Vision Language Model Distillation Using Partial Information Decomposition

Stephen D. Liang [1]

Abstract

Vision-Language Models (VLMs) have achieved remarkable success by integrating visual and textual modalities, enabling advancements in tasks like image captioning and multimodal retrieval. However, the substantial computational cost and large model sizes hinder their deployment in resource-constrained environments. This paper introduces a novel approach to VLM distillation by incorporating synergetic information—capturing emergent properties from the interaction between visual and textual modalities—into the distillation framework. By leveraging Partial Information Decomposition (PID), we decompose mutual information into unique, redundant, and synergistic components, explicitly optimizing the student model to retain critical multimodal interactions. Our proposed framework integrates contrastive loss, KL divergence, L2 regularization, and a synergy term into the total loss function. Experimental results demonstrate that incorporating synergetic information significantly enhances retrieval performance across image-to-text and text-to-image tasks compared to traditional distillation approaches. Although the student model (ResNet-34 with a 2-layer transformer) lags behind the teacher model (CLIP ViT-B/16) due to differences in capacity and lack of pretraining, the proposed method consistently narrows the performance gap. This work highlights the importance of synergetic information in VLM distillation and sets a foundation for future exploration into scaling student models, pretraining strategies, and optimizing synergy-driven objectives. Our findings underscore the transformative potential of synergy in developing lightweight, efficient VLMs without compromising multimodal understanding and performance.

[1]Hewlett Packard Enterprise, 6280 America Center Dr, San Jose, CA 95002, USA. Correspondence to: Stephen D. Liang <stephendliang@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Recent advances in vision-language models (VLMs) have led to powerful systems capable of understanding and generating descriptions of visual content in natural language. Models such as CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021), ALIGN (A Large-scale ImaGe and Noisy-text embedding) (Jia et al., 2021), ALBEF (Align Before Fuse) (Li et al., 2021), and BLIP (Bootstrapping Language-Image Pre-training) (Li et al., 2022) have demonstrated impressive performance on tasks including image captioning, visual question answering, and image-text retrieval. As these models become increasingly sophisticated, they also become larger and more computationally expensive, making them challenging to deploy in resource-constrained settings such as mobile devices, embedded systems, or real-time applications. To address these limitations, model compression techniques—particularly knowledge distillation (Hinton et al., 2015; Gou et al., 2021) have emerged as promising avenues to preserve model performance while reducing complexity and inference cost.

In knowledge distillation, a smaller student model learns from a larger teacher model by matching its outputs, intermediate representations, or latent distributions (Hinton et al., 2015; Gou et al., 2021; Sanh et al., 2019). Distillation has shown significant promise for language models (e.g., DistilBERT (Sanh et al., 2019)) and computer vision models, and has increasingly been applied to multimodal domains to create efficient vision-language students that inherit the capabilities of large-scale teacher models (Shen et al., 2022; Fang et al., 2021). However, the transfer of knowledge in multimodal settings is not limited to the alignment of final outputs or feature distributions; it also depends on how visual and textual information interact within the model. We posit that capturing the *synergy*, a form of complementary information that arises from the joint presence of multiple modalities beyond what is available from each modality alone, and is key to effective VLM distillation. More related work on VLM distillation is in Appendix A.

The concept of synergy has roots in information theory, where it is part of the partial information decomposition (PID) framework (Williams & Beer, 2010; Bertschinger et al., 2014). Synergy captures the emergent joint information that does not reside fully in either modality alone. In the

context of VLMs, high synergy implies that the joint visual-textual representations encode richer semantics than the separate vision-only or text-only streams. Preserving this synergy during distillation ensures that the student model can replicate the teacher's ability to integrate and interpret multi-modal cues effectively.

Our main contributions in this paper are as follows:

1. We conduct an in-depth analysis of the visual and textual modalities in VLM through PID and propose a distillation framework based on synergetic information.
2. We introduce an efficient method to compute synergetic information using cosine similarity, providing a practical and scalable solution.
3. We design a comprehensive framework that integrates contrastive loss, KL divergence, L2 regularization, and a synergy term into the overall loss function, enhancing the distillation process.
4. Our experimental results demonstrate that incorporating synergetic information significantly improves retrieval performance for both image-to-text and text-to-image tasks, surpassing the performance of traditional distillation approaches.

