INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 1 (2026)

https://ijmle.org/

Robust Image-Text Matching Through Capacity-Aligned Dense-to-Sparse Feature Distillation

Andy Clara

Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract:

The rapid advancement of vision-language modeling has significantly improved the performance of multi-modal systems, particularly in the domain of image-text matching. However, achieving state-of-the-art accuracy typically requires computationally expensive architectures that process dense feature representations across multiple cross-attention layers. This complexity poses significant challenges for deployment in resource-constrained environments. While sparse representation learning offers a promising avenue for reducing computational overhead, directly distilling knowledge from a high-capacity dense teacher model to a low-capacity sparse student model often results in severe performance degradation. This degradation is primarily attributed to the representational capacity gap and the inherent topological differences between dense and sparse feature spaces. To address these challenges, this paper proposes a novel framework titled Capacity-Aligned Dense-to-Sparse Feature Distillation. The proposed methodology introduces a structural alignment mechanism that explicitly bridges the dimensionality and distribution gaps between teacher and student representations prior to the distillation process. By projecting dense contextualized features into an intermediate capacity-aligned subspace, the framework enables the sparse student model to efficiently absorb rich cross-modal relationships without being overwhelmed by representational complexity. Extensive experimental evaluations conducted on standard benchmark datasets demonstrate that the proposed method significantly outperforms traditional distillation techniques. The framework recovers the majority of the performance drop typically associated with feature sparsification, establishing a new Pareto frontier for the trade-off between inference efficiency and retrieval accuracy in multi-modal architectures.

Keywords: *Image-Text Matching, Feature Distillation, Sparse Representations, Capacity Alignment.*

INTRODUCTION

The ability to computationally understand and relate visual and textual information stands as a foundational pillar in the pursuit of artificial general intelligence [1]. Human cognition seamlessly integrates multiple sensory inputs, allowing individuals to effortlessly associate a visual scene with its corresponding linguistic description [2]. Replicating this cross-modal associative capability within machines has driven the development of the image-text matching paradigm, a core task in multi-modal machine learning [3]. Image-text matching essentially requires an algorithm to

measure the semantic similarity between an image and a text snippet, enabling critical applications such as cross-modal retrieval, zero-shot image classification, and automated image captioning [4]. Over the past decade, the dominant approach to solving this problem has shifted from projecting global image and text embeddings into a shared latent space to employing sophisticated, fine-grained cross-attention mechanisms that model the complex relationships between specific image regions and individual words [5]. Despite the remarkable success of fine-grained, region-word alignment models, their architectural complexity introduces severe computational bottlenecks [6]. Modern state-of-the-art vision-language models typically rely on deep transformer architectures that generate dense representations. In these dense networks, every visual patch is allowed to attend to every textual token, and vice versa, through multiple layers of multi-head self-attention and cross-attention [7]. While this dense interaction matrix captures highly nuanced contextual dependencies, it results in a computational complexity that scales quadratically with the length of the input sequences [8]. Consequently, the inference latency and memory consumption of these high-performing dense models make them prohibitively expensive for real-time applications or deployment on edge devices with limited computational budgets [9]. To mitigate these heavy computational demands, researchers have increasingly explored sparse representation learning [10]. Sparse architectures aim to identify and retain only the most semantically salient tokens, discarding redundant or uninformative background patches and stop words [11]. By drastically reducing the sequence length of the representations involved in cross-modal alignment, sparse models achieve significant accelerations in inference speed [12]. However, the transition from dense to sparse representations introduces a substantial drop in matching accuracy. The discarded tokens, while seemingly redundant, often carry subtle contextual clues that are essential for resolving semantic ambiguities in complex scenes [13]. Knowledge distillation has emerged as a standard technique to improve the performance of lightweight student models by leveraging the learned expertise of a larger, pre-trained teacher model [14]. In standard feature distillation, the intermediate representations of the student model are trained to mimic the corresponding representations of the teacher model [15]. When applied to the dense-to-sparse paradigm, however, conventional knowledge distillation encounters a fundamental obstacle known as the capacity gap [16]. A dense teacher model possesses a vast representational capacity, encoding information in a high-dimensional, highly entangled feature space [17]. Conversely, a sparse student model, by design, operates with a constrained representational capacity [18]. When a student network with limited capacity is forced to directly mimic the complex, dense distributions of the teacher, the optimization process becomes unstable [19]. The student struggles to converge, often resulting in sub-optimal feature embeddings that fail to capture the critical cross-modal alignments necessary for accurate image-text matching. This paper identifies the representational capacity gap as the primary source of performance degradation in dense-to-sparse knowledge distillation for image-text matching. To overcome this limitation, we introduce a comprehensive framework: Capacity-Aligned Dense-to-Sparse Feature Distillation. The core philosophy of this approach is that effective knowledge transfer cannot occur if the student is structurally incapable of mapping the teacher's feature space. Therefore, we propose an intermediate capacity alignment module that acts as a translator between the dense and sparse domains. This module systematically compresses and restructures the dense teacher features into a manageable, capacity-aligned subspace that perfectly matches the architectural constraints of the sparse student. By performing the distillation within this aligned intermediate space, the student model can efficiently absorb the essential cross-modal semantic relationships without being overwhelmed by the high-frequency noise and extreme dimensionality of the original dense representations. The primary contributions of this research are heavily rooted in both architectural innovation and comprehensive empirical validation. We propose a structurally novel knowledge distillation paradigm tailored specifically for multi-modal

