SJEPA: Joint Embedding Predictive Architecture for Synthetic-to-Real Alignment

Anonymous CVPR submission

Paper ID 20

Abstract

Generative models have made it possible to synthesize highly realistic images, potentially providing an abundant data source for training machine learning models. However, despite the advantages of these synthesizable data sources, the indiscriminate use of generated images for training can cause mode collapse due to inherent modality discrepancies between real and synthetic domains. Consequently, state-of-the-art self-supervised learning methods like the Image Joint Embedding Predictive Architecture (I-JEPA) typically rely exclusively on real images, leaving the rich potential of synthetic distributions largely unexplored. In this paper, we introduce SJEPA, a novel framework designed to bridge this gap by explicitly treating generated images as a separate modality and aligning them with real images within a shared latent space. By harmonizing these representations, SJEPA effectively transfers the diversity of synthetic data to downstream tasks without overfitting to the peculiarities of synthetic data. Extensive experiments demonstrate that our method achieves superior performance compared to strong baselines, including notable enhancements for the large multimodal model, LLaVA. Specifically, SJEPA significantly improves performance in image captioning on COCO, zero-shot image retrieval on COCO and Flickr30k, zero-shot image classification across eight diverse datasets, and long caption retrieval on ShareGPT4V. Furthermore, we observe positive generated data scaling trends when training across datasets of increasing volume, including COCO, CC3M, and CC12M, highlighting the scalability and robustness of our approach.

1. Introduction

Generative models, such as diffusion models [10, 23] and GANs [9], have revolutionized computer vision by enabling the synthesis of highly realistic images. These generated images offer a rich, scalable data source that can significantly augment training datasets, enhance data diversity, and reduce the dependency on costly real-world data collection [28].

However, effectively leveraging this synthetic data in self-supervised learning paradigms remains a profound challenge. Current state-of-the-art methods, such as the Image Joint Embedding Predictive Architecture (I-JEPA) [1], typically rely exclusively on real images. When generated images are indiscriminately incorporated into these training pipelines, it often leads to a phenomenon known as mode collapse [25]. In such cases, the model's performance severely deteriorates due to an over-reliance on generated content that fails to generalize well to real-world scenarios.

The primary source of this collapse lies in the failure to recognize that generated images, despite their realism, represent a distinct data modality that deviates from real-world distributions in subtle but significant ways. Even when generated images appear visually convincing, they often contain underlying artifacts, domain-specific noise, and distinct high-frequency patterns introduced during the generative process. These discrepancies are particularly detrimental to predictive self-supervised frameworks like I-JEPA. Because these architectures learn by predicting the latent representations of target patches from contextual patches, they are highly sensitive to local spatial dependencies. When trained indiscriminately on synthetic data, the predictive model tends to exploit synthetic-specific local artifacts as learning shortcuts. Instead of extracting robust, high-level semantic features, the network learns to predict the generative noise. Consequently, this leads to a severe misalignment in the feature space; the learned representations become tightly coupled to the synthetic manifold and fail to generalize when exposed to real-world data distributions. Furthermore, generative models may inadvertently capture and amplify biases present in their original training data, producing synthetic images that diverge unpredictably from real domains. Addressing this generated-to-real modality gap is therefore critical. Without explicitly decoupling and aligning these modalities, models struggle with overfitting, reduced robustness, and degraded performance on down-