2. PID for VLM Distillation

2.1. Introduction to PID

Mutual information (MI) quantifies the shared information between two random variables. In the context of VLMs, mutual information between vision (V) and text (T) modalities with respect to a ground truth label Y can be expressed as (Liang, 2021; Cover & Thomas, 2006):

$$I(V;T) = \sum_{v\in V}\sum_{t\in T} p(v,t)\log\frac{p(v,t)}{p(v)p(t)} \tag{1}$$

By definition, $I(V;T)$ is non-negative, ensuring $I(V;T) \geq 0$.

Partial Information Decomposition (PID) further dissects mutual information into unique, redundant, and synergistic contributions, which is particularly useful in VLM distillation (Williams & Beer, 2010). PID enables us to analyze how visual and textual features interact to provide information about the target ground truth label Y. The decomposition of mutual information for the pair (V,T) with respect to Y is defined as (Williams & Beer, 2010):

$$\begin{aligned} I(V,T;Y) \quad = \quad & R(V,T;Y) + S(V,T;Y) \\ & + U(V;Y|T) + U(T;Y|V) \end{aligned} \tag{2}$$

Where:

- $R(V,T;Y)$ measures redundant information about Y that is captured by both V and T.
- $S(V,T;Y)$ represents the synergetic information about Y that emerges only when V and T are combined.
- $U(V;Y|T)$ quantifies the unique contribution of V in predicting Y that is not present in T.
- $U(T;Y|V)$ describes the unique information T provides about Y that is absent from V.

Figure 1 illustrates the breakdown of mutual information $I(V,T;Y)$. The red region corresponds to redundancy $R(V,T;Y)$, while the blue area represents synergy $S(V,T;Y)$. The orange and green areas highlight the unique contributions of vision and text modalities, $U(V;Y|T)$ and $U(T;Y|V)$, respectively (Williams & Beer, 2010).

The mathematical formulation for each component is given by (Williams & Beer, 2010):

$$R(V,T;Y) = I(V;T) \tag{3}$$

$$\begin{aligned} S(V,T;Y) =& I(V,T;Y) - I(V;Y|T) \\ & - I(T;Y|V) - I(V;T) \end{aligned} \tag{4}$$

$$U(V;Y|T) = I(V;Y|T) \tag{5}$$

$$U(T;Y|V) = I(T;Y|V) \tag{6}$$

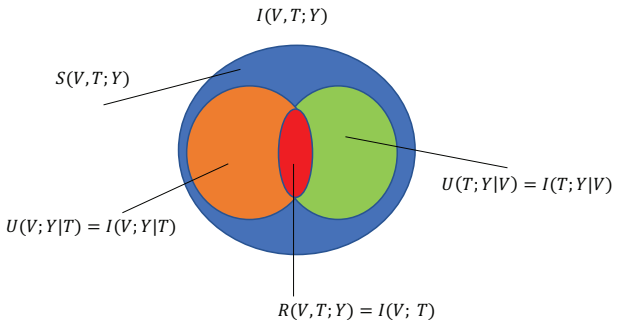

Figure 1. Partial Information Decomposition (PID) for VLM distillation. The figure shows redundant information $R(V,T;Y)$, synergistic information $S(V,T;Y)$, and unique components $U(V;Y|T)$ and $U(T;Y|V)$ (Williams & Beer, 2010).

2.2. VLM Distillation Using PID

VLMs have achieved significant success by effectively combining visual and textual modalities, enabling advanced capabilities such as image captioning, visual question answering, and multimodal retrieval. However, these models are often computationally expensive, limiting their deployment in resource-constrained environments. Distillation

offers a solution by transferring the knowledge of a large teacher model to a smaller student model while preserving performance. This section introduces a PID-driven distillation framework using synergetic information, which integrates multiple objectives to enhance the effectiveness of VLMs while maintaining computational efficiency. Synergetic information plays a crucial role in vision-language model distillation as it captures the emergent properties of multimodal interactions, which are central to the success of VLMs.

