AGT: Improving Vision-Language Alignment with Generative Target

Nam Kha Nguyen[1] **Wei Emma Zhang**[1] **Xiang Dai**[2] **Mong Yuan Sim**[1,2]

[1]Adelaide University [2]CSIRO

{namkha.nguyen;wei.e.zhang}@adelaide.edu.au

Abstract

Contrastive vision-language models frequently suffer from the modality gap, where image and text embeddings occupy separate regions within a shared representation space. When these frozen, misaligned models are used as encoders to build generative Vision-Language Models (VLMs), current architectures typically rely on a simple projector layer to map visual features into the input embedding space of the language decoder. However, because the backbone visual encoder remains frozen, these projection layers struggle to fully bridge the modality gap. To address these limitations, we introduce **A**lignment with **G**enerative **T**arget (AGT), a framework that transforms textual semantics into the native visual space and enforces an alignment between real and generative visual features, which differs from the existing works that push visual features into the language space. We adopt the diffusion model to obtain the generative visual features and design a distillation framework to align the visual features with the generative targets. Extensive experiments demonstrate that AGT improves captioning performance over baselines, while exhibiting robust zero-shot transfer under domain shifts. Furthermore, the framework enhances zero-shot image-text retrieval and image classification, proving that aligning visual features with generative targets structurally improves the shared representation space and preserves fine-grained discriminative power. [1]

1 Introduction

Vision-language models have become the backbone of modern multimodal systems, supporting tasks from image retrieval and classification to open-ended captioning and visual dialogue. Foundational models, including CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), have established robust paradigms for these tasks. Despite demonstrating remarkable multi-modal alignment capabilities, the shared representation space learned by these foundational models suffer from a fundamental limitation formally identified as the "modality gap" (Liang et al., 2022). Empirical analysis shows that image and text embeddings do not fully overlap; instead, they occupy separate, narrow regions within the shared hypersphere. As a result, the feature space remains disconnected, which significantly restricts smooth cross-modal interpolation and detailed semantic integration, which can cause "blindness" problem in VLMs (Sim et al., 2025a).

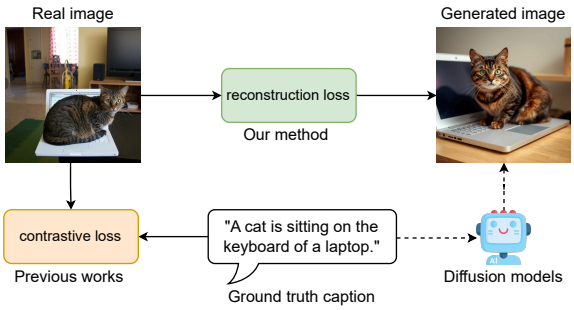

Figure 1: Illustration of ground truth image-caption pair and image generated from the corresponding caption. While the human caption is sparse, the real image contains dense details that text does not capture. Previous works use contrastive learning to pull those slightly different features into the same space. Our proposed method is inspired by this observation and performs alignment natively in the visual space using generative targets to bridge this gap.

To build generative Vision-Language Models (VLMs) using these misaligned encoders, research has mainly added intermediate "connector" layers that transform visual features into the text embedding space of a Large Language Model (LLM). Frameworks like BLIP-2 (Li et al., 2023) deploy a Querying Transformer (Q-Former) to extract and compress text-aligned visual features via cross-attention mechanisms. Alternatively. architectures

[1]Code is available at github.com/noonamkha/AGT

such as LLaVA (Liu et al., 2023) utilise a straightforward Multi-Layer Perceptron (MLP) to project visual patches directly into the LLM's discrete token space. Despite their empirical success, these methodologies rely solely on late-stage mapping layers. Because the foundational visual encoder remains fixed, the underlying visual feature space is never fundamentally reshaped to match the complex, compositional granularity of text (Li et al., 2025), potentially limiting effective visual grounding and contributing to hallucinations in LVLMs (Sim et al., 2026a). Consequently, projection layers struggle to fully bridge the semantic divide, forcing the LLM to decipher misaligned visual mappings (Sim et al., 2025b).

Real-world images contain dense, high-frequency details, whereas human-written captions are inherently sparse and abstract (Urbanek et al., 2024; Viltoriano et al., 2026), often lacking the fine-grained regional information needed for effective visual-grounding (Sim et al., 2026b). As illustrated in Figure 1, a brief human caption cannot capture every nuanced visual detail from the real image, such as the complex background landscape. Previous works often use contrastive learning to pull these weakly aligned features into the same space, which struggles to fully bridge the divide. Conversely, an image generated by a text-to-image model might align better with the provided caption, as it is generated directly from those specific textual semantics. Our proposed method performs alignment natively in the visual space, using these highly aligned generative targets to bridge the gap effectively.