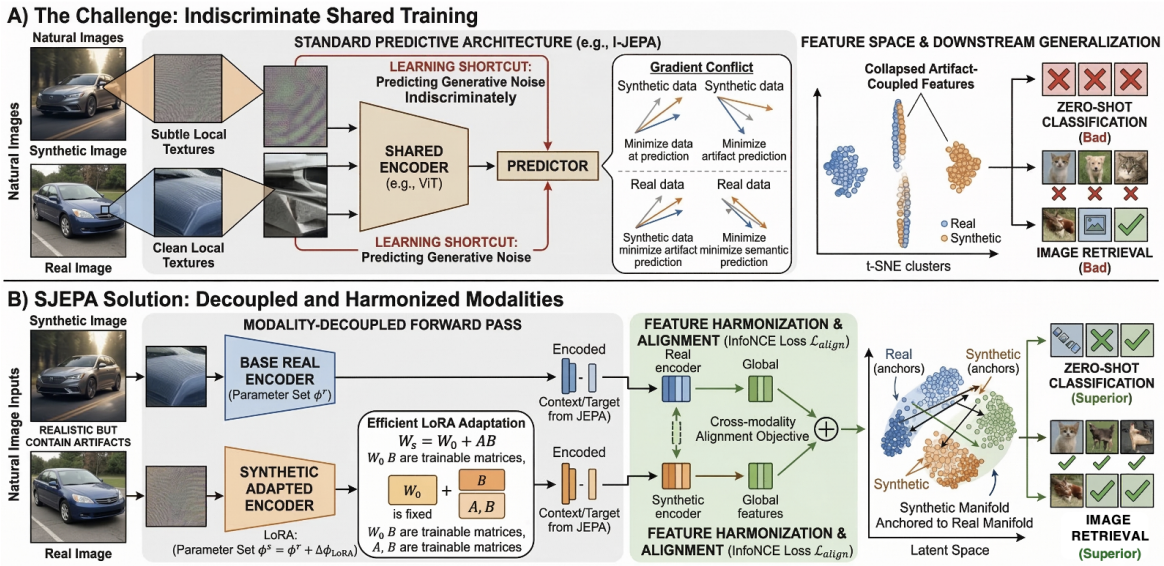

Figure 1. **The Synthetic-to-Real Modality Gap in Predictive Architectures.** While generated images are visually highly realistic, they inherently contain domain-specific noise and subtle generative artifacts. When a standard predictive architecture processes both real and synthetic images through a shared encoder, the network exploits these local artifacts as learning shortcuts, learning to predict generative noise rather than extracting robust, high-level semantic features. This indiscriminate training leads to severe feature-space misalignment and mode collapse, severely degrading downstream generalization on real-world tasks. To address this, our proposed SJEPA explicitly decouples the modalities and harmonizes their latent spaces.

stream tasks, severely limiting the practical utility of synthetic data.

To tackle this challenge, we introduce SJEPA, a novel framework designed for the discriminative use of generated images within a predictive architecture. SJEPA explicitly treats generated images as a separate modality from real images. Unlike conventional methods that mix generated and real data indiscriminately, our approach bridges the two distinct modalities by explicitly aligning their representations within a shared latent space. By harmonizing these representations, SJEPA allows the model to effectively transfer the diversity and scale of synthetic data to downstream tasks, while actively preventing the network from overfitting to the peculiarities of the synthetic domain.

Through extensive experiments across a wide range of tasks, we demonstrate the effectiveness of our SJEPA framework. Specifically, our approach significantly enhances image captioning on COCO, zero-shot image retrieval on COCO and Flickr30k, and zero-shot image classification across eight widely used datasets. SJEPA also enables notable enhancements in the captioning performance of the large multimodal model, LLaVA, and long caption retrieval on ShareGPT4V. Furthermore, we observe positive generated data scaling trends across diverse datasets such as COCO, CC3M, and CC12M, highlighting the scalability and robustness of our method as the volume of training data increases.

Our main contributions are summarized as:

- We introduce SJEPA, a novel self-supervised learning framework that explicitly treats generated images as a distinct modality within a predictive architecture. By decoupling the patch-level prediction process and employing a cross-modality alignment objective, we effectively bridge the modality gap and prevent mode collapse.
- We demonstrate the effectiveness of our framework through extensive experiments across a diverse set of benchmarks, achieving superior performance in image captioning, zero-shot image retrieval, and zero-shot image classification.
- We explore the generated data scaling trends of our framework using large-scale synthetic datasets, demonstrating that SJEPA's robust representation learning consistently improves as the volume of training data increases.

2. Related Work

Self-Supervised Representation Learning. Self-supervised learning has emerged as a dominant paradigm for learning powerful visual representations without manual annotations. Recent advancements have shifted from contrastive learning approaches to joint-embedding predictive architectures (JEPAs), such as I-JEPA [1]. These predictive models learn high-level semantic representations by predicting the embeddings of missing target patches from contextual patches within a latent space. While these architectures excel at modeling local and global spatial

dependencies, they are predominantly designed for, and trained exclusively on, real-world image distributions. When naively exposed to synthetic images, these predictive models tend to overfit to generative noise, treating localized artifacts as predictive shortcuts. Our proposed SJEPA builds upon the foundational predictive architecture but introduces a dual-modality framework to decouple real and synthetic patch processing, ensuring robust, artifact-free representation learning.

