RETHINKING TEXT SEMANTICS IN SUPERVISED UNPAIRED MULTIMODAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised *Unpaired Multimodal Representation Learning* (UMRL) improves visual classification using text representations, requiring neither *prior cross-modal alignment* nor *sample-level correspondence*. Under such conditions, what drives the performance gains remains unclear, particularly the role of text semantics. In this paper, we revisit supervised UMRL and identify learnable visual projector and text-derived classifier initialization as the main contributors among the examined factors, while increasing the number of text representations and introducing a text classification loss provide only limited benefits. Interestingly, shuffling the correspondence between text representations and class labels has little effect on the performance. Motivated by this, we introduce a semantics-agnostic representation synthesis strategy for classifier initialization. Extensive experiments show that semantics-agnostic classifier initialization performs comparably to its text-derived counterpart. It also eliminates the need for an additional text encoder, reducing text encoding time and model storage overhead. Our observations suggest that text semantics may not be the primary driver of the performance gains in supervised UMRL. The codes are available at `https://anonymous.4open.science/r/RethinkUMRL`.

1 INTRODUCTION

Multimodal representation learning (Radford et al., 2021; Dufumier et al., 2025; Cicchetti et al., 2025; Yin et al., 2026) typically leverages large-scale paired multimodal samples to learn aligned representations across modalities. Such cross-modal alignment facilitates effective knowledge transfer and has achieved remarkable success in various downstream tasks, including visual classification (Roth et al., 2025; Zhang et al., 2025; Guo & Gu, 2025) and vision-language understanding (Deitke et al., 2025; Tong et al., 2024; Dai et al., 2023). Without such alignment, independently pretrained unimodal encoders learn modality-specific representations, making direct cross-modal knowledge transfer challenging. Recent work (Gupta et al., 2026) formalizes Unpaired Multimodal Representation Learning (UMRL), demonstrating that independently pretrained unimodal encoders can improve performance on the target modality without prior cross-modal alignment and sample-level correspondence (see Figure 1). This motivates further investigation into UMRL.

In this work, we focus on supervised UMRL, where auxiliary text representations have been shown to improve the performance of visual classification tasks. Gupta et al. (2026) suggest that these improvements may stem from the semantic information contained in text representations. However, the extent to which these gains are driven by semantic information remains unclear, as text representations are obtained from independently pretrained encoders without prior cross-modal alignment. This raises a question:

What factors mainly contribute to the performance gains in supervised UMRL?

To answer this question, we revisit supervised UMRL and systematically analyze four factors behind its performance gains: (1) the number of text representations per class, (2) the text classification loss, (3) the learnable visual projector g_v, and (4) text-derived classifier initialization W_t. Our analysis finds that the observed gains are mainly associated with g_v and W_t. The former enables effective adaptation of visual representations, while the latter provides a favorable initialization for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

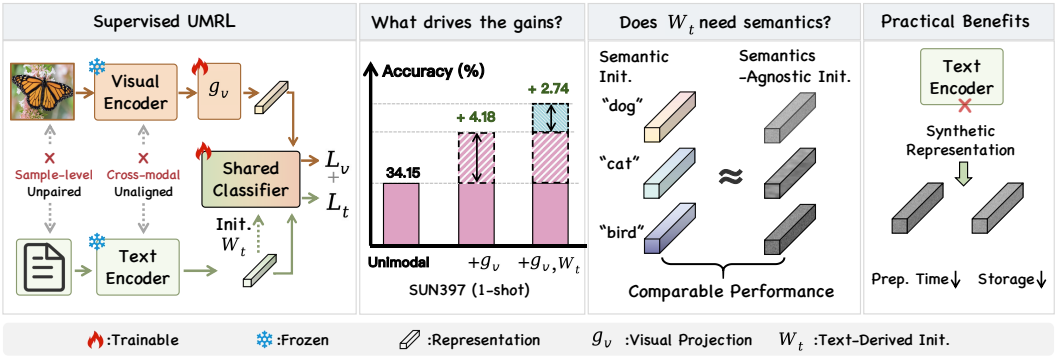


Figure 1: Overview of our analysis of supervised UMRL. Among the factors examined, the performance gains are mainly associated with learnable visual projector and text-derived classifier initialization. For the latter, semantics-agnostic representations achieve performance comparable to text representations while eliminating the need for an additional text encoder, thus reducing text encoding time and model storage overhead.

the classifier. We further investigate whether the performance gains of W_t rely on text semantics. By replacing text representations with semantics-agnostic alternatives to initialize classifier weights, we achieve comparable performance, suggesting that the performance gains are not mainly driven by text semantics. Furthermore, using semantics-agnostic representations eliminates the time overhead of extracting text representations and the storage overhead of maintaining a text encoder. Figure 1 provides an overview of our analysis and main observations.

Taken together, those observations provide a different perspective on supervised UMRL. When auxiliary representations are obtained from independently pretrained encoders without prior cross-modal alignment, their benefits may not primarily arise from semantic transfer. We summarize our main contributions as follows.

- We conduct a systematic analysis of supervised UMRL and identify visual adaptation and classifier initialization as the main contributors to the performance gains.
- We show that the benefits of text-derived classifier initialization do not appear to primarily depend on text semantics.
- We introduce a semantics-agnostic representation synthesis strategy that achieves comparable performance without a text encoder, reducing computation and storage overhead.

2 Related Work

In this section, we review several lines of research closely related to our work, including UMRL, text-derived classifier initialization, and semantics-agnostic representation learning.

UMRL. Gupta et al. (2026) formalize UMRL, which studies whether an auxiliary modality can benefit a target modality without paired samples, using independently pretrained and unaligned encoders. They study UMRL under both self-supervised and supervised settings. We focus on the latter, where unpaired text is incorporated to improve visual classification over image-only unimodal baselines. *While this work demonstrates the potential of unpaired auxiliary modalities, the source of the gains in supervised UMRL remains unresolved. We address this by systematically disentangling the supervised pipeline to determine which components account for the observed improvement.*

Text-Derived Classifier Initialization. Text representations have been widely used to provide class-level priors for visual classification. CLIP (Radford et al., 2021) constructs class-specific text representations from text prompts and directly uses them to define a zero-shot classifier. Lin et al. (2023) use similar class-specific text representations to initialize a linear classifier. LP++ (Huang et al., 2024) further uses such text representations to guide classifier learning together with class-specific visual prototypes. In supervised UMRL, Gupta et al. (2026) initialize the classifier using class-specific text prototypes, computed by averaging the text representations within each class. *In*

this work, we analyze whether the benefit of text-derived classifier initialization in supervised UMRL depends on semantic content.