Computing synergetic information directly using mutual information, which is theoretically appealing but presents practical challenges, especially in high-dimensional settings like multimodal embeddings. We propose to compute the synergy as:

$$\mathcal{R}_{\text{synergy}} = \cos(\mathbf{s}_{\text{joint}}, \mathbf{t}_{\text{joint}}) - \frac{1}{2}\left(\cos(\mathbf{s}_I, \mathbf{t}_I) + \cos(\mathbf{s}_T, \mathbf{t}_T)\right), \tag{7}$$

where $\cos(u, v)$ denotes the cosine similarity between embeddings u and v, computed as:

$$\cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}. \tag{8}$$

Here, $\mathbf{s}_{\text{joint}}$ and $\mathbf{t}_{\text{joint}}$ represent the concatenated embeddings of image and text for the student and teacher models, respectively. Similarly, $\mathbf{s}_I$, $\mathbf{t}_I$ are the individual image embeddings, and $\mathbf{s}_T$, $\mathbf{t}_T$ are the text embeddings. This formulation captures the additional information (synergy) gained by jointly processing image and text modalities compared to treating them separately.

Our proposed distillation framework integrates PID and multiple loss functions to effectively transfer knowledge from the teacher to the student model. More loss functions are described in Appendix B.

3. Experiments

We use CLIP ViT-B/16 as the teacher model, comprising a 12-layer Vision Transformer (86M params) and a text encoder (63M params), totaling 149M parameters (OpenAI, 2021; Dosovitskiy et al., 2020). The vision encoder processes 16×16 image patches, while the text encoder produces 512-dimensional embeddings for tokenized inputs.

The student model is based on ResNet-34 (21.8M params) and a 2-layer transformer text encoder (31M params), totaling 52.8M parameters. The text encoder uses a vocabulary size of 49,408, context length of 77, and embedding dimension of 512. This makes the student a lightweight alternative suitable for knowledge distillation.

We use the MS COCO dataset, consisting of 82,783 training and 40,504 validation images (Lin et al., 2014). Each image is annotated with object segmentations and five captions, supporting multimodal learning and retrieval tasks.

In the process of VLM distillation, a widely adopted metric for this purpose is Recall@K, which assesses the model's ability to retrieve relevant items in cross-modal retrieval tasks (Patel et al., 2022). Recall@K quantifies the frequency at which the correct match appears within the top K predictions. Mathematically, Recall@K is defined as (Patel et al., 2022):

$$\text{Recall@K} = \frac{\text{Number of relevant items in top-K}}{\text{Total number of relevant items}}. \tag{9}$$

Recall@1, for instance, measures the percentage of queries where the correct result is the top-ranked output, while Recall@5 and Recall@10 expand this to the top 5 and top 10 results, respectively.

VLMs are often evaluated through two primary retrieval tasks: image-to-text and text-to-image retrieval. The training of the student model was performed for 10 epochs based on the loss function in (20). The objective was to minimize contrastive loss, KL-divergence, and L2 loss while maximizing the synergy gain. We selected hyperparameters as $\tau = 0.05$ (in contrastive loss), $\alpha = 1.0$, $\beta = 100$ (due to the small magnitude of L2 distance), and $\gamma = 1.2$.

In Fig. 2, the plot illustrates the progression of losses and synergy gain across epochs. It can be observed that the overall loss decreases consistently, indicating effective learning and convergence. We set $\tau = 0.05$ to sharpen the contrastive softmax distribution, improving separation between positive and negative pairs (Radford et al., 2021). $\beta = 100$ was used to scale the small-magnitude L2 loss, and $\gamma = 1.2$ balanced the synergy term. These values were selected based on loss trends in Fig. 2, ensuring improved alignment and stable convergence of the student model.

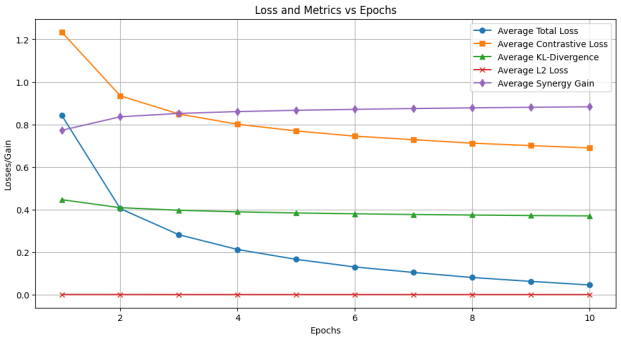

Figure 2. Loss and metric progression over 10 epochs during the training of the student model. The figure shows average loss, contrastive loss (CL), KL-divergence, L2 loss, and synergy gain.