Learning from Generated Images. The rapid evolution of generative models, particularly diffusion models [10, 26], has enabled the synthesis of highly realistic images that closely mimic real-world distributions. Prominent models like Stable-Diffusion [23] and the DALL-E series [2, 22] have provided a rich and scalable source of synthetic data, prompting researchers to explore generated image learning to augment traditional training paradigms. Recent works, such as SynCLR [28], have demonstrated that self-supervised frameworks can utilize synthetic data to pre-train visual representations, achieving competitive results compared to models trained on real data. However, a critical limitation in this domain is mode collapse, where an over-reliance on unaligned synthetic data leads to severe performance degradation on real-world tasks. Furthermore, recent studies highlight the inherent risks of training models on recursively generated data [25], emphasizing that models can inherit and amplify errors or artifacts present in synthetic data. SJEPA directly addresses these challenges by proposing a novel strategy that explicitly treats generated images as a distinct modality and aligns them with real images in the same latent space. This explicitly mitigates the risk of mode collapse and enhances model robustness.

Vision-Language Models and Cross-Modal Alignment. Vision-language models, such as CLIP [21], have revolutionized cross-modal understanding by learning joint representations of images and text through contrastive learning objectives. Building on this foundation, models such as Long-CLIP [32] and LLaVA [15] have extended these capabilities to handle longer, more descriptive texts and complex multimodal tasks like visual question answering. Despite their success with large-scale real-world data, these models often struggle when fine-tuned on generated images due to the inherent modality gap. Existing approaches typically integrate generated images into the training process without adequately addressing this discrepancy, mixing generated and real data indiscriminately. By introducing an explicit cross-modality alignment objective into the predictive training pipeline, SJEPA directly addresses the modality discrepancies that limit model performance. This provides robust, aligned representations that significantly boost cross-modal learning across diverse downstream vision-language tasks, including zero-shot retrieval and image captioning.

3. Method

In this section, we present the Synthetic Joint Embedding Predictive Architecture (SJEPA), a novel self-supervised learning framework designed to effectively leverage synthetic images without suffering from mode collapse. We first review the foundational predictive architecture that our method builds upon. Next, we introduce our dual-modality approach for processing real and generated images independently. Finally, we formulate the cross-modality alignment objective that harmonizes these distinct distributions within a shared latent space.

3.1. Preliminaries

Our framework builds upon the Image Joint Embedding Predictive Architecture (I-JEPA), a self-supervised approach that learns semantic representations by predicting the embeddings of missing image regions from contextual patches. Given an input image x, the method extracts a set of target blocks y and a context block x_c. The context block is constructed by masking out the regions corresponding to the target blocks.

The I-JEPA framework consists of three core components: a context encoder f_ϕ, a target encoder f_θ, and a predictor g_ψ. The context encoder processes x_c to produce a latent representation, while the target encoder independently processes the target blocks y. The predictor is then tasked with predicting the target representations from the context representation, conditioned on positional embeddings p corresponding to the target regions. The standard predictive loss is defined as the squared ℓ_2 distance between the predicted and actual target representations:

$$\mathcal{L}_{JEPA} = \mathbb{E}_x \left[\sum_i \left\| g_\psi(f_\phi(x_c), p_i) - f_\theta(y_i) \right\|_2^2 \right]$$

While highly effective for real-world distributions, applying this standard objective indiscriminately to a mixture of real and generated images causes the predictor g_ψ to exploit synthetic-specific high-frequency artifacts as learning shortcuts, severely degrading the learned semantic representations.

3.2. Dual-Modality Predictive Architecture

To address the inherent modality gap between real and synthetic data, SJEPA explicitly treats the real dataset $\mathcal{D}_r = \{x_r\}$ and the synthetic dataset $\mathcal{D}_s = \{x_s\}$ as distinct data distributions. Rather than feeding both domains into a single, shared encoder sequence—which forces the network to reconcile conflicting spatial dependencies and generative noise—we introduce a modality-decoupled forward pass.