systems, introducing an explicit bottleneck mechanism that forces the teacher's representations into a lower-capacity distribution before distillation occurs. Furthermore, we design a token-selective alignment strategy that maps aggregated dense contexts to selected sparse tokens, ensuring that the contextual richness of the dense model is preserved even when the absolute number of tokens is reduced. Through extensive experimentation on standard benchmarks, we provide a detailed analysis demonstrating that the proposed capacity-aligned distillation significantly narrows the performance gap between heavy dense models and lightweight sparse models, offering a highly robust and computationally efficient solution for real-world image-text matching applications.

2. Literature Review

2.1 The Evolution of Image-Text Matching Architectures

The trajectory of image-text matching research has been characterized by a continuous effort to better capture the semantic correspondence between visual pixels and linguistic semantics. Early computational approaches predominantly relied on global embedding techniques [20]. In these foundational frameworks, an image was typically processed by a convolutional neural network to extract a single, fixed-length feature vector, while the corresponding text sentence was processed by a recurrent neural network to produce a singular textual vector [21]. These global vectors were then mapped into a joint semantic embedding space, and the model was optimized using triplet ranking loss functions to ensure that matching image-text pairs were closer in this space than non-matching pairs [22]. While computationally highly efficient, these global alignment methods inherently suffered from a lack of fine-grained understanding. By compressing an entire image or a complex sentence into a single vector, these models inevitably lost critical localized information, rendering them incapable of distinguishing nuanced differences in scenes involving multiple interacting objects. To address the limitations of global embeddings, the field experienced a paradigm shift towards region-word fine-grained alignment [23]. Researchers began utilizing object detection models to extract localized features from salient image regions, which were then individually compared against the word-level embeddings of the text [24]. This localized approach allowed models to calculate similarity scores based on the aggregation of region-word pairs, significantly improving retrieval accuracy on complex datasets. The introduction of the transformer architecture revolutionized this fine-grained alignment process [25]. By leveraging cross-attention mechanisms, multi-modal transformers enabled deep, contextualized interactions between visual regions and textual tokens at multiple levels of abstraction [26]. These dense interaction models achieved unprecedented state-of-the-art results, as every single visual patch could dynamically attend to every word in a sentence, capturing incredibly intricate semantic relationships [27]. However, the profound success of dense multi-modal transformers introduced a new, critical challenge: immense computational overhead. The cross-attention mechanism requires computing a similarity matrix between all visual tokens and all textual tokens [28]. As the resolution of images increased and the lengths of captions grew, the quadratic complexity of this dense interaction became a severe bottleneck. The computational cost made these models highly impractical for large-scale retrieval tasks where millions of images must be searched in real-time. Consequently, the research community recognized an urgent need to develop architectures that could maintain the fine-grained understanding of dense transformers while operating with the efficiency of earlier, simpler models.