Table 1 summarizes the recall performance at different levels (Recall@1, Recall@5, Recall@10) for both image-to-text

Table 1. Performance Comparison of Student and Teacher Models in VLM Distillation (Recall %)

Model (Loss Configuration)	Image-to-Text Retrieval		
	Recall@1	Recall@5	Recall@10
Student (Contrastive Loss)	4.36	13.69	20.55
Student (Contrastive Loss + $\beta\times$ L2 Loss)	5.04	15.70	23.35
Student (Contrastive Loss + $\alpha\times$ KL Loss)	5.73	16.95	25.24
Student (Contrastive Loss - $\gamma\times$ Synergy)	5.99	17.27	25.40
Student (Contrastive + $\alpha\times$ KL Loss - $\gamma\times$ Synergy)	6.84	19.29	27.86
Student (Contrastive + α KL Loss + β L2 Loss - γ Synergy)	7.20	19.94	28.83
Teacher (ViT-B/16)	17.81	34.09	42.46

Model (Loss Configuration)	Text-to-Image Retrieval		
	Recall@1	Recall@5	Recall@10
Student (Contrastive Loss)	3.55	11.60	17.97
Student (Contrastive Loss + $\beta\times$ L2 Loss)	4.11	13.11	19.93
Student (Contrastive Loss + $\alpha\times$ KL Loss)	5.13	15.50	23.04
Student (Contrastive Loss - $\gamma\times$ Synergy)	4.68	14.55	21.87
Student (Contrastive + $\alpha\times$ KL Loss - $\gamma\times$ Synergy)	5.66	16.76	24.64
Student (Contrastive + α KL Loss + β L2 Loss - γ Synergy)	5.79	17.08	25.14
Teacher (ViT-B/16)	14.69	29.87	38.28

and text-to-image retrieval tasks. The results demonstrate the progressive improvement achieved by adding components to the loss function. For comparison and references, we also summarize the pretrained ViT-B/16 model performance for MSCOCO dataset in Table 1.

From Table 1, we observe that the baseline student model using only contrastive loss achieves the lowest recall scores across all retrieval tasks. By introducing L2 loss ($\beta = 100$), there is a noticeable improvement, with Recall@1 increasing by approximately 0.68% for image-to-text and 0.56% for text-to-image retrieval. However, this gain remains modest compared to the introduction of synergy penalties.

The model trained with $\alpha\times$ KL loss improves performance further, particularly in Recall@1 metrics. Nevertheless, the most significant performance boost is seen when synergetic information is integrated into the loss function. The combination of contrastive loss and synergy ($\gamma = 1.0$) increases Recall@1 by 1.63% for image-to-text and 1.13% for text-to-image retrieval over the baseline.

A more refined model, incorporating KL divergence, L2 loss, and synergy penalties, achieves the highest recall scores. This configuration results in Recall@1 of 7.20% for image-to-text retrieval and 5.79% for text-to-image retrieval—both representing substantial gains over simpler loss formulations.

4. Conclusions and Future Work

In this paper, we have introduced a novel approach to VLM distillation by incorporating synergetic information, a key element often overlooked in conventional distillation methods. By decomposing mutual information using PID, we successfully isolated and leveraged the synergetic interactions between visual and textual modalities, resulting in significant performance improvements for student models.

Our primary contribution lies in highlighting the role of synergy in multimodal learning, demonstrating that while contrastive loss and KL divergence are essential, they fail to capture the emergent properties that arise from the joint processing of visual and textual inputs. The addition of synergetic penalties in the distillation framework ensures that the student model not only mimics the teacher's outputs but also learns the intricate cross-modal dependencies that drive state-of-the-art VLM performance.

The experimental results validate the effectiveness of this approach, showing consistent gains across image-to-text and text-to-image retrieval tasks. Although the teacher model (ViT-B/16) maintains a considerable advantage, our synergy-driven framework significantly narrows the gap, providing a promising direction for future research.