Semantics-Agnostic Representation Learning. Related studies show that structured but semantics-agnostic representations can provide useful learning signals. Hyperspherical Prototype Networks (Mettes et al., 2019) construct data-independent class prototypes with large margin separation. A similar geometric structure emerges in neural collapse (Papyan et al., 2020; Zhu et al., 2021), where class means and classifier weights converge to a simplex equiangular tight frame (ETF). Building on this geometry, Yang et al. (2022) fix the classifier to a predefined simplex ETF. Additionally, Yao et al. (2025; 2026) show that synthetic noise with suitable discriminative structure can yield positive transfer. *Unlike those studies, we use semantics-agnostic representations to isolate the role of semantic content in text-derived classifier initialization in supervised UMRL.*

3 SUPERVISED UMRL

In this section, we introduce the supervised UMRL setting and its key components. Let $\mathcal{D}_v = \{(\mathbf{x}_i^v, y_i^v)\}_{i=1}^{n_v}$ and $\mathcal{D}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$ denote the visual and textual datasets, respectively. The two modalities share the same C-class label space but *have no sample-level correspondences*, and their encoders are independently pretrained *without prior cross-modal alignment*. Supervised UMRL jointly optimizes the visual and textual classification losses:

$$\mathcal{L}_{v+t} = \mathcal{L}_v + \mathcal{L}_t. \tag{1}$$

To bridge heterogeneous representation spaces, it utilizes a learnable visual projector g_v to map visual representations into the textual representation space. It further initializes the classifier using W_t, constructed from class prototypes of text representations. This design introduces several components that may contribute to the observed gains: (1) the number of text representations per class; (2) the text classification loss $\mathcal{L}_t$; (3) the learnable visual projector g_v; and (4) the text-derived classifier initialization W_t. Next, we analyze the contribution of each component.

4 WHAT DRIVES THE PERFORMANCE GAINS IN SUPERVISED UMRL?

In this section, we first present the experimental setup, then analyze the sources of the gains, and finally summarize the key observations.

4.1 EXPERIMENTAL SETUP

Datasets and Backbones. We evaluate on eight widely used classification datasets: DTD (Cimpoi et al., 2014), UCF-101 (Soomro et al., 2012), Stanford Cars (Krause et al., 2013), SUN397 (Xiao et al., 2010), Caltech-101 (Fei-Fei et al., 2004), FGVC-Aircraft (Maji et al., 2013), Oxford Pets (Parkhi et al., 2012), and Oxford 102 Flowers (Nilsback & Zisserman, 2008). We additionally use ImageNet-1K (Deng et al., 2009) for large-scale analysis. Following Gupta et al. (2026), we use DINOv2 ViT-S/14 (Oquab et al., 2024) as the visual encoder and OpenLLaMA-3B-v2 (Geng & Liu, 2023) as the text encoder unless otherwise specified.

Training and Evaluation. We focus on the 1-, 2-, and 4-shot settings, corresponding to 1, 2, and 4 labeled images per class, respectively. Both visual and text encoders are kept frozen throughout those experiments. Hyperparameters are selected via grid search on the validation set. We report the mean classification accuracy and standard deviation over three random seeds. Further setup details, including dataset statistics and hyperparameter search spaces, are provided in Appendix A.

4.2 ANALYZING THE SOURCES OF THE GAINS

Q1. How many text representations are needed per class? Figure 2 examines the effect of varying the number of text representations per class from 1 to 30. Across datasets and few-shot settings, performance remains largely stable, with no consistent improvement as more text representations are introduced. Notably, a single representation per class already performs comparably to using 30. A possible explanation is that the visual and text encoders are independently pretrained without prior cross-modal alignment, so additional textual representations may not consistently translate into

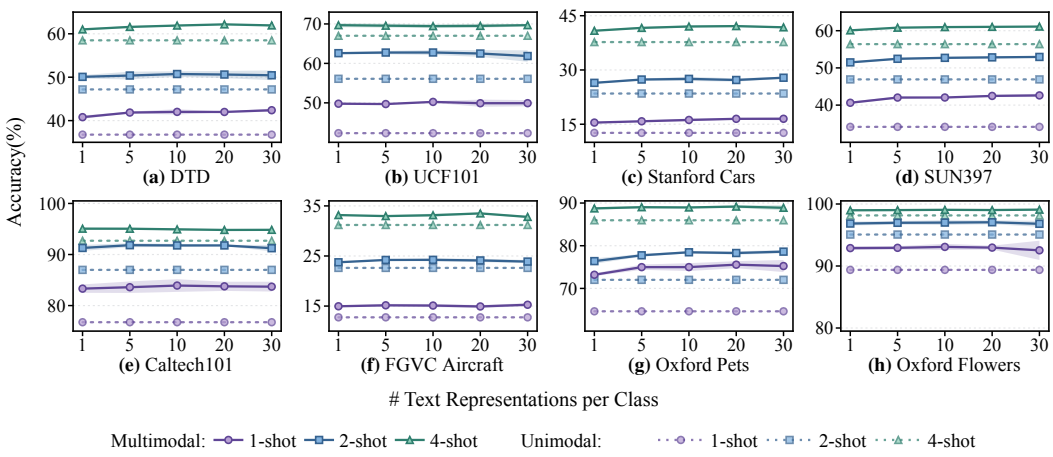


Figure 2: Effect of the number of text representations per class in supervised UMRL.