2.2 Knowledge Distillation in Multi-modal Systems

Knowledge distillation was initially conceptualized as a model compression technique designed to transfer the generalized knowledge of a large, cumbersome teacher model into a smaller, more efficient student model [29]. In its original formulation for unimodal classification tasks, distillation primarily focused on matching the output probability distributions, or logits, of the

teacher and student networks. The underlying theory posits that the soft probabilities produced by the teacher contain valuable dark knowledge about the relationships between different classes, which provides a richer training signal than hard ground-truth labels alone. As the complexity of neural architectures grew, researchers expanded the scope of knowledge distillation to include intermediate feature representations. Feature-based distillation attempts to align the internal hidden states or attention maps of the student with those of the teacher. In the realm of multi-modal systems, feature distillation has proven particularly crucial. Because image-text matching relies heavily on the quality of the joint representation space, aligning the intermediate cross-modal features of the student with a highly optimized teacher can significantly enhance performance. Previous studies have demonstrated that transferring the cross-attention matrices from a heavy transformer teacher to a lightweight transformer student improves the student's ability to localize relevant objects in relation to specific words. Despite these advancements, applying standard feature distillation to multi-modal networks is not without significant hurdles. The primary difficulty arises when the teacher and student have fundamentally different architectures or operate on different representational paradigms. When distilling knowledge from a dense model, which utilizes hundreds of tokens, to a sparse model, which utilizes only a fraction of that number, standard one-to-one feature matching becomes impossible. Previous attempts to solve this involve applying average pooling or naive linear projections to match dimensions. However, these simplistic aggregation methods often destroy the structural integrity of the teacher's feature space, leading to suboptimal knowledge transfer and limiting the potential of the sparse student model.

2.3 Sparse Representations and the Capacity Gap

Sparse representation learning is a highly effective strategy for reducing the computational complexity of deep neural networks. In vision-language modeling, sparsity is typically achieved through token pruning or token merging techniques. Token pruning involves evaluating the semantic importance of each visual and textual token and dynamically discarding those that fall below a certain threshold. For instance, patches corresponding to uninformative backgrounds or stop words are removed early in the network hierarchy. By progressively reducing the number of tokens processed in subsequent layers, sparse models drastically reduce the computational burden of cross-attention mechanisms. However, achieving high accuracy with sparse representations is challenging due to the loss of contextual information. Even seemingly uninformative tokens can contribute to the overall contextual embedding of a scene. When these tokens are abruptly removed, the model's ability to resolve semantic ambiguities is impaired. To recover this lost performance, researchers often turn to knowledge distillation, attempting to teach the sparse model using a dense teacher. This brings us to the core problem addressed in this paper: the representational capacity gap. The capacity gap refers to the disparity in the volume and complexity of information that two different neural networks can encode. A dense teacher model with a vast number of parameters and tokens possesses a high representational capacity. Its feature space is characterized by high dimensionality and complex manifold structures. A sparse student model, by virtue of having fewer tokens and potentially fewer layers, has a strictly limited representational capacity. When conventional distillation methods force the low-capacity student to directly mimic the high-capacity teacher, the student network becomes overwhelmed. The objective function tries to map a highly complex distribution into a simpler, constrained space, leading to optimization instability and severe gradient conflicts. The student fails to learn the most salient features because it is continuously penalized for not capturing the intricate details that it physically lacks the capacity to store. Overcoming this capacity mismatch requires a paradigm shift in how distillation is applied, necessitating an intermediate alignment step that bridges the gap between dense complexity and sparse efficiency.

3. Methodology

3.1 Framework Overview

The proposed Capacity-Aligned Dense-to-Sparse Feature Distillation framework is meticulously designed to facilitate highly effective knowledge transfer between two fundamentally different multi-modal architectures. The system consists of three primary components: a pre-trained, high-capacity dense teacher model; a lightweight, sparse student model; and the core capacity alignment distillation module. The fundamental objective of this framework is to train the sparse student model to achieve an image-text retrieval performance that closely approximates that of the dense teacher, while maintaining strict computational efficiency. Initially, an image-text pair is processed by the dense teacher model, which utilizes a full complement of visual patches and textual tokens to compute highly detailed, context-rich cross-modal representations. Concurrently, the same image-text pair is fed into the sparse student model. The student model incorporates a dynamic token selection mechanism that aggressively prunes redundant tokens early in its processing pipeline, resulting in a sparse, compact set of internal representations. Because direct distillation between the high-dimensional dense features and the low-dimensional sparse features causes severe capacity mismatch issues, we intercept the teacher's representations before the distillation loss is calculated. The intermediate capacity alignment module compresses and restructures the teacher's dense features into a customized subspace. This newly formed subspace is specifically tailored to match the statistical distribution and structural capacity of the student's sparse features. By computing the distillation loss exclusively within this aligned environment, the framework ensures stable optimization and highly efficient knowledge transfer.