Future work will focus on the following directions:

1. Pretraining ResNet-34 before distillation may bridge the performance gap and provide a stronger initialization for downstream VLM tasks.

2. Investigating the resilience of synergistic distillation under noisy or incomplete data scenarios could enhance the applicability of this approach in real-world environments.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Chen, Z., Wang, W., Zhao, Z., Su, F., Men, A., and Meng, H. Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23501–23511, 2024.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2006.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fang, H., Wang, S., Zhang, Y., Zhang, H., Liu, Z., Liu, H., Wang, Y., and Xie, X. Compressing large-scale pretrained models via tiny distilled models. *arXiv preprint arXiv:2104.08481*, 2021.

Fischer, A. and Rätsch, G. The rise of information bottleneck in deep learning. *arXiv preprint arXiv:1212.2247*, 2012.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4914. PMLR, 2021.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, 2020.

Kamath, A., Singh, M., LeCun, Y., Misra, I., Synnaeve, G., and Carion, N. Mdetr–modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Kim, T., Oh, J., Kim, N., Cho, S., and Yun, S.-Y. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. *arXiv preprint arXiv:2105.08919*, 2021a. URL https://arxiv.org/abs/2105.08919.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5583–5594, 2021b.

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12888–12900, 2022.

Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., and Kong, L. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023a.

Li, M., Zhou, F., and Song, X. Bild: Bi-directional logits difference loss for large language model distillation. *arXiv preprint arXiv:2406.13555*, 2024a. URL https://arxiv.org/abs/2406.13555.

Li, X., Fang, Y., Liu, M., Ling, Z., Tu, Z., and Su, H. Distilling large vision-language model with out-of-distribution generalizability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2492–2503, 2023b.

Li, Y., Gu, Y., Dong, L., Wang, D., Cheng, Y., and Wei, F. Direct preference knowledge distillation for large language models. *arXiv preprint arXiv:2406.19774*, 2024b. URL https://arxiv.org/abs/2406.19774.

Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., and Yang, J. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26617–26626, 2024c.

Liang, S. D. Variational autoencoder for data analytics in internet of things based on transfer entropy. *IEEE Internet of Things Journal*, 8(20):15267–15275, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. URL https://arxiv.org/abs/1405.0312.

Liu, Y., Wu, C., Tseng, S.-y., Lal, V., He, X., and Duan, N. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021.

Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., and Salimans, T. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, 2023.

Naeem, M. F., Xian, Y., Zhai, X., Hoyer, L., Van Gool, L., and Tombari, F. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Najibi, M., Ji, J., Zhou, Y., Qi, C. R., Yan, X., Ettinger, S., and Anguelov, D. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8602–8612, 2023.

OpenAI. Clip model card. https://github.com/openai/CLIP/blob/main/model-card.md, 2021. Accessed: 2024-12-26.

Patel, Y., Tolias, G., and Matas, J. Recall@k surrogate loss with large batches and similarity mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7502–7511, 2022. URL https://arxiv.org/abs/2108.11179.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Sameni, S., Kafle, K., Tan, H., and Jenni, S. Building vision-language models on solid foundations with masked distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14216–14226, 2024.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Shen, Y., Liu, L., Yao, J., and Guan, Y. Multimodal knowledge distillation for efficient vision-language understanding. *arXiv preprint arXiv:2209.01573*, 2022.

Tsai, Y.-S., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. Multimodal routing: Improving information flow for vision-and-language navigation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 143–154, 2019.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

Wang, T., Zhou, W., Zeng, Y., and Zhang, X. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. *arXiv preprint arXiv:2210.07795*, 2022.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Williams, P. L. and Beer, R. D. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Wu, T., Tao, C., Wang, J., Zhao, Z., and Wong, N. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*, 2024. URL https://arxiv.org/abs/2404.02657.

Wu, X., Zhang, B., Deng, Z., and Russakovsky, O. Vision-language dataset distillation. *arXiv preprint arXiv:2308.07545*, 2023.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015.

Yang, C., An, Z., Huang, L., Bi, J., Yu, X., Yang, H., Diao, B., and Xu, Y. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15952–15962, 2024.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhao, Y., Zhao, L., Zhou, X., Wu, J., Chu, C.-T., Miao, H., Schroff, F., et al. Distilling vision-language models on millions of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13106–13116, 2024.