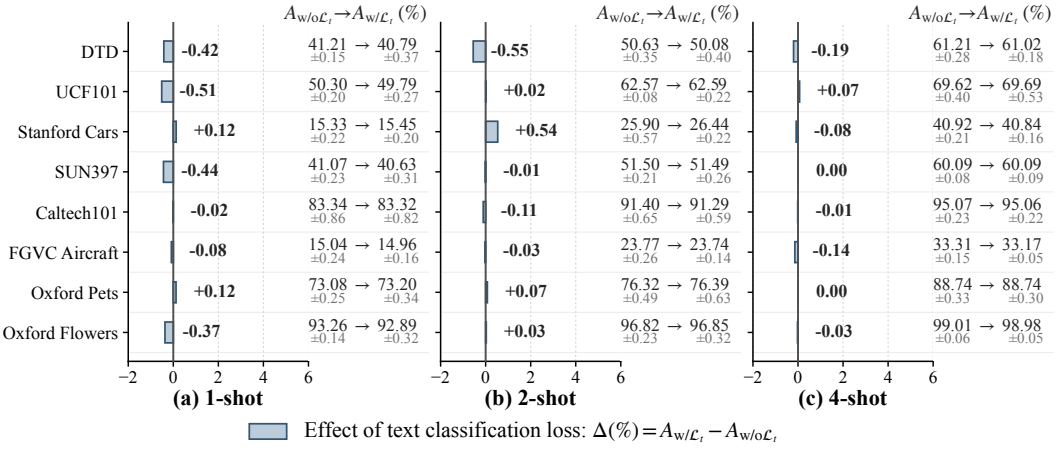


Figure 3: Effect of text classification loss in supervised UMRL. Here, $A_{\mathrm{w}/\mathcal{L}_t}$ and $A_{\mathrm{w/o}\mathcal{L}_t}$ denote accuracies with and without the text classification loss $\mathcal{L}_t$, respectively.

better visual performance. These results suggest that the number of text representations is unlikely to be a key factor behind the gains in supervised UMRL. ***We thus use one text representation per class in the following analyses, which is used for both classifier initialization and the text classification loss $\mathcal{L}_t$.***

Q2. Is the text classification loss necessary? Following the above analysis, we further examine whether explicit text supervision is necessary. Figure 3 compares performance with and without the text classification loss $\mathcal{L}_t$. As can be seen, removing $\mathcal{L}_t$ leads to only minor performance changes, with no consistent improvement or degradation. A possible explanation is that the two encoders are not cross-modally aligned, so training the classifier on text representations may not lead to substantial gains in visual classification. These results suggest that the text classification objective is unlikely to be a key factor behind the observed gains. ***We thus remove $\mathcal{L}_t$ in the following analyses, such that the text representation is used only for classifier initialization.***

Q3. Does learnable visual projector contribute to the gains? We investigate the role of the learnable visual projector g_v. To isolate its effect, the two settings differ only in whether g_v is used. Without g_v, training updates only the classifier over frozen visual representations; with g_v, the visual representations are additionally adapted during training. As shown in Figure 4, introducing g_v improves performance in most evaluated settings. Those results suggest that learnable visual projector provides an important source of performance gain beyond classifier optimization alone.

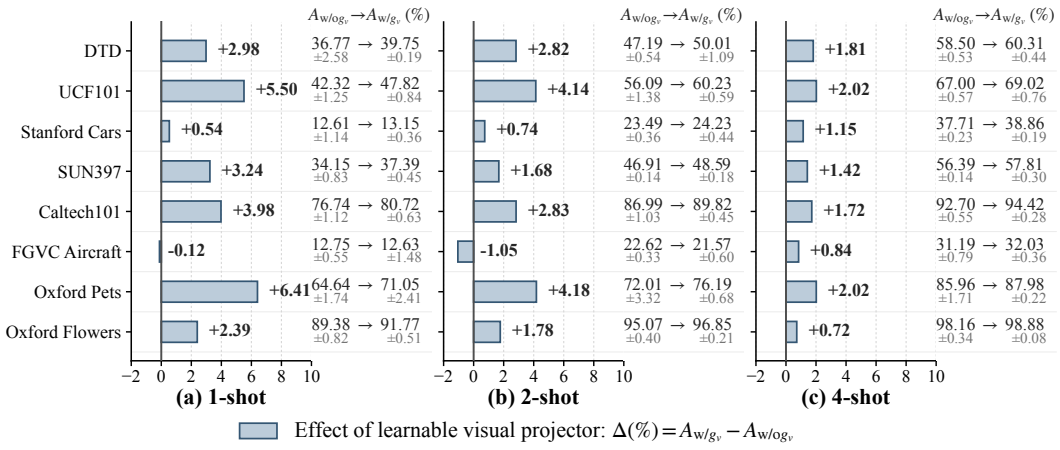


Figure 4: Effect of learnable visual projector in supervised UMRL. Here, A_{w/g_v} and $A_{\mathrm{w/o}g_v}$ denote accuracies with and without the learnable visual projector, respectively.

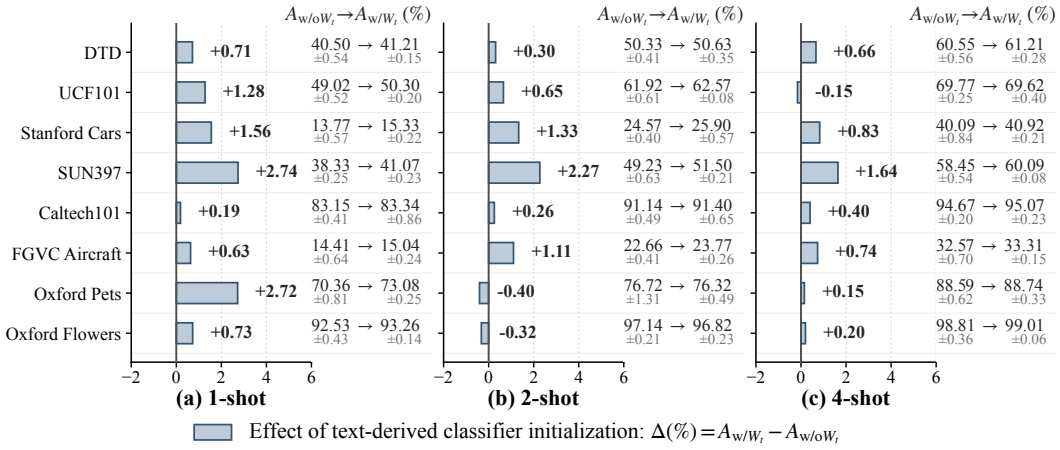


Figure 5: Effect of text-derived classifier initialization in supervised UMRL. Here, A_{w/W_t} and $A_{\mathrm{w/o}W_t}$ denote accuracies with and without text-derived classifier initialization, respectively.