To overcome the profound limitation caused by contrastive loss, unidirectional cross-modal projection and modality gap, we propose AGT (**A**lignment with **G**enerative **T**arget). Our proposed framework fundamentally rethinks the alignment target: rather than forcing real images into an incompatible textual space, we pull textual semantics into the native visual space and enforce an alignment between real and generative visual features. Because modern diffusion models possess a latent space that is inherently conditioned on and aligned with textual semantics, leveraging these models to synthesise targets effectively translates the cross-modal alignment problem into an intra-modal (visual-to-visual) alignment problem.

In summary, our contributions are as follows:

- We introduce AGT, an alignment framework that eliminates modality gap at the point of feature extraction by utilising diffusion-generated images as supervisory targets.
- We implement multimodal alignment natively within the visual feature space. By using a reconstruction loss on the vision encoder, we bypass the need for computationally heavy contrastive loss and simultaneously reduce the burden of transmitting visual information for the projection layer.
- We demonstrate that AGT establishes better performance on image captioning tasks compared to other baselines. Furthermore, it exhibits robustness in zero-shot cross-model retrieval and image classification tasks, proving that the vision encoder's enhanced feature extraction capabilities remain effective.
- We demonstrate that AGT inherently mitigates the modality gap geometrically by pulling real visual features toward generative targets, thereby structurally reorganizing the visual manifold to be closer to the textual space.

2 Related Work

Recent research has increasingly utilised diffusion models to resolve the modality gap. Some approaches exploit this generative prior to synthesise highly aligned training data. ICSD (Ma et al., 2024) employs a LLM to aggregate diverse perspectives into comprehensive descriptions, which then guide a diffusion model to generate intricate, multi-context images. Similarly, SaCap (Liu et al., 2025) uses an iterative retrieval module to swap real images with optimised synthetic proxies to construct hard prompts. Beyond pre-training, recent frameworks like LoRA-Loop (Wang et al., 2025) leverage Stable Diffusion to generate synthetic replay samples for continual learning, using task-specific low-rank adapters to bridge the domain gap between real and generated data distributions.

Moving beyond using synthetic images as pure data substitutes, some frameworks manipulate synthetic features to enforce cross-modal alignment. SynTIC (Liu et al., 2024) utilises global feature contrastive learning to optimise the pseudo-features of synthetic images toward the distribution of real image features within the CLIP space, subsequently utilising these aligned features as training prefixes for language decoders. To handle defective details in synthetic images, PCM-Net (Luo et al., 2024)

utilizes a Patch-wise Cross-modal feature Mix-up mechanism. This approach detects salient visual concepts based on CLIP space similarity and selectively fuses patch-wise visual features with textual features. Rather than synthesizing images, GenDA (Nguyen et al., 2026) conditions a diffusion model on visual tokens concatenated with the caption, back-propagating its reconstruction loss into the vision encoder to capture details the caption omits.

While the mentioned methods successfully leverage synthesis for cross-modal alignment, they primarily treat synthetic images as direct data substitutes (Liu et al., 2025), rely on heuristic feature mixing to mask generative artifacts (Luo et al., 2024), or use contrastive loss, which maybe very computationally heavy (Liu et al., 2024). Furthermore, they continue to align modalities by pulling visual features toward a textual space or by operating in an external projected space. Our proposed framework addresses these limitations by aligning directly within the image feature space. By utilizing high-fidelity images generated from ground-truth captions as supervisory targets, the vision encoder is forced to structurally reshape its native visual features. Minimising the geometric distance between real and synthetic visual features with reconstruction loss natively eliminates the modality gap at the point of feature extraction, mitigating the representational bottleneck of projection layers without the need of heavy contrastive loss.

3 Method

3.1 Overview

As illustrated in Figure 2, the main architecture of the model during inference follows standard VLMs architecture with three components, including a Visual Encoder (f_θ), an MLP Connector (ϕ) and a language decoder. During training, the framework utilises a frozen generative prior and a dual-branch distillation setup that operates on a triplet structure: a real image x_{real}, its ground-truth caption y, and a set of generated images $\mathcal{X}_{gen}$ generated from y. In the dual-branch architecture, a trainable Student Encoder processes real images, while a frozen, momentum-updated Teacher Encoder processes synthetic images. The model is optimised via a compound objective: a global alignment loss that pulls the Student's representation of the real image toward the Teacher's representation of the generative target, and a captioning loss that ensures the preservation of fine-grained local features for text generation.

3.2 Diffusion Images as Generative Target

To establish the generative visual targets essential for the proposed alignment framework, we use a pre-trained, frozen diffusion model to generate a set of images for every real image-text pair (x_{real}, y) in the training dataset, transform the standard paired dataset into a triplet configuration (x_{real}, x_{gen}, y) for training.