Zhou, A., Wang, J., Wang, Y.-X., and Wang, H. Distilling out-of-distribution robustness from vision-language foundation models. *Advances in Neural Information Processing Systems*, 36:32938–32957, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

A. Related Work

A.1. Vision-Language Models

The past few years have witnessed a surge of interest in models that jointly process visual and textual information. An overview of VLM development was presented in (Zhang et al., 2024). Early approaches focused on image captioning (Vinyals et al., 2015; Xu et al., 2015) and visual question answering (Antol et al., 2015; Anderson et al., 2018), while more recent methods leverage large-scale pretraining to learn general-purpose VLMs. Models like CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and ALBEF (Li et al., 2021) align image and text embeddings through contrastive learning, achieving state-of-the-art results in zero-shot image recognition, retrieval, and multimodal reasoning tasks. Subsequent efforts have extended these methods to integrate vision and language in more flexible ways, incorporating transformers (Kim et al., 2021b), improved objectives (Li et al., 2022), and richer multimodal datasets. Learning to promopt for VLM was discussed in (Zhou et al., 2022b), and conitional prompt learning for presented in (Zhou et al., 2022a). In the domain of video-based vision-language learning, the work by Zhao and colleagues emphasizes the importance of large-scale video datasets for distillation, showcasing how millions of videos can contribute to effective model compression (Zhao et al., 2024). Meanwhile, Chen et al. introduce a hybrid approach to domain generalization through perturbation distillation, targeting the limitations of vision-language models in multi-domain settings (Chen et al., 2024).

A.2. Knowledge Distillation in Multimodal Settings

Knowledge distillation has emerged as a powerful technique to compress large neural networks into more compact and efficient student models (Hinton et al., 2015; Gou et al., 2021). While distillation has been widely studied in single-modal tasks such as language modeling (Sanh et al., 2019; Jiao et al., 2020) and image classification (Romero et al., 2015), its application to VLMs remains relatively underexplored. Recent works have begun to address this gap by introducing methods to distill multimodal representations (Shen et al., 2022; Wang et al., 2022). For example, TinyBERT (Jiao et al., 2020) and MiniLM (Wang et al., 2020) distill language models for computational efficiency, while approaches like MDETR (Kamath et al., 2021) show how joint vision-language models can benefit from distillation techniques. One notable approach, CLIP-KD, investigates various techniques to distill knowledge from CLIP models, achieving promising results in compressing large-scale models without compromising performance (Yang et al., 2024). A different perspective is presented by Li et al., who focus on enhancing the out-of-distribution robustness of distilled models, ensuring generalizability across diverse datasets (Li et al., 2023b). These methods often focus on aligning final logits or intermediate features, but few explicitly consider the intrinsic multimodal interactions that arise from jointly modeling vision and language.

Meng et al. explore diffusion-based models by investigating the distillation of guided diffusion models, offering insights into improving the efficiency of these computationally intensive frameworks (Meng et al., 2023). Unsupervised prompt distillation for vision-language models (VLMs) was proposed in (Li et al., 2024c), while masked distillation-based VLM distillation was studied in (Sameni et al., 2024). Self-distillation was applied to vision-language pretraining in (Naeem et al., 2024), and 2D VLM distillation was leveraged for unsupervised 3D perception in autonomous driving in (Najibi et al., 2023). Vision-language dataset distillation was examined in (Wu et al., 2023), whereas preference distillation for large-scale VLMs, aimed at enhancing their ability to generate helpful and faithful responses based on visual context, was introduced in (Li et al., 2023a). Additionally, object knowledge distillation was explored to improve end-to-end VLM pretraining in (Liu et al., 2021), and out-of-distribution robustness distillation from VLMs was investigated in (Zhou et al., 2023). Collectively, these studies highlight the diverse strategies employed to advance VLM distillation, underscoring the dynamic evolution of this research area.

A.3. Information-Theoretic Perspectives and Synergy

Information-theoretic tools have been employed to understand and quantify the contributions of different modalities and their interactions. Partial Information Decomposition (PID) (Williams & Beer, 2010; Bertschinger et al., 2014) provides a framework to decompose the mutual information among multiple variables into unique, shared (redundant), and synergistic components. Synergy captures emergent information contributed by the combination of modalities that is not present in each modality alone. In the context of multimodal representation learning (Tsai et al., 2019), synergy has been linked to improved generalization and robustness. However, incorporating synergy directly into training or distillation objectives remains an open problem.