Q4. Does text-derived classifier initialization outperform random initialization? We retain the learnable visual projector g_v and further isolate the effect of text-derived classifier initialization W_t. Specifically, we compare W_t with random classifier initialization while keeping all other components unchanged. As shown in Figure 5, text-derived initialization improves performance on most datasets, with more pronounced gains under the 1-shot setting. This trend is consistent with the findings of Gupta et al. (2026). These results suggest that, beyond visual adaptation, text-derived classifier initialization provides an additional source of performance gain. Since W_t is derived from text representations, we next investigate whether this benefit depends on the semantic information encoded in those representations.

Q5. Does text-derived classifier initialization require correct text–class correspondence? We now examine whether the benefit of text-derived classifier initialization depends on the semantic correspondence between text representations and class labels. Specifically, we randomly shuffle the correspondence between text representations and class labels and use the shuffled text representations to initialize the classifier. We repeat this procedure using nine different random shuffles. As shown in Figure 6, performance remains largely stable across these random assignments and stays close to that obtained with the correct assignment. Those results suggest that the effectiveness of text-derived classifier initialization may not primarily require the correct semantic correspondence between text representations and class labels. This motivates us to further examine whether similar gains can be reproduced using fully semantics-agnostic representations.

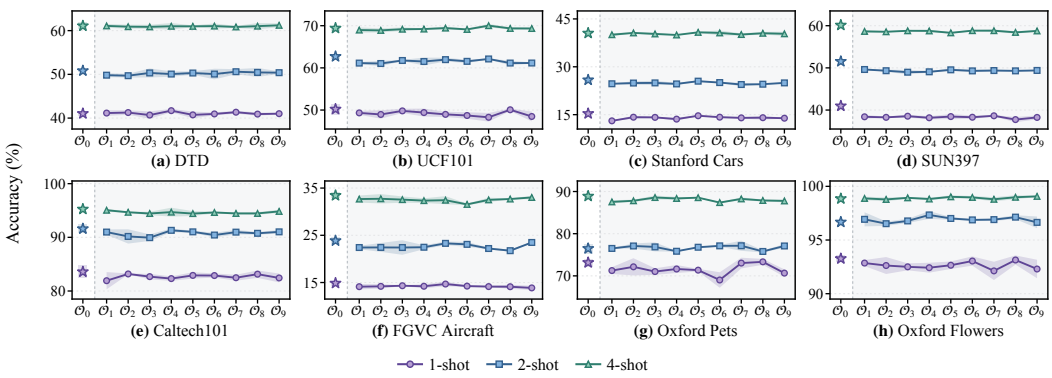


Figure 6: Effect of changing the class assignment of text representations in supervised UMRL. Here, $\mathcal{O}_0$ denotes the correct assignment, while $\mathcal{O}_1$–$\mathcal{O}_9$ denote nine random reassignment schemes, where the text representation of each class is reassigned to another class.

4.3 Summary

In summary, those results suggest the following observations.

Observation 1: Increasing the number of text representations or using explicit text supervision provides only limited additional gains.

Observation 2: learnable visual projector and text-derived classifier initialization are the main factors associated with the observed performance gains.

Observation 3: The benefits of text-derived classifier initialization do not appear to rely primarily on the correct correspondence between text representations and class labels.

5 Can the Gains Be Reproduced without Text Semantics?

The above results suggest that the correct text-class correspondence plays only a limited role in the effectiveness of classifier initialization. However, the representations used for initialization are still produced by a text encoder and may retain semantic information. We therefore ask whether the observed gains can be reproduced without relying on text semantics. To this end, we introduce a semantics-agnostic representation synthesis strategy, evaluate whether it can reproduce the benefits of text-derived initialization, and finally summarize the key observations.

5.1 Semantics-Agnostic Representation Synthesis

A straightforward way to remove text semantics is to replace text representations with randomly sampled representations. As shown in **Q10**, direct random sampling may not consistently match the performance of text-derived representations, especially in low-shot settings. We thus consider a simple synthesis strategy with controllable expected inter-class cosine similarity, denoted by $m \in [0, 1]$. Specifically, for each class c, we construct the representation as

$$\mathbf{v}_c = \sqrt{m}\mathbf{q} + \sqrt{1-m}\mathbf{r}_c, \tag{2}$$

where $\mathbf{q}$ is a shared unit vector and $\mathbf{r}_c$ is a class-specific unit vector independently and uniformly sampled from the unit sphere in the subspace orthogonal to $\mathbf{q}$. The resulting representations satisfy $\|\mathbf{v}_c\|_2 = 1$ and for $i \neq j$, $\mathbb{E}[\cos(\mathbf{v}_i, \mathbf{v}_j)] = m$.

5.2 Empirical Evaluation

For simplicity, in the following evaluation, we refer to classifier initialization using the original text representations as *Semantic Init.*, and initialization using the synthesized semantics-agnostic representations as *Semantics-Agnostic Init.*.

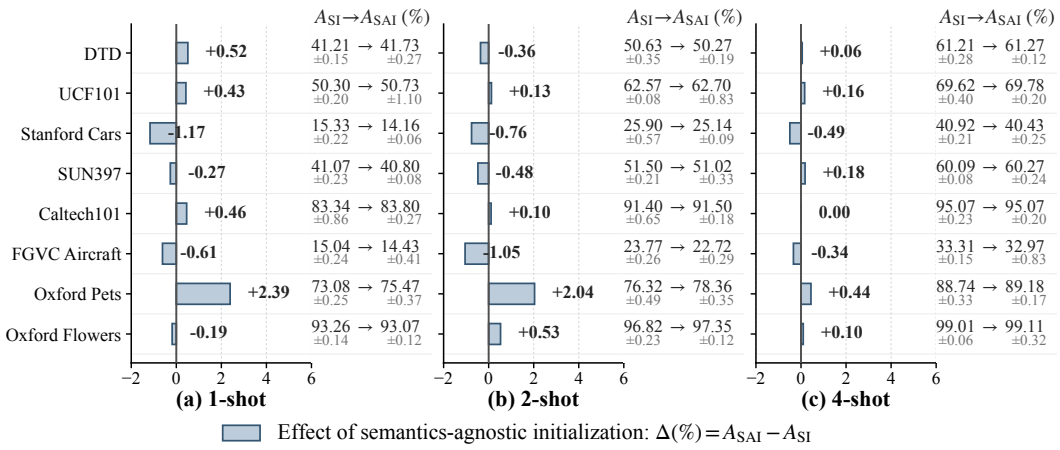


Figure 7: Effect of semantics-agnostic initialization in supervised UMRL. Here, A_{SI} and A_{SAI} denote accuracies with Semantic Init. and Semantics-Agnostic Init., respectively.