A single generated image, however, represents only one possible visual realisation of a caption. Diffusion models are inherently stochastic, which means a single text prompt (i.e., the ground truth caption) y (i.e., the ground truth caption) could generate one-to-many mapping that can describe an infinite permutation of visual scenes (Ho et al., 2020). Relying on a single synthetic image (x_{gen}) as an alignment target introduces noise into the alignment process. This can force the Student encoder to overfit to random generation artifacts or layout quirks caused by a specific noise seed. To mitigate this, we use a multi-sample semantic integration strategy. For every caption y, we generate a set of K generative variations: $\mathcal{X}_{gen} = \{x^1_{gen}, x^2_{gen}, \ldots, x^K_{gen}\}$. We hypothesise that while individual images contain variance and noise, the expected value of their feature representations constitutes a more balanced visual representation of the text semantics. By aggregating features across these K samples, we marginalise out the generative noise, yielding a robust semantic target.

3.3 Distillation Framework for Alignment

To execute the representational alignment between the real observations x_{real} and the generative target $\mathcal{X}_{gen}$ without inducing dimensional collapse or catastrophic forgetting, we employ a Student-Teacher structure inspired by BYOL (Grill et al., 2020), SimSiam (Chen and He, 2021) and DINO (Caron et al., 2021) to enforce alignment without the need for negative pairs.

Student encoder (trainable) parameterized by θ_s, processes the real image x_{real}. As a Vision Transformer, it outputs two distinct feature sets: a global token, denoted as $v^{[CLS]}_{real}$, serves as the aggregated representation of the image which will be used for semantic alignment; and a sequence of local patch tokens, denoted as v^{patch}_{real}, captures fine-grained spatial details used for caption reconstruction.

Teacher encoder (EMA frozen) is parameterized by θ_t, processes the set of generated images $\mathcal{X}_{gen}$.

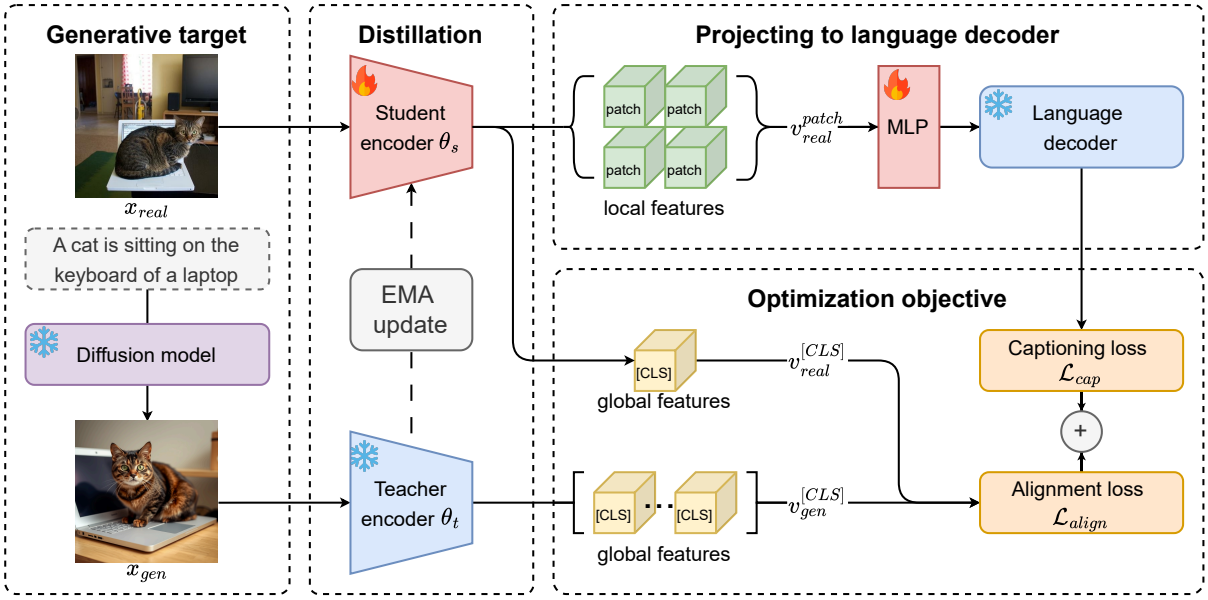

Figure 2: The AGT framework utilizes a distillation setup. A trainable Student encoder processes the real image (x_{real}) to produce local patch features for the language decoder and a global token ($v_{real}^{[CLS]}$) for alignment. Simultaneously, a frozen Teacher encoder processes a set of generated images (x_{gen}) to produce a balanced semantic target representation (v_{gen}).

Since the synthetic images are solely utilised to provide a global semantic compass for alignment, not for spatial reconstruction, the local patch tokens from the Teacher are deliberately discarded. To obtain the balanced target representation, each image x_{gen}^k in the set is passed through the Teacher to extract its corresponding global token. The final target representation is computed as the normalised average of these tokens:

$$v_{gen} = \frac{1}{K} \sum_{k=1}^{K} f_{\theta_t}(x_{gen}^k).$$

This Teacher branch operates strictly in an evaluation mode without computing gradients, serving entirely as a stable feature extractor for the semantic targets.