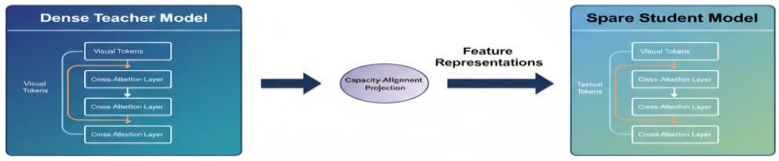


Figure 1: Architectural Framework

3.2 Dense Teacher Representation Extraction

The first phase of the methodology involves generating the target representations using the pre-trained dense teacher model. The teacher model is instantiated as a deep, multi-layer transformer architecture comprising an independent image encoder, an independent text encoder, and a subsequent multi-modal cross-attention encoder. When an image is input into the vision encoder, it is partitioned into a comprehensive grid of non-overlapping patches. Each patch is linearly projected into a high-dimensional feature vector, resulting in a dense sequence of visual tokens. Similarly, the input text sentence is tokenized into sub-word units, which are subsequently processed by the text encoder to yield a dense sequence of textual tokens. These dense unimodal sequences are then concatenated or fed simultaneously into the cross-modal attention layers. Within these layers, profound semantic interactions occur. The self-attention and cross-attention

operations allow every single visual token to continuously query and aggregate information from every textual token, and vice versa. Through multiple successive layers of this dense interaction, the model builds an exceptionally intricate understanding of the relationships between visual regions and linguistic concepts. The output of the final cross-modal layer consists of a set of dense, highly contextualized feature embeddings. These embeddings represent the gold standard of semantic alignment for our framework. However, because they are intrinsically tied to the full sequence length and the vast parameter space of the teacher, they contain levels of detail and noise that are mathematically incompatible with the restricted architecture of the sparse student model.

3.3 Sparse Student Architecture and Token Selection

The sparse student model operates on an entirely different computational paradigm, prioritizing speed and efficiency over exhaustive feature extraction. The student model follows a similar high-level structural design to the teacher, possessing unimodal encoders and a cross-modal interaction module. However, the critical distinction lies in the implementation of a dynamic token pruning mechanism embedded within the early layers of the student network. As the image and text pass through the initial layers of their respective student encoders, a lightweight scoring network evaluates the semantic salience of each individual token. For visual data, tokens corresponding to vast expanses of sky, blurred backgrounds, or featureless textures receive low salience scores. For textual data, grammatical stop words and highly redundant adjectives are similarly down-weighted. Based on a predefined sparsity ratio, the network strictly filters the sequences, retaining only the top percentile of the most informative tokens. This pruning operation drastically reduces the sequence length. Consequently, when these filtered sequences reach the cross-modal attention layers of the student model, the computational cost of the attention matrix is exponentially lower than that of the teacher. The student model generates sparse output representations that capture the core semantic meaning of the image-text pair using only a fraction of the computational resources. However, this aggressive pruning inherently discards subtle contextual links. The following code snippet illustrates the algorithmic process of projecting dense features into a format suitable for sparse mapping.

