703**, 2025.

[16] Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In **Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)**, pp. 4256–4266, June 2025.

[17] Xi Yang, Pai Peng, Wulin Xie, Xiaohuan Lu, and Jie Wen. Cross-modal mapping: Mitigating the modality gap for few-shot image classification. **arXiv preprint arXiv:2412.20110**, 2024.

[18] Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In **2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 4050–4059, 2025.

[19] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, **International Conference on Representation Learning**, Vol. 2025, pp. 82070–82088, 2025.

[20] Amit Sofer, Yoav Goldman, and Shlomo E. Chazan. Pull It Together: Reducing the Modality Gap in Contrastive Learning. In **Interspeech 2025**, pp. 196–200, 2025.

[21] Xiang Ma, Xuemei Li, Lexin Fang, and Caiming Zhang. Bridging the modality gap: Dimension information alignment and sparse spatial constraint for image-text matching. In **ACM Multimedia 2024**, 2024.

[22] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 7661–7671, June 2023.

A　分類体系の表

表 1 は，モダリティギャップの存在を問題として提起し，解消するための手法として提案されているものの分類である．主に，後処理的にギャップを解消する手法，追加の学習を必要とする手法，一から学習する手法に分類できる．

表 1　既存のモダリティギャップ補正手法の分類

分類	引用	概要
後処理	Liang et al. [2]	各モダリティの平均ベクトルに重みをつけて加減
後処理	An et al. [13]	中心化+正規化の処理でモダリティギャップを補正
後処理	Maystre et al. [14]	Procrustes 変換で両埋め込みを近づける
後処理	Role et al. [15]	埋め込みの類似度グラフを用いて新しい共通空間に再埋め込み
後処理	Role et al. [15]	画像分布からテキスト分布に変換する輸送行列を学習して埋め込みを変換
追加学習	Yamaguchi et al. [16]	モダリティ間分布を近づける＋ zero-shot 性能を保持する
追加学習	Yang et al. [17]	線形変換＋トリプレット損失で大域的・局所的に空間関係を最適化
追加学習	An et al. [13]	バッチ正規化層の重みとバイアス項を用いて各モダリティの埋め込みを計算
追加学習	Fahim et al. [11]	モダリティを跨いだ Uniformity 関数を導入
追加学習	Lee et al. [18]	拡散モデルを用いて視覚埋め込みをテキスト分布へ写像する
事前学習	Eslami et al. [19]	Transformer と Projection の重みを共有する
事前学習	Eslami et al. [19]	上記に加えて画像内の埋め込みを適度に押し広げる
事前学習	Yaras et al. [12]	τ のスケジューリングにより逆温度 β の増大を抑えてギャップ閉鎖を促進
事前学習	Yaras et al. [12]	1/τ =ev に変わってより収束速度が速くなる別のパラメータ 1/τ =log(1+ev) を使う
事前学習	Yaras et al. [12]	温度の急激な変化を防ぎ，より長く大きい温度を維持する
事前学習	Yaras et al. [12]	温度を最初から大きな値で固定する
事前学習	Yaras et al. [12]	ランダムに画像とテキストのペアを選び特徴ベクトルを完全に入れ替える
事前学習	Yaras et al. [12]	画像とテキストの特徴を線形混合した新しい特徴を作る
事前学習	Amit et al. [20]	Gradient Reverse Layer を用いて，特徴量をモダリティで区別されないようにする
事前学習	Xiang et al. [21]	次元情報アラインメントとスパース空間相関アルゴリズムの併用

B　モダリティギャップベクトルの分解

テキスト埋め込みを行に持つ，分類タスクの重み行列 $W \in \mathbb{R}^{C\times d}$ の特異値分解

$$W = U\Sigma V^{\top}$$

を考える．ここで $U \in \mathbb{R}^{C\times C}$，$V \in \mathbb{R}^{d\times d}$ は直交行列，$\Sigma \in \mathbb{R}^{C\times d}$ は特異値を対角成分にもつ非負行列である．このとき，V の列ベクトル $v_1, \ldots, v_d$ は $\mathbb{R}^d$ の直交基底を与え，とくに特異値が 0 でない成分に対応する $v_1, \ldots, v_r$ は W の行空間の直交基底となる．したがって，タスク有効部分空間は

$$\mathcal{S}_{\text{effective}} = \text{span}\{\, v_1, \ldots, v_r \,\}$$

と表せる．つぎに，その直交補空間

$$\mathcal{S}_{\text{ineffective}} = \mathcal{S}_{\text{effective}}^{\perp}$$

をタスクにとって無効な部分空間とみなす．これらは直交分解になっているため，

$$\mathbb{R}^d = \mathcal{S}_{\text{effective}} \oplus \mathcal{S}_{\text{ineffective}}$$

が成り立つ．

画像埋め込みとテキスト埋め込みの平均の差として定義されるモダリティギャップベクトルを $\mathbf{g} \in \mathbb{R}^d$ とすると，任意のベクトルは直交分解により一意に

$$\mathbf{g} = \mathbf{g}_{\text{effective}} + \mathbf{g}_{\text{ineffective}},\quad \mathbf{g}_{\text{effective}} \in \mathcal{S}_{\text{effective}},\ \mathbf{g}_{\text{ineffective}} \in \mathcal{S}_{\text{ineffective}}$$

と書ける．このとき，直交性から，

$$\|\mathbf{g}\|^2 = \|\mathbf{g}_{\text{effective}}\|^2 + \|\mathbf{g}_{\text{ineffective}}\|^2$$

であり，各成分の比率は

$$\frac{\|\mathbf{g}_{\text{effective}}\|^2}{\|\mathbf{g}\|^2} \qquad \frac{\|\mathbf{g}_{\text{ineffective}}\|^2}{\|\mathbf{g}\|^2}$$

で表される．