3.4 Optimization Objective

The global training objective is formalised as a compound loss function that balances the generative feature alignment with the preservation of natural language generation capabilities. This ensures the model learns cross-modal alignment without sacrificing its fundamental identity as a vision-language system. The total loss is defined as:

$$\mathcal{L}_{total} = \lambda \cdot \mathcal{L}_{align} + \mathcal{L}_{cap},$$

where λ is a scalar hyperparameter balancing the gradient of the two contributing losses: Alignment loss $\mathcal{L}_{align}$ and Captioning loss $\mathcal{L}_{cap}$.

Alignment loss $\mathcal{L}_{align}$. To align the real and generated visual manifolds, the loss minimises the distance between their respective global representations. In accordance with the BYOL formulation (Grill et al., 2020), we use the Squared Euclidean distance computed over L_2-normalized features:

$$\begin{aligned}\mathcal{L}_{align} &= \|v_{real}^{[CLS]} - v_{gen}^{[CLS]}\|_2^2 \\ &= 2 - 2 \cdot \text{sim}(v_{real}^{[CLS]}, v_{gen}^{[CLS]}),\end{aligned}$$

where $\text{sim}(\cdot, \cdot)$ represents the cosine similarity. By minimising this loss, the Student encoder is forced to map the real image directly into the subspace occupied by the generative target. This formulation closes the modality gap by performing the alignment natively within the visual feature space, entirely bypassing the geometric restrictions of the text encoder's latent cone.

Captioning loss $\mathcal{L}_{cap}$. While global alignment ensures high-level semantic grounding, the visual encoder must retain dense, descriptive features to function effectively as a VLM. To enforce this, we adapt LLaVA-style (Liu et al., 2023) to project the Student's local patch features v_{real}^{patch} through a trainable MLP connector ϕ. The resulting projected tokens are prepended as visual prefixes to the input space of a frozen language decoder. The language model then reconstructs the ground-truth caption

y auto-regressively. The standard auto-regressive language modelling objective (cross-entropy loss over the vocabulary space) is applied:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, \phi(V_{real}^{patch}))$$

Crucially, only the Student branch feeds the captioner. This ensures that the encoder learns to extract features from real images that are sufficient for text generation, preventing the model from collapsing into a representation that is only good for alignment but ineffective for downstream tasks.

4 Experiments Setup

We conduct extensive experiments on a diverse collection of datasets to evaluate our model across multiple tasks, including image captioning, zero-shot image retrieval, and zero-shot image classification. We also analyse the capability of our framework to reduce the modality gap.

Datasets. We evaluate our approach on two widely adopted image captioning benchmarks: MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014). For data partitioning, we follow the standard Karpathy split (Karpathy and Fei-Fei, 2015).

Evaluation Metrics. For image captioning, we employ BLEU (B@n) (Papineni et al., 2002), METEOR (M) (Denkowski and Lavie, 2014), ROUGE-L (R-L) (Lin and Och, 2004), CIDEr (C) (Vedantam et al., 2015), and SPICE (S) (Anderson et al., 2016). For zero-shot image–text retrieval, we evaluate both image-to-text and text-to-image retrieval tasks using Recall@1, Recall@5, and Recall@10. In the zero-shot image classification setting, we report Top-1 accuracy on unseen categories.

Model Architecture. For our architectural framework, we adopt CLIP (ViT-B/32) (Radford et al., 2021) as the vision encoder and GPT-2 (Radford et al., 2019) as the text decoder. This selection serves two primary purposes. First, this configuration is a widely accepted baseline in VLM research, ensuring fair comparisons with prior works. Second, GPT-2 is a pure autoregressive completion model; its lack of instruction tuning makes it particularly well-suited for standard image captioning tasks, as it relies on the conditional input rather than learned chat behaviours.

For the text-to-image model, we use `black-forest-labs/FLUX.1-schnell` (Labs, 2024) as our frozen generator thanks to the ability to achieve visual quality and text alignment in an accelerated trajectory of only 1 to 4 diffusion steps.

Implementation Details. For the image generation pipeline, we employ the diffusion model to generate a set of K=8 images for each caption in the MSCOCO and Flickr30k datasets. To balance image quality with inference time, we set the number of inference steps to 4. Additionally, we utilise a relatively high guidance scale of 10.0 to ensure the generated images strictly adhere to the captions. Training is conducted for 10 epochs with a batch size of 128. The MLP connector is fully trained, whereas the vision encoder is updated using LoRA (Hu et al., 2022) with a rank of 4. For optimisation, we use AdamW. Using the cyclical learning rate technique (Smith, 2017), we identified an optimal learning rate of 1.5×10^{-3} for the MLP connector, alongside a lower rate of 1×10^{-4} for the vision encoder to preserve its pretrained visual features. We also applied the same method to determine λ at the beginning of training for each configuration. We apply a weight decay of 0.01 to the vision encoder and a reduced weight decay of 1×10^{-5} to the MLP connector. We employ a cosine annealing schedule with a warm-up ratio of 0.03.