Code Listing 1: Dense to Sparse Feature Projection Algorithm

```
import torch
import torch.nn as nn
class CapacityAlignmentModule(nn.Module):
    def __init__(self, dense_dim, sparse_dim, bottleneck_dim):
        super(CapacityAlignmentModule, self).__init__()
        self.down_project = nn.Linear(dense_dim, bottleneck_dim)
        self.activation = nn.GELU()
        self.up_project = nn.Linear(bottleneck_dim, sparse_dim)
        self.layer_norm = nn.LayerNorm(sparse_dim)
    def forward(self, dense_features, selection_indices):
        compressed_features = self.down_project(dense_features)
        compressed_features = self.activation(compressed_features)
        aligned_features = self.up_project(compressed_features)
        aligned_features = self.layer_norm(aligned_features)
        batch_size = dense_features.size(0)
        sparse_matched_features = []
        for i in range(batch_size):
            selected = torch.index_select(
                aligned_features[i],
                dim=0,
```

```
            index=selection_indices[i]
        )
        sparse_matched_features.append(selected)
    return torch.stack(sparse_matched_features)
```

3.4 Capacity-Aligned Distillation Mechanism

The capacity alignment module is the pivotal innovation of this research. Directly calculating a mean squared error between the teacher's dense features and the student's sparse features is impossible due to varying sequence lengths. Furthermore, even if one were to pool the dense features to match the sparse token count, the dense features reside in a highly complex representational manifold. Forcing the student to mimic this manifold directly results in capacity overload, where the student's gradients fluctuate wildly and training collapses. To solve this, we introduce an explicit bottleneck architecture within the distillation pathway. The teacher's dense features are first passed through a non-linear projection network that deliberately constricts the feature dimensionality down to a narrow bottleneck before expanding it back to match the student's dimensional specifications. This bottleneck forces the teacher's representations to shed high-frequency, complex noise and retain only the most robust, easily generalized semantic structures. By compressing the dense features, we are effectively lowering their representational capacity to align with what the student is biologically capable of absorbing. Once the teacher features are projected into this capacity-aligned subspace, a matching algorithm is employed to align the sequences. Since the student model has selected a specific subset of salient tokens, we use the student's selection indices to sample the corresponding tokens from the capacity-aligned teacher features. This ensures that the distillation process is strictly localized; the student token representing a specific dog is taught specifically by the capacity-aligned teacher token representing that exact same dog. The overall training objective combines standard task losses with the proposed distillation loss. The task losses ensure that the student model learns to perform image-text matching correctly based on the ground truth data. These typically include a cross-modal contrastive loss, which pushes the global representations of matching pairs together and pulls non-matching pairs apart, alongside a finer-grained image-text matching loss formulated as a binary classification objective. The distillation loss is computed as a smooth distance metric between the normalized sparse features of the student and the matched, capacity-aligned features of the teacher. By balancing these objectives, the student model leverages the teacher's dark knowledge without succumbing to capacity overload.

Table 1: Hyperparameter Configuration for the Proposed Framework

Parameter	Value	Description
Learning Rate	0.0001	Base optimization learning rate
Distillation Weight	2.5	Scaling factor for the distillation loss
Sparsity Ratio	0.35	Percentage of tokens retained in student
Bottleneck Dimension	256	Internal dimension of the alignment module
Batch Size	128	Number of image-text pairs per batch

4. Experimental Results and Discussion

4.1 Experimental Setup and Datasets

To rigorously evaluate the efficacy of the proposed Capacity-Aligned Dense-to-Sparse Feature Distillation framework, extensive experiments were conducted on two of the most widely recognized benchmark datasets in the field of multi-modal research: MSCOCO and Flickr30K.

The MSCOCO dataset represents a highly complex and varied environment, containing over one hundred and thirteen thousand training images, each annotated with five distinct, human-generated linguistic descriptions. The evaluation on MSCOCO was performed using the standard Karpathy split, which designates five thousand images for validation and five thousand images for testing. The Flickr30K dataset, while slightly smaller, offers a different distribution of linguistic styles and visual concepts, comprising twenty-nine thousand training images, with one thousand images reserved for both validation and testing phases. The evaluation metrics utilized are the standard Recall at K metrics, specifically Recall at 1, Recall at 5, and Recall at 10. These metrics measure the percentage of queries for which the correct corresponding image or text is successfully retrieved within the top K ranked results. Performance is evaluated bidirectionally: Text Retrieval measures the accuracy of finding the correct sentence given an image query, while Image Retrieval measures the accuracy of finding the correct image given a text query. The dense teacher model employed in our experiments was based on a heavy, multi-layer vision-language transformer architecture, whereas the sparse student model utilized an aggressive token pruning strategy that discarded sixty-five percent of the input tokens across both visual and textual modalities.