Q6. Can semantics-agnostic representations be as effective as text representations? Figure 7 compares *Semantics-Agnostic Init.* with *Semantic Init.* across different datasets and shot settings. Overall, replacing text representations with semantics-agnostic representations does not lead to a systematic performance drop. The performance differences are generally small, with *Semantics-Agnostic Init.* performing slightly better on some datasets and slightly worse on others. In several settings, *Semantics-Agnostic Init.* even outperforms *Semantic Init.*, with particularly clear gains on Oxford Pets under the 1-shot and 2-shot settings. Moreover, the performance gap becomes smaller under the 4-shot setting, where the two initialization schemes achieve very similar results across all evaluated datasets. Those results suggest that the benefits of *Semantic Init.* may not rely on text semantics and can be largely reproduced by semantics-agnostic representations.

Q7. Do semantics-agnostic representations remain effective on large-scale ImageNet-1K? We conduct experiments on ImageNet-1K (Deng et al., 2009) with different numbers of labeled images per class. Specifically, we compare *Semantics-Agnostic Init.* with *Semantic Init.*, together with the original unimodal baseline. Figure 8 shows that *Semantics-Agnostic Init.* achieves performance comparable to *Semantic Init.* across all evaluated settings. Moreover, both variants consistently outperform the original unimodal baseline across different numbers of labeled images per class. Those results show that the effectiveness of *Semantics-Agnostic Init.* extends to ImageNet-1K.

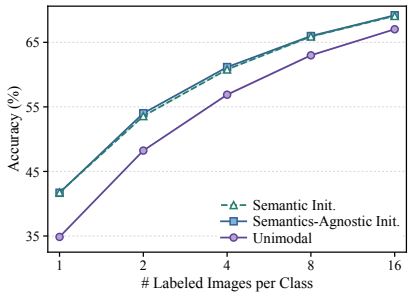


Figure 8: Effect of semantics-agnostic representations on ImageNet-1K.

Q8. Does *Semantics-Agnostic Init.* remain effective across DINO/DINOv2 backbones? As shown in Figure 9, on DINOv2 ViT-B/14 (Oquab et al., 2024), *Semantics-Agnostic Init.* remains comparable to *Semantic Init.* across most settings. Similar trends are observed on DINOv2 ViT-L/14 (Oquab et al., 2024), DINO ViT-B/16 (Caron et al., 2021), and DINO ViT-B/8 (Caron et al., 2021), with additional results provided in Appendix B. Those results suggest that *Semantics-Agnostic Init.* remains effective across the evaluated DINO/DINOv2 backbones.

Q9. Does the observation hold under full fine-tuning settings? We further evaluate *Semantics-Agnostic Init.* under full fine-tuning, where the visual backbone and classification head are jointly optimized. As shown in Table 1, the gains over the unimodal baseline become relatively modest on many datasets. Meanwhile, *Semantics-Agnostic Init.* remains comparable to *Semantic Init.*. Those results suggest that our main observation continues to hold under full fine-tuning settings.

Q10. Do distinct synthesis strategies affect *Semantics-Agnostic Init.*? We compare our synthesis strategy with Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, Uniform $\mathcal{U}(-\sqrt{3}, \sqrt{3})$, and Laplace $\mathcal{L}(0, 1/\sqrt{2})$ sampling. As shown in Table 2, different synthesis strategies can lead to performance gaps of several percentage

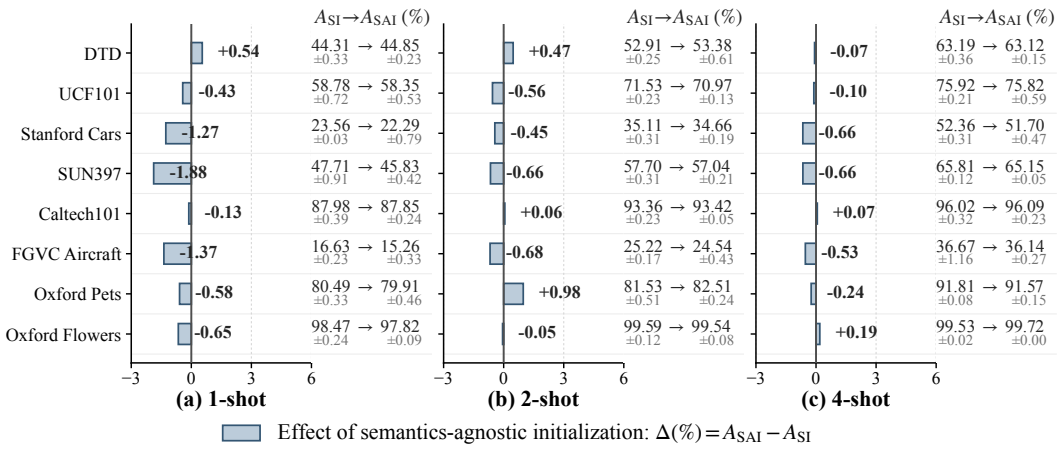


Figure 9: Effect of semantics-agnostic initialization in supervised UMRL using DINOv2 ViT-B/14. Here, A_{SI} and A_{SAI} denote accuracies with Semantic Init. and Semantics-Agnostic Init., respectively.

Table 1: Full fine-tuning evaluation of *Semantic Init.* and *Semantics-Agnostic Init.*.