5 Result

5.1 Evaluation on Image Captioning

We evaluate our method's performance on the MSCOCO and Flickr30k datasets against prior works that also use diffusion models for image captioning, including SynTIC (Liu et al., 2024), PCM-Net (Luo et al., 2024), and SaCap (Liu et al., 2025). We also compare against two controlled baselines that share the same architecture, backbone and training data. *Baseline* removes the alignment loss $\mathcal{L}_{align}$ and is trained with $\mathcal{L}_{cap}$ only. *Baseline + gen images* additionally uses the same generated images as AGT, but treats them as extra training data: each generated image is paired with its caption and trained with $\mathcal{L}_{cap}$ only. This variant has access to the same generative information as AGT. With these baselines, we isolates whether AGT's gains stem from the alignment objective itself or merely from the additional information provided by the diffusion model.

In-domain captioning. As detailed in Table 1, our method establishes better performance on metrics that heavily weight semantic accuracy and scene graph alignment, specifically CIDEr and SPICE. On MSCOCO, AGT achieves a CIDEr score of

Table 1: In domain image captioning task on MSCOCO and Flickr30k. The metrics include BLEU@4 (B@4), METEOR (M), ROUGE-L (R-L), CIDEr (C), and SPICE (S).

Method	MSCOCO					Flickr30k				
	B@4	M	R-L	C	S	B@4	M	R-L	C	S
SynTIC	29.9	25.8	53.2	101.1	19.3	22.3	22.4	47.3	56.6	16.6
PCM-Net	**31.5**	25.9	53.9	103.8	19.7	26.9	23.0	50.1	61.3	16.8
SaCap	31.0	25.9	53.4	104.2	19.7	**27.1**	22.4	49.3	64.3	15.9
Baseline	24.6	25.3	52.8	102.7	19.5	24.3	22.4	49.4	64.1	17.8
Baseline + gen images	24.8	25.8	53.2	103.1	19.7	24.8	22.5	49.6	64.4	18.0
AGT (ours)	26.9	**26.4**	**54.1**	**106.4**	**21.9**	26.0	**24.7**	**50.2**	**67.2**	**19.8**

Table 2: Cross-domain image captioning task on MSCOCO and Flickr30k.

Method	MSCOCO → Flickr30k					Flickr30k → MSCOCO				
	B@4	M	R-L	C	S	B@4	M	R-L	C	S
SynTIC	17.9	18.6	42.7	38.4	11.9	14.6	19.4	40.9	47.0	11.9
PCM-Net	20.8	19.2	45.2	45.5	12.9	17.1	19.6	43.2	54.9	12.8
SaCap	21.2	20.2	**45.7**	50.9	13.2	17.5	19.7	42.8	62.5	13.5
Baseline	19.7	19.5	42.7	51.3	12.5	16.4	19.7	41.4	62.9	12.7
Baseline + gen images	20.0	19.8	42.9	51.5	12.6	16.8	19.6	41.7	63.2	13.3
AGT (ours)	**21.6**	**20.4**	43.5	**52.7**	**13.9**	**18.3**	**19.8**	**43.4**	**63.5**	**13.9**

106.4 and SPICE of 21.9, outperform the closest competitor, SaCap (104.2 and 19.7, respectively). This trend holds on Flickr30k, validating that aligning real features to generative targets enriches the visual representation with highly descriptive semantics. We observe that AGT yields lower performance in exact n-gram matching metrics compared to prior works. Specifically, on MSCOCO, our BLEU@4 (26.9) fell behind PCM-Net (31.5) and SaCap (31.0). A similar gap is present in Flickr30k (26.0 vs. 27.1). We attribute this discrepancy to the nature of our alignment objective. Because AGT aligns the native visual features to a multi-sample semantic expectation (averaged across $K = 8$ diffusion generations) rather than textual tokens, the model is encouraged to generate semantically rich and diverse captions rather than memorizing the exact n-gram phrasing of the training data. Consequently, while AGT occasionally fell behind from the exact lexical structure rewarded by BLEU@4, the substantial gains in CIDEr and SPICE demonstrate that the underlying meaning, objects, and relationships are captured with higher fidelity.

Additionally, we want to test whether AGT benefits simply from the extra information provided by diffusion model. Table 1 shows that adding generated images as training data yields only marginal improvements over the baseline (+0.4 CIDEr and +0.2 SPICE on MSCOCO; +0.3 CIDEr and +0.2 SPICE on Flickr30k). In contrast, AGT, which uses the same generated images, improves CIDEr by 3.7 and SPICE by 2.4 on MSCOCO, and by 3.1 and 2.0 on Flickr30k. We attribute this to the different roles of the generated images: as training data, they are simply additional image-caption pairs, whereas as alignment targets, their representation provides a generative target that actively reshapes the encoder's visual feature space.