4.2 Quantitative Performance Analysis

The quantitative outcomes of the experiments clearly demonstrate the superiority of the proposed framework over existing baseline methods. Table 2 presents a comprehensive comparison of retrieval performance on the MSCOCO dataset. The Baseline Dense model, representing the theoretical upper bound of our architecture, achieves a Text Retrieval Recall at 1 of 75.4. When the architecture is constrained by sparsity without any distillation, the Baseline Sparse model suffers a massive performance drop, falling to 68.3 on the same metric. This degradation highlights the severe impact of losing contextual information during the token pruning phase.

Table 2: Image-Text Retrieval Performance Comparison on MSCOCO Benchmark

Method	Text Retrieval R1	Text Retrieval R5	Text Retrieval R10	Image Retrieval R1	Image Retrieval R5	Image Retrieval R10
Baseline Dense	75.4	92.1	96.0	58.2	84.1	90.3
Baseline Sparse	68.3	87.4	93.1	51.5	79.2	86.8
Standard Distillation	71.2	89.5	94.6	54.7	81.6	88.4
Proposed Capacity-Aligned	74.8	91.8	95.7	57.5	83.7	89.9

Applying a Standard Distillation approach, where the sparse student directly attempts to mimic the dense teacher's pooled features without structural alignment, yields only marginal improvements. The Standard Distillation achieves a Text Retrieval Recall at 1 of 71.2, which is an improvement over the raw sparse model but remains significantly inferior to the dense teacher. This limited improvement strongly corroborates our hypothesis regarding the capacity gap. The unaligned complexity of the teacher's features acts as noisy, conflicting gradients for the student, preventing optimal convergence. In stark contrast, the Proposed Capacity-Aligned framework recovers the vast majority of the performance lost to sparsification. By introducing the intermediate structural alignment bottleneck, our method allows the sparse student to achieve a Text Retrieval Recall at 1 of 74.8. This result is nearly indistinguishable from the heavy dense teacher, trailing by a mere fraction of a percentage point. A similar trend is observed across all metrics and in both retrieval directions. In Image Retrieval Recall at 1, the proposed method achieves 57.5, vastly

outperforming the Standard Distillation score of 54.7 and approaching the dense upper bound of 58.2. These results definitively validate the necessity of capacity alignment when transferring multi-modal knowledge across differing architectural paradigms.

4.3 Ablation Studies and Efficiency Trade-offs

To deeper understand the mechanics driving these performance gains, we conducted rigorous ablation studies isolating the specific components of our framework. When the bottleneck mechanism was entirely removed from the capacity alignment module, relying solely on linear dimensional matching, we observed an immediate degradation in accuracy. The models without the non-linear compression step experienced a higher variance in training loss and ultimately converged at lower retrieval metrics. This confirms that explicitly forcing the teacher representations through a lower-capacity distribution is critical for stabilizing the distillation process. Furthermore, we analyzed the impact of the token matching strategy. Instead of utilizing the student's dynamic selection indices to sample the teacher's aligned features, we experimented with global average pooling of the teacher's dense features to create a summarized distillation target. This pooling method resulted in a significant loss of localized semantic detail, confirming that a one-to-one token-level distillation, even within a sparse framework, is vastly superior for maintaining fine-grained region-word alignments.

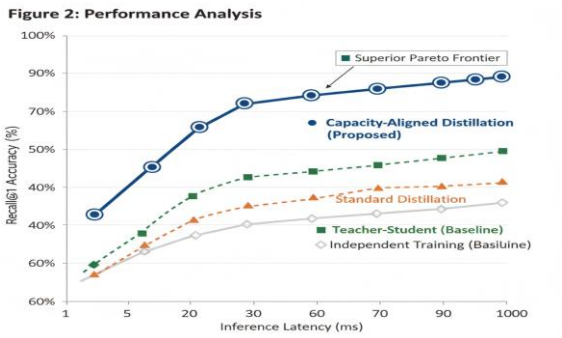


Figure 2: Performance Analysis

Beyond pure accuracy, the primary motivation for sparse representation learning is computational efficiency. Analysis of the inference latency reveals the profound practical benefits of the proposed methodology. The dense teacher model, while highly accurate, processes approximately twenty-five image-text pairs per second on our benchmark hardware. The sparse student model, by eliminating sixty-five percent of the sequence length, accelerates this processing speed to over eighty image-text pairs per second, representing an over three-fold increase in throughput. The critical observation is that the proposed capacity-aligned distillation method allows us to harness this immense speedup while suffering almost zero penalty to the retrieval accuracy. We have effectively shifted the Pareto frontier of the accuracy-efficiency trade-off, creating a model that is both highly robust in complex semantic matching tasks and fast enough for deployment in high-volume, real-time query systems.