A.4. Synergy-Based Distillation Approaches

While synergy is conceptually acknowledged in multimodal learning, few works have explicitly incorporated synergy into model optimization. Preliminary attempts to incorporate information-theoretic measures into representation learning (Fischer & Rätsch, 2012; Achille & Soatto, 2018) have demonstrated that optimizing for information decomposition can enhance representation quality and interpretability. Yet, to the best of our knowledge, no existing distillation frameworks explicitly target the synergy component of teacher and student models' representations. Our work addresses this gap by integrating a synergy-based objective into VLM distillation, guiding the student model to replicate not only the teacher's output or hidden state distributions but also the emergent information that arises from the combined presence of both vision and language modalities.

B. Loss Functions for VLM Distillation

B.1. Contrastive Loss

Contrastive loss plays a pivotal role in aligning the embeddings of visual and textual modalities, ensuring that corresponding image-text pairs are brought closer in the latent space while non-matching pairs are pushed apart. This alignment facilitates the student model's ability to learn robust cross-modal representations by mimicking the teacher model's output.

The contrastive loss is formulated as (Yang et al., 2024):

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}\left(\mathcal{L}_{\text{image-to-text}} + \mathcal{L}_{\text{text-to-image}}\right), \tag{10}$$

where the loss is computed symmetrically for both image-to-text and text-to-image directions, ensuring bi-directional consistency.

The image-to-text loss component is expressed as (Yang et al., 2024):

$$\mathcal{L}_{\text{image-to-text}} = -\sum_{i=1}^{N} \log \frac{\exp(\mathbf{s}_i^\top \mathbf{t}_i/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{s}_i^\top \mathbf{t}_j/\tau)}, \tag{11}$$

where N denotes the batch size, $\mathbf{s}_i$ is the normalized embedding of the i-th image, and $\mathbf{t}_i$ is the corresponding text embedding. The temperature parameter τ controls the sharpness of the softmax distribution, which modulates the separation between positive and negative pairs.

Similarly, the text-to-image loss is defined as (Yang et al., 2024):

$$\mathcal{L}_{\text{text-to-image}} = -\sum_{i=1}^{N} \log \frac{\exp(\mathbf{t}_i^\top \mathbf{s}_i/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^\top \mathbf{s}_j/\tau)}. \tag{12}$$

In this formulation:

- $\mathbf{s}_i$ and $\mathbf{t}_i$ represent the image and text embeddings for the i-th sample, extracted from the student model.
- The numerator $\exp(\mathbf{s}_i^\top \mathbf{t}_i/\tau)$ measures the similarity between the matching image-text pair.
- The denominator aggregates similarities across all N text embeddings, effectively normalizing the score and penalizing non-matching pairs.
- The softmax operation ensures that positive pairs receive higher probabilities, driving the model to learn discriminative embeddings.

The contrastive loss encourages the model to map corresponding images and text to nearby points in the embedding space, thereby reducing the modality gap. By optimizing this loss, the student model gradually learns to capture the semantic relationships between vision and language, which is crucial for tasks like image-caption retrieval and cross-modal understanding.

The temperature τ plays a crucial role in controlling the distribution's smoothness. A lower τ sharpens the distribution, increasing the contrast between positive and negative pairs, while a higher τ results in a softer distribution, promoting smoother alignment. Empirical studies suggest that fine-tuning τ can significantly affect the convergence and performance of contrastive learning frameworks (Radford et al., 2021; Jia et al., 2021). When CLIP model was trained, τ was selected as 0.07.

Larger batch sizes provide more negative samples, enhancing the quality of contrastive learning by increasing the difficulty of the task. However, computational constraints often necessitate a trade-off between batch size and model complexity. In this paper, we choose batch size as 64.

B.2. KL-Divergence Loss

KL-divergence (Kullback-Leibler divergence) is a key measure in probability theory and statistics that quantifies the divergence between two probability distributions. In the context of vision-language model (VLM) distillation, KL-divergence is employed to minimize the difference between the output distributions of the student and teacher models (Wu et al., 2024)(Li et al., 2024b). This ensures that the student model approximates the teacher's predictions accurately.