Cross-domain captioning. To assess robustness against domain shifts, we conduct cross-dataset evaluations (MSCOCO → Flickr30k and vice versa) without any target-domain fine-tuning. As shown in Table 2, AGT demonstrates better zero-shot transfer capabilities. When transferring from Flickr30k to MSCOCO, AGT dominates across all five metrics. In the MSCOCO → Flickr30k setting, AGT achieves the highest scores in BLEU@4, METEOR, CIDEr, and SPICE, although we note a minor drop in ROUGE-L (43.5) compared to SaCap (45.7). This strong cross-domain performance highlights an advantage of our framework: by resolving the modality gap at the point of the vision encoder, AGT relies less on the fragile, dataset-specific memorization of the MLP projection layer. Reshaping the visual feature space via generative target allows AGT to extract highly generalized, text-aligned features that remain robust across visual domains. Consistent with the in-domain results, using generated images as additional training data brings only small changes in the cross-domain setting (at most +0.6 SPICE, and a slight METEOR drop on Flickr30k → MSCOCO), while

Table 3: Zero-shot Image-to-Text (I → T) and Text-to-Image (T → I) retrieval tasks on COCO and Flickr30k benchmark. The metrics include Recall@1,5,10.

Method	Task	MSCOCO			Flickr30k		
		R@1 (↑)	R@5 (↑)	R@10 (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)
CLIP	I → T	51.8	76.8	84.3	44.1	68.2	77.0
CLIP + AGT	I → T	**52.4**	**77.3**	**84.8**	**46.5**	**68.9**	**77.6**
CLIP	T → I	32.7	57.7	68.2	24.7	45.1	54.6
CLIP + AGT	T → I	**34.5**	**58.4**	**69.2**	**26.4**	**45.7**	**55.3**

Table 4: Zero-shot Classification tasks on 3 benchmarks. The metrics used is top-1 accuracy.

Method	ImageNet	Caltech 101	SUN397
CLIP	63.4	88.1	63.2
CLIP + AGT	**64.7**	**89.3**	**65.3**

AGT yields larger gains, most notably in SPICE (+1.3 and +0.6 over Baseline + gen images).

5.2 Evaluation on Downstream Tasks

Beyond generation, We also assess AGT's impact on zero-shot discriminative tasks using the frozen vision encoder. For these experiments, we extract the global feature $v_{real}^{[CLS]}$ from the AGT-aligned visual encoder and compute its cosine similarity against standard CLIP text embeddings.

Zero-Shot Image-Text Retrieval. Table 3 reports the performance of our framework on Image-to-Text (I → T) and Text-to-Image (T → I) retrieval. By utilizing the AGT-aligned visual features in conjunction with standard CLIP text features, we observe consistent improvements across all Recall metrics on both MSCOCO and Flickr30k. For instance, T → I R@1 on MSCOCO improves from 32.7 (base CLIP) to 34.5 (CLIP + AGT). This proves that pulling real visual features toward generative targets structurally reorganizes the visual manifold, inherently bridging the visual regions and their corresponding textual regions.

Zero-Shot Image Classification. A common failure mode of contrastive alignment is the degradation of dense visual features in favour of coarse semantic concepts. To verify that AGT preserves fine-grained discriminative power, we evaluate zero-shot classification on ImageNet, Caltech 101, and SUN397. As reported in Table 4, our alignment strategy not only preserves but consistently improves native classification capabilities across all three benchmarks. CLIP + AGT achieves 64.7 Top-1 accuracy on ImageNet, a gain of 1.3 over the standard CLIP baseline. By using multi-sample synthetic images ($\mathcal{X}_{gen}$) to marginalise out generative noise, the alignment loss acts as a powerful regularizer. It guides the vision encoder to separate tightly clustered visual concepts in the latent space without relying on heavy contrastive loss or negative sampling.

5.3 Evaluation on Modality Gap

In this section, we evaluate the effectiveness of AGT in reducing the modality gap. Specifically, we conduct UMAP-based visualization and similarity assessments on embeddings from the MSCOCO validation set. To facilitate this analysis, we denote $\mathbf{z}_{text}$ as the text embedding from CLIP text encoder, $\mathbf{z}_{img}$ as the image embedding from baseline CLIP image encoder and $\widetilde{\mathbf{z}}_{img}$ as the image embedding from our CLIP + AGT.