	Accuracy (%)			Δ (%)	
Dataset	Unimodal	Semantic Init.	Semantics-Agnostic Init.	vs. Unimodal	vs. Semantic Init.
DTD	71.77 ± 3.57	71.26 ± 1.36	72.61 ± 0.57	$+0.84$	$+1.35$
UCF101	82.76 ± 0.58	85.06 ± 1.91	84.75 ± 0.50	$+1.99$	-0.31
Cars	79.48 ± 1.36	84.41 ± 0.78	85.60 ± 0.04	$+6.12$	$+1.19$
SUN397	65.81 ± 0.24	67.84 ± 1.59	67.25 ± 0.67	$+1.44$	-0.59
Caltech101	96.96 ± 0.42	96.94 ± 0.47	97.05 ± 0.18	$+0.09$	$+0.11$
Aircraft	67.36 ± 1.13	71.45 ± 0.19	73.78 ± 0.16	$+6.42$	$+2.33$
Pets	91.75 ± 0.68	92.13 ± 0.59	92.22 ± 0.21	$+0.47$	$+0.09$
Flowers	99.23 ± 0.11	99.43 ± 0.07	99.54 ± 0.08	$+0.31$	$+0.11$

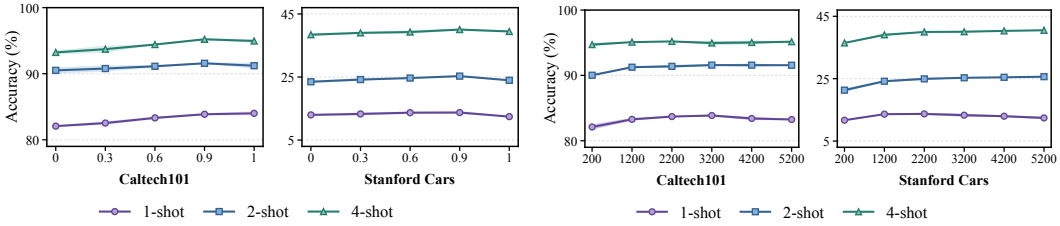
points in low-shot settings. Despite these differences, Gaussian, Uniform, and Laplace sampling exhibit broadly similar trends, and the gaps across synthesis strategies become smaller as more labeled images are available.

Table 2: Comparison of different semantics-agnostic construction strategies and *Semantic Init.*.

Dataset	**Method**	**1-shot**	**2-shot**	**4-shot**
DTD	Gaussian Sampling	37.50 ± 0.85	48.99 ± 1.05	60.03 ± 0.40
	Uniform Sampling	38.73 ± 1.20	49.19 ± 0.93	59.69 ± 0.36
	Laplace Sampling	39.05 ± 0.81	49.08 ± 0.65	60.15 ± 0.19
	Semantic Init.	41.21 ± 0.15	50.63 ± 0.35	61.21 ± 0.28
	Semantics-Agnostic Init.	41.73 ± 0.27	50.27 ± 0.19	61.27 ± 0.12
SUN397	Gaussian Sampling	36.09 ± 0.36	47.30 ± 0.18	57.01 ± 0.09
	Uniform Sampling	35.04 ± 0.57	47.21 ± 0.49	56.92 ± 0.11
	Laplace Sampling	36.05 ± 0.37	47.35 ± 0.27	57.00 ± 0.18
	Semantic Init.	41.07 ± 0.23	51.50 ± 0.21	60.09 ± 0.08
	Semantics-Agnostic Init.	40.80 ± 0.08	51.02 ± 0.33	60.27 ± 0.24

Q11. How sensitive is *Semantics-Agnostic Init.* to inter-class cosine similarity m? We evaluate $m \in \{0, 0.3, 0.6, 0.9, 1\}$, which controls the expected inter-class cosine similarity among synthesized class representations at initialization. As shown in Figure 10(a), performance remains relatively stable across the tested values on Caltech101 and Stanford Cars. Overall, those results suggest limited sensitivity of the performance to the tested values of m under the evaluated settings.

Q12. Does representation dimensionality influence *Semantics-Agnostic Init.*? We evaluate synthesized representations with dimensions ranging from 200 to 5200. As shown in Figure 10(b), performance remains largely stable on both Caltech101 and Stanford Cars. Those results indicate



(a) Inter-class cosine similarity (b) Representation dimensionality

Figure 10: Sensitivity to (a) inter-class cosine similarity and (b) representation dimensionality.

that *Semantics-Agnostic Init.* is not sensitive to representation dimensionality across a wide range on evaluated settings.

Q13. How much time and storage overhead does *Semantics-Agnostic Init.* save over *Semantic Init.*? The original supervised UMRL pipeline relies on the OpenLLaMA-3B-v2 text encoder, requiring approximately 6.85 GB of storage for its FP16 weights. *Semantics-Agnostic Init.* eliminates this dependency and reduces text encoding time, as shown in Table 3. Overall, it simplifies the supervised UMRL pipeline by reducing both computation time and model storage overhead.

Table 3: Time overhead (in milliseconds) of text extraction and semantics-agnostic synthesis.

Runtime Component	UCF101	Oxford Pets	DTD	SUN397
Text Representation Extraction	2,215	1,031	878	7,736
Semantics-Agnostic Representation Synthesis	194	185	170	247

5.3 Summary

In summary, those results suggest the following observations.

Observation 4: *Semantics-Agnostic Init.* achieves performance comparable to *Semantic Init.* across datasets, shot settings, visual backbones, and full fine-tuning settings.

Observation 5: The effectiveness of *Semantic Init.* does not appear to depend strongly on the semantic information encoded in text representations, as comparable performance can be achieved using semantics-agnostic representations.

Observation 6: Different synthesis strategies may affect performance to some extent in low-shot settings, while their influence generally decreases as more labeled images become available.

Observation 7: *Semantics-Agnostic Init.* is relatively insensitive to inter-class cosine similarity and representation dimensionality, especially with more labeled images.

Observation 8: By removing the text encoder, *Semantics-Agnostic Init.* simplifies the supervised UMRL pipeline and reduces the associated time and storage overhead.

6 Conclusion

In this paper, we investigate what drives the performance gains from text representations in supervised UMRL. Among the factors examined, we find that these gains are mainly associated with the learnable visual projector and text-derived classifier initialization. Meanwhile, text semantics appear to play only a limited role in the effectiveness of this initialization. We further introduce a semantics-agnostic representation synthesis strategy that achieves performance comparable to text-derived classifier initialization. This strategy eliminates the additional text encoder, reducing both text encoding time and model storage overhead. Our observations suggest that, without prior cross-modal alignment, the benefits of text representations may not primarily arise from semantic transfer. In the future, we plan to explore how to bridge independently pretrained modality encoders to utilize semantic information more effectively.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, manuscript organization, and the refinement of experimental framing and interpretation. All AI-assisted content was carefully reviewed and revised by the authors. The authors take full responsibility for the final content, analyses, and conclusions of this work.