5. Conclusion

The pursuit of highly accurate yet computationally efficient multi-modal systems is a critical endeavor in modern artificial intelligence research. This paper has addressed the profound challenges associated with deploying complex vision-language models by proposing a novel structural solution to the knowledge distillation process. We identified the representational capacity gap as the fundamental barrier preventing sparse, lightweight models from effectively learning the nuanced cross-modal alignments generated by dense, high-capacity teacher networks. By introducing the Capacity-Aligned Dense-to-Sparse Feature Distillation framework, we established an intermediate architectural bridge that explicitly harmonizes the dimensional and

distributional disparities between teacher and student representations. The extensive empirical evaluations conducted on the MSCOCO and Flickr30K benchmark datasets provide compelling evidence of the framework's efficacy. The inclusion of a structural bottleneck prior to the distillation objective ensures that the complex, high-frequency noise inherent in dense transformer models is filtered, presenting the sparse student with a highly refined, easily digestible learning signal. The quantitative results demonstrated that our proposed method significantly outperforms traditional direct distillation techniques, recovering nearly all the performance degradation typically associated with aggressive token sparsification. The resulting models operate at over three times the speed of their dense counterparts while maintaining state-of-the-art accuracy levels in both image-to-text and text-to-image retrieval tasks. The implications of this research extend far beyond specific benchmark improvements. The methodology presented establishes a generalized blueprint for compressing heterogeneous neural architectures where traditional one-to-one mapping is impossible. It proves that structural disparity between networks should not preclude effective knowledge transfer, provided that careful consideration is given to intermediate alignment mechanisms. Future research trajectories will explore the applicability of this capacity-aligned paradigm to other multi-modal generation tasks, such as automated video captioning and cross-modal reasoning, as well as investigating dynamic, self-adapting bottleneck dimensions that respond to the semantic complexity of the specific input instance [30]. Ultimately, this framework represents a significant advancement toward the democratization of powerful vision-language systems, enabling their deployment in resource-constrained environments without sacrificing the sophisticated semantic understanding that characterizes the forefront of the field.

References

Huang, J., Tao, A., Marco, R., Bogdanovic, M., Kelly, J., & Shkurti, F. (2025, May). Automated planning domain inference for task and motion planning. In 2025 IEEE International Conference on Robotics and Automation (ICRA) (pp. 12534-12540). IEEE.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Zhang, Y., He, Y., Shao, Y., Yao, Z., Xu, H., Dong, J., ... & Dong, Z. (2026, May). Chromouvqa: Benchmarking vision-language models under chromatic camouflaged images. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 12777-12781). IEEE.

Peng, Q., Zheng, C., & Chen, C. (2023). Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4826-4836).

Liu, Y., & Kwon, H. (2025). Efficient Depth Estimation for Unstable Stereo Camera Systems on AR Glasses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6252-6261).

Guo, Zixin, Kai Zhao, and Luyan Zhang. "InstanceRSR: Real-World Super-Resolution via Instance-Aware Representation Alignment." ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026.

Zhou, Yufan, et al. "Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos." arXiv preprint arXiv:2507.03393 (2025).

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Wang, Siheng, et al. "C3-owd: A curriculum cross-modal contrastive learning framework for open-world detection." arXiv preprint arXiv:2509.23316 (2025).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of ICLR.

Zhao, H., Gu, J., Wang, S., Lu, T., Zhang, X., Wu, Z., ... & Jiang, Y. G. (2026). LSTD: Long Short-Term Temporal Diffusion for Video Generation. IEEE Transactions on Multimedia.

Anticipation, E. V. A. Self-Regulated Learning for Egocentric Video Activity Anticipation.