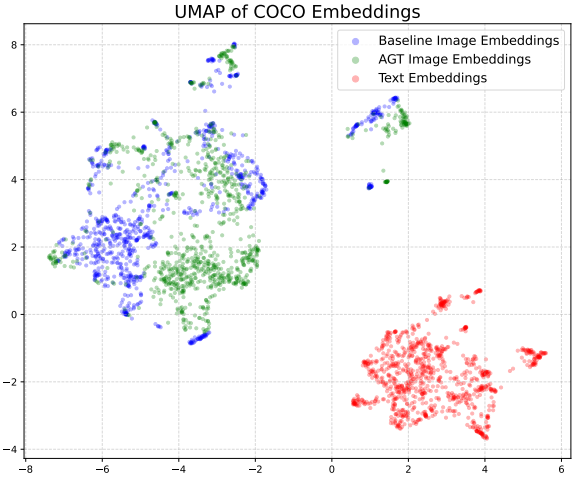

Figure 3: UMAP visualization of COCO validation set embeddings. Red points represent text embeddings, blue points denote baseline image embeddings, and green points show image embeddings produced by AGT.

UMAP Visualization of the Embedding Space. Figure 3 presents a UMAP (McInnes and Healy, 2018) projection of the COCO embeddings, showing three types of embeddings: text embeddings $\mathbf{z}_{text}$ (red), image embeddings extracted by base-

line $\mathbf{z}_{img}$ (blue) and image embeddings extracted by our AGT method $\widetilde{\mathbf{z}}_{img}$ (green). The distinct clusters in Figure 3 show the inherent modality gap between visual and text data. However, the image embeddings extracted by our AGT method are clustered noticeably closer to the text embeddings than those extracted by the baseline model.

Table 5: Average cosine similarity scores between image and text embedding pairs

Comparison Pair	Cosine Similarity
$\mathbf{z}_{text} \cdot \mathbf{z}_{text}$	0.6527
$\mathbf{z}_{text} \cdot \mathbf{z}_{img}$	0.3750
$\mathbf{z}_{text} \cdot \widetilde{\mathbf{z}}_{img}$	0.4369

Cosine Similarity Analysis of Modality Alignment. To quantify this visual observation, we measure the average cosine similarity between embedding pairs. As detailed in Table 5, the cosine similarity between text embeddings of captions belonging to the same images ($\mathbf{z}_{text} \cdot \mathbf{z}_{text}$) achieves the highest score of 0.6527, serving as an upper-bound reference. When examining the cross-modal alignments, the similarity between the text embedding and the image embedding extracted by our method ($\mathbf{z}_{text} \cdot \widetilde{\mathbf{z}}_{img}$) reaches 0.4369, demonstrating a clear improvement over the 0.3750 similarity score achieved by the baseline image embedding ($\mathbf{z}_{text} \cdot \mathbf{z}_{img}$). Together, these findings confirm that aligning visual representations with generative targets successfully draws the visual feature space closer to the textual space, structurally improving the modality gap.

5.4 Ablation Studies

To analyse the effect of different types of alignment target on our framework, we conduct ablation studies comparing the primary AGT architecture against variants conditioned on alternative visual targets. Specifically, we compare our method against a version that employs SDXL-Turbo (Stability AI, 2025), which is a weaker diffusion model for image generation, as well as a version that completely replaces generated images with standard augmented images.

As illustrated in Figure 4, the quality and type of the target images significantly affect the framework's overall performance. The variant utilising SDXL-Turbo for generation demonstrates an improvement over the baseline across metrics such as BLEU@4 (25.8 vs. 24.6) and SPICE (20.7 vs. 19.5). However, it remains inferior to the main AGT method using FLUX.1-schnell, which achieves higher scores of 26.9 on BLEU@4 and 21.9 on SPICE. Conversely, the model variant trained with augmented images performs very similarly to the baseline, yielding a BLEU@4 score of 24.4 and a SPICE score of 19.2. We hypothesize this stagnation occurs because augmented images remain semantically and geometrically similar to the original images. Consequently, the alignment loss fails to introduce any meaningful new learning signals. Without a strong alignment target to pull the representation space, the model's performance relies entirely on the captioning loss, thereby mirroring the behaviour of the baseline architecture.

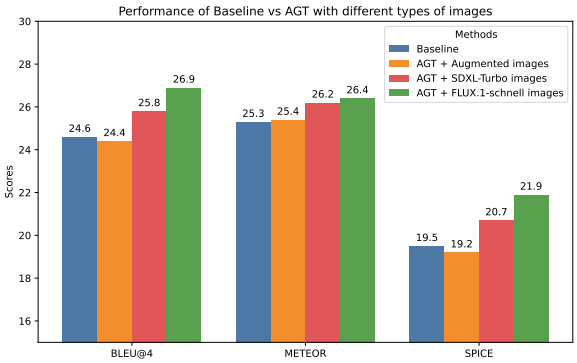

Figure 4: Performance of Baseline vs AGT with different types of alignment target

6 Conclusion

In this work, we introduced a new paradigm to mitigate the modality gap in VLMs. We designed AGT, a framework to pull text into visual space and evaluated it on different tasks. AGT establishes better performance on image captioning metrics, and shows strong zero-shot cross-domain transfer capabilities. Beyond generative tasks, AGT structurally improves the shared representation space, preserving and enhancing fine-grained discriminative power for zero-shot image-text retrieval and image classification. Ultimately, our modality gap analysis confirms that pulling real visual observations toward integrated generative targets successfully bridges the representational divide between visual and textual spaces.