ETHICS STATEMENT

This work focuses on the methodological analysis of multimodal representation learning and does not involve human subjects or the collection of personal or sensitive data. Our experiments are conducted using existing benchmark datasets and pretrained models. We are not aware of any specific ethical concerns arising from this work beyond those generally associated with the datasets and pretrained models used.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. To this end, we make our source code available at `https://anonymous.4open.science/r/RethinkUMRL`. The experimental setup, including dataset splits, model architectures, and training hyperparameters, is described in Section 4.1 and Appendix A. With the released code and documentation, all reported results can be readily reproduced by the research community.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, pp. 446–461, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *ICLR*, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven C. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, volume 36, pp. 49250–49267, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pp. 91–104, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *ICLR*, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *CVPRW*, 2004.

Xinyang Geng and Hao Liu. OpenLLaMA: An open reproduction of LLaMA. GitHub repository, May 2023.

Yuncheng Guo and Xiaodong Gu. MMRL: Multi-modal representation learning for vision-language models. In *CVPR*, pp. 25015–25025, 2025.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. In *ICLR*, 2026.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *CVPR*, pp. 23773–23782, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCVW*, pp. 554–561, 2013.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *CVPR*, pp. 19325–19337, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Subhransu Maji, Juho Kannala, Esa Rahtu, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Pascal Mettes, Elise van der Pol, and Cees G. M. Snoek. Hyperspherical prototype networks. In *NeurIPS*, volume 32, 2019.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pp. 722–729, 2008.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *PNAS*, 117(40):24652–24663, 2020.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pp. 3498–3505, 2012.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pp. 15691–15701, 2023.

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 *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Karsten Roth, Zeynep Akata, Dima Damen, Ivana Balažević, and Olivier J Hénaff. Context-aware multimodal pretraining. In *CVPR*, pp. 4267–4279, 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision, University of Central Florida, November 2012.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *NeurIPS*, volume 37, pp. 87310–87356, 2024.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pp. 3485–3492, 2010.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *NeurIPS*, volume 35, pp. 37991–38002, 2022.

Yuan Yao, Xiaopu Zhang, Yu Zhang, Jian Jin, and Qiang Yang. Noise may contain transferable knowledge: Understanding semi-supervised heterogeneous domain adaptation from an empirical perspective. *arXiv preprint arXiv:2502.13573*, 2025.

Yuan Yao, Jin Song, Huixia Li, Tongtong Yuan, Jiaqi Wu, and Yu Zhang. Semi-supervised noise adaptation: Transferring knowledge from noise domain. In *ICML*, 2026.

Wenzhe Yin, Pan Zhou, Zehao Xiao, Jie Liu, Shujian Yu, Jan-Jakob Sonke, and Efstratios Gavves. Towards uniformity and alignment for multimodal representation learning. In *ICML*, 2026.

Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models. In *CVPR*, pp. 14604–14614, 2025.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *NeurIPS*, volume 34, pp. 29820–29834, 2021.

A Additional Experimental Settings

A.1 Dataset Details

We evaluate on the following datasets, generally following the splits of Gupta et al. (2026). We further clean the data to reduce potential split overlap. The dataset statistics are summarized below:

- **DTD-47** (Cimpoi et al., 2014): 5,623 texture images from 47 classes, split into 2,807 training, 1,127 validation, and 1,689 test images.
- **Food-101** (Bossard et al., 2014): 100,940 images from 101 food categories, split into 50,461 training, 20,187 validation, and 30,292 test images.
- **UCF101** (Soomro et al., 2012): 13,300 action images from 101 classes, split into 7,626 training, 1,897 validation, and 3,777 test images.
- **Stanford Cars-196** (Krause et al., 2013): 16,146 images from 196 fine-grained car classes, split into 6,482 training, 1,631 validation, and 8,033 test images.
- **SUN397** (Xiao et al., 2010): 39,700 images from 397 scene categories in our benchmark split, with 15,880 training, 3,970 validation, and 19,850 test images.
- **Caltech-101** (Fei-Fei et al., 2004): 8,242 images from 100 object classes retained in our benchmark split, with 4,128 training, 1,649 validation, and 2,465 test images.
- **FGVC Aircraft** (Maji et al., 2013): 10,000 aircraft images from 100 fine-grained aircraft classes, split into 3,334 training, 3,333 validation, and 3,333 test images.
- **Oxford Pets** (Parkhi et al., 2012): 7,349 images from 37 cat and dog breeds, split into 2,944 training, 736 validation, and 3,669 test images.
- **Oxford Flowers-102** (Nilsback & Zisserman, 2008): 8,185 images from 102 flower categories in our processed benchmark split, with 4,090 training, 1,632 validation, and 2,463 test images.
- **ImageNet-1K** (Deng et al., 2009): contains 1,000 object classes. We follow the data split adopted by Gupta et al. (2026), with 1,024,936 training images, 256,231 validation images, and 50,000 test images.

A.2 Implementation Details

We build our implementation on the official Unpaired Multimodal Representation Learning (UMRL) codebase released by Gupta et al. (2026) and conduct all experiments on NVIDIA A800 GPUs.

Vision and text encoders. We use DINOv2 ViT-S/14 as the default vision backbone and further evaluate other DINOv2 (Oquab et al., 2024) and DINO (Caron et al., 2021) encoders. Experiments with semantic text representations use a frozen OpenLLaMA-3B-v2 (Geng & Liu, 2023) text encoder. The vision and text encoders are independently pretrained without prior cross-modal alignment. Text descriptions are generated using the CuPL prompting strategy (Pratt et al., 2023).

Few-shot linear probing. We follow the training protocol of Gupta et al. (2026), freezing the pretrained vision backbone and optimizing only the downstream classification module. All methods use cross-entropy loss and AdamW (Loshchilov & Hutter, 2019). Hyperparameters are selected on the validation set, with the complete configuration summarized in Table 4.

Full fine-tuning. For full fine-tuning, we jointly optimize the vision backbone and downstream classification module using DINOv2 ViT-S/14 (Oquab et al., 2024). Hyperparameters are selected based on validation performance, with the complete configuration provided in Table 4.

B Additional Results on Visual Backbones

We provide additional results on DINOv2 ViT-L/14 (Oquab et al., 2024), DINO ViT-B/16, and DINO ViT-B/8 (Caron et al., 2021), as shown in Figures 11–13. Across these backbones, *Semantics-Agnostic Init.* generally achieves performance comparable to *Semantic Init.*, further showing that its effectiveness is not limited to a specific visual backbone.