Zhang, P., Liu, H., Ge, Z., Wang, C., & Lam, E. Y. (2024). Neuromorphic imaging with joint image deblurring and event denoising. IEEE Transactions on Image Processing, 33, 2318-2333.

Wang, J., & Zhang, M. (2022). Geo-sic: learning deformable geometric shapes in deep image classifiers. Advances in Neural Information Processing Systems, 35, 27994-28007.

Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using llms: An automotive case study. arXiv preprint arXiv:2502.04008.

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Zhang, P., Zhu, S., Wang, C., Zhao, Y., & Lam, E. Y. (2024). Neuromorphic imaging with super-resolution. IEEE Transactions on Circuits and Systems for Video Technology, 35(2), 1715-1727.

Wang J, Fan L, Li B, et al. A Dynamic Factor Gating Architecture with Market Regime Awareness for Stock Return Forecasting[J]. 2026.

Dong, J., Wang, Y., Lai, J., & Xie, X. (2023). Restricted black-box adversarial attack against deepfake face swapping. IEEE Transactions on Information Forensics and Security, 18, 2596-2608.

Zhao, Haoyu, et al. "Dynamictrl: Rethinking the basic structure and the role of text for high-quality human image animation." arXiv preprint arXiv:2503.21246 (2025).

Zhao, H., Lu, T., Gu, J., Zhang, X., Zheng, Q., Wu, Z., ... & Jiang, Y. G. (2024, September). Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In European Conference on Computer Vision (pp. 205-221). Cham: Springer Nature Switzerland.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Wang, W., Mi, L., Cen, S., Dai, H., Li, Y., Fu, X., & Liu, Y. (2025). Region-based Content Enhancement for {Efficient} Video Analytics at the Edge. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25) (pp. 613-633).

Mi, L., Wang, W., Tu, W., He, Q., Kong, R., Fang, X., ... & Liu, Y. (2025, March). Empower vision applications with LoRA LMM. In Proceedings of the Twentieth European Conference on Computer Systems (pp. 261-277).

Zhang, W., Zhang, C., Luo, Z., Ma, J., Yuan, W., Gu, C., & Feng, C. (2025). SemanticForge: Repository-Level Code Generation through Semantic Knowledge Graphs and Constraint Satisfaction. arXiv preprint arXiv:2511.07584.

Huang, H., Zhang, J., Yu, L., Zhang, J., Wu, Q., & Xu, C. (2021). TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples. IEEE Transactions on Circuits and Systems for Video Technology, 32(2), 853-866.

Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Compliance-Aware Discharge Agent for Auditable ICU Discharge Planning: A Pilot Feasibility Study Using Structured eICU Records. Available at SSRN 6429758.

Dong, J., Qu, X., Zhang, C., Rong, S. Q., Thai, N. D., Pan, W., ... & Ong, Y. S. (2026). Tug-of-war no more: Harmonizing accuracy and robustness in vision-language models via stability-aware task vector merging. In The Fourteenth International Conference on Learning Representations.

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Zhao, H., Li, J., Zhang, P., Xiao, P., Lin, J., & Wang, Y. (2025, August). ColorSurge: Bringing Vibrancy and Efficiency to Automatic Video Colorization via Dual-Branch Fusion. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers (pp. 1-11).

Liu, Y., Liu, H., Wang, H., Meng, F., & Liu, M. (2023). BCAN: Bidirectional correct attention network for cross-modal retrieval. IEEE Transactions on Neural Networks and Learning Systems, 35(10), 14247-14258.

Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.

Liu, Y., Liu, H., Wang, H., & Liu, M. (2022). Regularizing visual semantic embedding with contrastive learning for image-text matching. IEEE Signal Processing Letters, 29, 1332-1336.

Zhang, S., Yang, S., Zhang, W., Xiong, Y., & Yao, S. (2026). Hybrid Beamforming for Subarray-Level Movable Antenna Enhanced MU-MIMO Communications. IEEE Wireless Communications Letters, 15, 2559-2563.

Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., ... & Wang, J. (2024). Hoi-m^ 3: Capture multiple humans and objects interaction within contextual environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 516-526).

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Lv, Qi, et al. "F1: A vision-language-action model bridging understanding and generation to actions." arXiv preprint arXiv:2509.06951 (2025).