Mathematically, KL-divergence is expressed as (Cover & Thomas, 2006):

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{13}$$

where $P(i)$ represents the teacher model's probability distribution, and $Q(i)$ represents the student model's probability distribution over the same outcomes.

In the implementation, KL-divergence is computed separately for both image and text modalities. For the image modality, the student model's logits $\mathbf{z}_I^s$ are passed through a softmax to compute log probabilities:

$$\mathbf{q}_I^s = \log \mathrm{Softmax}(\mathbf{z}_I^s) \tag{14}$$

Similarly, the teacher model's logits $\mathbf{z}_I^t$ are passed through softmax:

$$\mathbf{p}_I^t = \mathrm{Softmax}(\mathbf{z}_I^t) \tag{15}$$

The KL-divergence for the image modality is then calculated as (Kim et al., 2021a)(Li et al., 2024a):

$$D_{\mathrm{KL}}(\mathbf{p}_I^t \parallel \mathbf{q}_I^s) = \sum_i \mathbf{p}_I^t(i) \log \frac{\mathbf{p}_I^t(i)}{\mathbf{q}_I^s(i)} \tag{16}$$

For the text modality, the process mirrors the image calculation (Kim et al., 2021a):

$$D_{\mathrm{KL}}(\mathbf{p}_T^t \parallel \mathbf{q}_T^s) = \sum_i \mathbf{p}_T^t(i) \log \frac{\mathbf{p}_T^t(i)}{\mathbf{q}_T^s(i)} \tag{17}$$

where $\mathbf{p}_T^t$ and $\mathbf{q}_T^s$ denote the teacher and student distributions for text inputs, respectively.

The overall KL-divergence loss is computed as the average of the image and text modality losses (Kim et al., 2021a)(Li et al., 2024a):

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2} \left(D_{\mathrm{KL}}(\mathbf{p}_I^t \parallel \mathbf{q}_I^s) + D_{\mathrm{KL}}(\mathbf{p}_T^t \parallel \mathbf{q}_T^s) \right) \tag{18}$$

Minimizing $\mathcal{L}_{\mathrm{KL}}$ aligns the output distributions of the student model with those of the teacher, promoting effective distillation. This method allows the student model to learn the intricate vision-language relationships encoded by the teacher, enhancing the performance and generalizability of the distilled model.

B.3. L2 Distance Loss (Mean Squared Error)

L2 distance loss, also referred to as Mean Squared Error (MSE) loss, minimizes the feature-level discrepancy between the embeddings produced by the teacher and student models. This promotes tighter alignment between the two, ensuring that

the distilled student model learns representations close to those of the teacher. The L2 loss can also be expressed as (Yang et al., 2024)(Kim et al., 2021a):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left((\mathbf{s}_T^{(i)} - \mathbf{s}_S^{(i)})^2 + (\mathbf{t}_T^{(i)} - \mathbf{t}_S^{(i)})^2\right), \tag{19}$$

which represents the mean squared error between the teacher and student embeddings over a batch of size N.

This loss penalizes large deviations quadratically, encouraging the student model to closely approximate the teacher's embeddings. A smaller L2 (MSE) loss indicates higher alignment between teacher and student representations, reducing feature-level errors.

B.4. Loss Functions in VLM Distillation

The overall objective is to minimize the discrepancy between the teacher and student outputs while encouraging synergistic learning across modalities. The total loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrastive}} + \alpha\mathcal{L}_{\text{KL}} + \beta\mathcal{L}_{\text{L2}} - \gamma\mathcal{R}_{\text{synergy}}, \tag{20}$$

where each component serves a unique role in optimizing the student's performance. The hyperparameters α, β, and γ balance the contributions of different losses and synergy rewards.

The hyperparameters α, β, and γ are empirically tuned to balance the contribution of each loss component. While contrastive loss drives primary alignment between modalities, KL divergence promotes distribution matching, L2 loss ensures feature-level imitation, and synergy enhances joint learning. This comprehensive loss formulation improves the robustness and generalizability of the distilled student VLM.