Limitations

We identify the following limitations in our current approach. Firstly, AGT yields lower performance in image captioning task when evaluated using ex-

act n-gram matching metrics compared to baselines. The framework aligns native visual features to a multi-sample semantic expectation rather than directly mapping to discrete textual tokens, the model is encouraged to produce semantically diverse and rich captions. Furthermore, AGT relies on a multi-sample semantic integration strategy to mitigate the inherent stochasticity of diffusion models. This requires generating a set of $K = 8$ high-fidelity synthetic images for every ground-truth caption in the training dataset, which introduces a substantial computational and storage overhead during the dataset preparation phase compared to standard VLM training pipelines.

Acknowledgments

This study was funded by the Australian Research Council Early Career Industry Fellowships (IE230100119) and Linkage Project (LP230200821).

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V, volume 9909 of Lecture Notes in Computer Science, pages 382–398. Springer.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 9630–9640. IEEE.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 15750–15758. Computer Vision Foundation / IEEE.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA, pages 376–380. The Association for Computer Linguistics.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 4904–4916. PMLR.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 3128–3137. IEEE Computer Society.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`. Accessed: 2026-02-20.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of Proceedings of Machine Learning Research, pages 19730–19742. PMLR.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulić, and Anders Søgaard. 2025. Lost in embeddings: Information loss in vision–language models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22676–22693, Suzhou, China. Association for Computational Linguistics.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information

Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain, pages 605–612. ACL.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, volume 8693 of Lecture Notes in Computer Science, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Zhiyue Liu, Jinyuan Liu, Xin Ling, Qingbao Huang, and Jiahai Wang. 2025. Synthesize then align: Modality alignment augmentation for zero-shot image captioning with synthetic data. Knowl. Based Syst., 315:113274.

Zhiyue Liu, Jinyuan Liu, and Fanrong Ma. 2024. Improving cross-modal alignment with synthetic pairs for text-only image captioning. In Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 3864–3872. AAAI Press.

Jianjie Luo, Jingwen Chen, Yehao Li, Yingwei Pan, Jianlin Feng, Hongyang Chao, and Ting Yao. 2024. Unleashing text-to-image diffusion prior for zero-shot image captioning. In Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVII, volume 15115 of Lecture Notes in Computer Science, pages 237–254. Springer.

Feipeng Ma, Yizhou Zhou, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. 2024. Image captioning with multi-context synthetic data. In Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 4089–4097. AAAI Press.

Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. CoRR, abs/1802.03426.

Nam Kha Nguyen, Wei Emma Zhang, Xiang Dai, Mong Yuan Sim, and Weitong Chen. 2026. Beyond caption loss: Generative distillation for modality alignment in vision-language models. In Findings of the Association for Computational Linguistics: EMNLP 2026, pages 40021–40034, Budapest, Hungary. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA, pages 311–318. ACL.

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. 2021. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025a. Can vlms actually see and read? A survey on modality collapse in vision-language models. In Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, volume ACL 2025 of Findings of ACL, pages 24452–24470. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, Biaoyan Fang, Sarbin Ranjitkar, Arjun Burlakoti, Jamie Taylor, and Haojie Zhuang. 2025b. The more, the better? A critical study of multimodal context in radiology report summarization. In Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025, pages 19116–19131. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Byron Fang. 2026a. The model knows where to look: Visual-grounding-guided head suppression for mitigating hallucination in LVLMs. In Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3347–3367, Hengqin, China. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, Byron Fang, and Miao Liu. 2026b. Learning pathology-localized medical vision–language representations with atomic and grounded disease supervision. In Findings of the Association for Computational Linguistics: EMNLP 2026, pages 49550–49564, Budapest, Hungary. Association for Computational Linguistics.

Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. In 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017, pages 464–472. IEEE Computer Society.

Stability AI. 2025. SDXL-Turbo. https://huggingface.co/stabilityai/sdxl-turbo. Accessed: 2026-02-20.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 26690–26699. IEEE.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 4566–4575. IEEE Computer Society.

Robin Viltoriano, Wei Emma Zhang, Hu Wang, Mong Yuan Sim, and Yanjun Shu. 2026. Peacap: Patch-level retrieval for lightweight retrieval-augmented image captioning. In Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, MelbourneVICAustralia, July 20-24, 2026, pages 4216–4220. ACM.

Kaihong Wang, Donghyun Kim, and Margrit Betke. 2025. Lora-loop: Closing the synthetic replay cycle for continual vlm learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 445–454.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics, 2:67–78.