Table 4: Hyperparameter configurations and search spaces for linear probing and full fine-tuning.

Hyperparameter	Linear Probe	Full Fine-tuning
Optimizer	AdamW	AdamW
Learning rate	$\{10^{-3}, 10^{-4}\}$	5×10^{-5}
Weight decay	$\{0, 10^{-2}, 10^{-3}\}$	0
Batch size	32	64
Maximum iterations	5,000	12,800
LR scheduler	Cosine	Cosine
Warmup type	Linear	Linear
Warmup iterations	50	50
Warmup initial LR	10^{-5}	10^{-5}
Dropout	0	0
Learnable logit scaling	Enabled	Disabled
Validation interval	100 iterations	100 iterations
Early-stopping patience	10 validations	10 validations
Inter-class cosine similarity	{0, 0.3, 0.6, 0.9, 1.0}	{0, 0.3, 0.6, 0.9, 1.0}

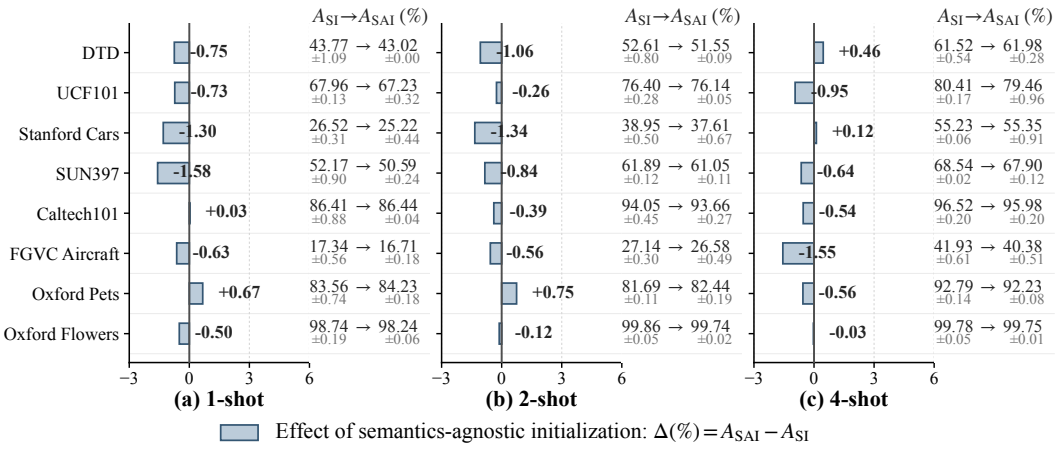


Figure 11: Effect of semantics-agnostic initialization in supervised UMRL using DINOv2 ViT-L/14. Here, A_{SI} and A_{SAI} denote accuracies with Semantic Init. and Semantics-Agnostic Init., respectively.

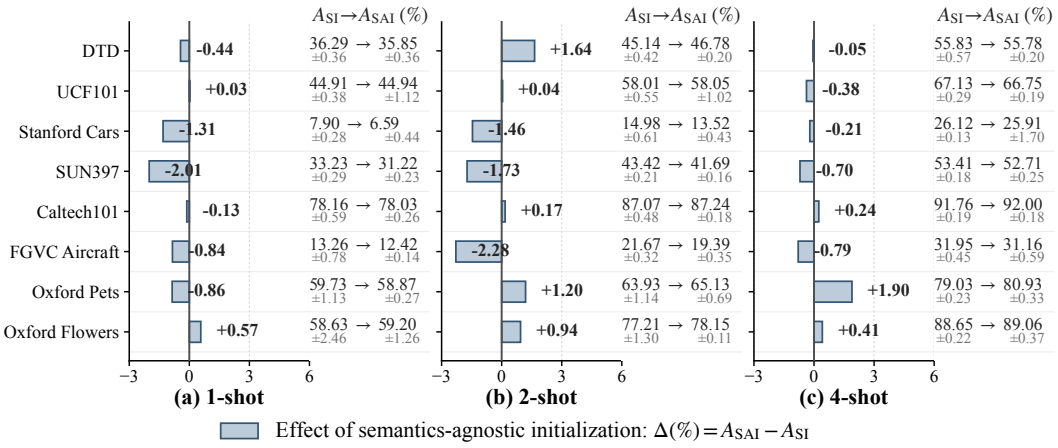


Figure 12: Effect of semantics-agnostic initialization in supervised UMRL using DINO ViT-B/8. Here, A_{SI} and A_{SAI} denote accuracies with Semantic Init. and Semantics-Agnostic Init., respectively.

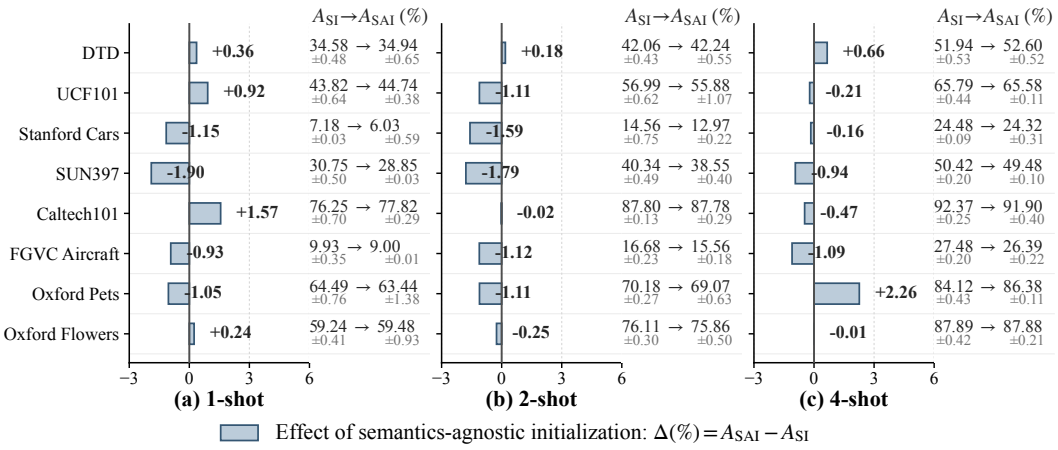


Figure 13: Effect of semantics-agnostic initialization in supervised UMRL using DINO ViT-B/16. Here, A_{SI} and A_{SAI} denote accuracies with Semantic Init. and Semantics-Agnostic Init., respectively.