Formally, an input image $x \in \mathbb{R}^{H \times W \times C}$ is divided into a sequence of N non-overlapping patches. We sample a

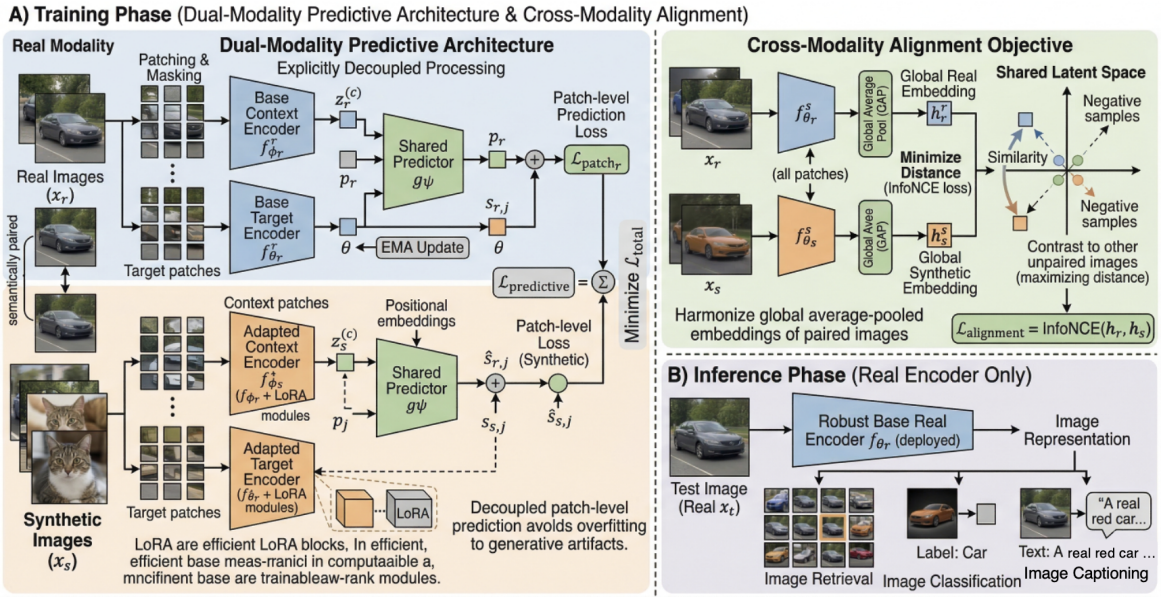

Figure 2. Illustration of the proposed SJEPA framework for self-supervised representation learning via Synthetic-to-Real Alignment. Our approach features a *Dual-Modality Predictive Architecture* that explicitly decouples the processing of real and synthetic images. Real images are processed using base context and target encoders, while synthetic images are processed via computationally efficient Low-Rank Adaptation (LoRA) modules. This decoupled patch-level prediction prevents the shared predictor from overfitting to high-frequency generative artifacts. To ensure the learned synthetic representations enhance real-world understanding, a *Cross-Modality Alignment Objective* explicitly harmonizes the global average-pooled embeddings of semantically paired real and synthetic images within a shared latent space. During inference on downstream tasks, only the robust base real encoder is deployed, circumventing modality discrepancies.

context mask M_c and a set of target masks $\{M_{t,k}\}_{k=1}^{K}$. The context patches are denoted as $x^{(c)} = \{x_i\}_{i \in M_c}$, and the target patches for the k-th block as $x^{(t,k)} = \{x_j\}_{j \in M_{t,k}}$.

We instantiate a base context encoder f^r_ϕ and target encoder f^r_θ for real images. To process synthetic images without catastrophic forgetting of real-world semantics, we introduce computationally efficient adapted encoders, f^s_ϕ and f^s_θ. These synthetic encoders are parameterized using Low-Rank Adaptation (LoRA) applied to the frozen base real encoders. For any linear projection layer within the transformer with a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times d}$, the forward pass for a synthetic token z is formulated as:

$$h_s = zW_0 + z\Delta W = zW_0 + zAB$$

where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$ are trainable low-rank matrices with rank $r \ll d$. Consequently, the synthetic context encoder can be mathematically expressed as $f^s_\phi(\cdot) = f^r_\phi(\cdot; W_0) + \Delta f_\phi(\cdot; A, B)$.

During training, a sampled batch is divided into real and synthetic subsets. For the real modality, the context encoder maps $x^{(c)}_r$ to a latent representation $z^{(c)}_r = f^r_\phi(x^{(c)}_r)$. The predictor g_ψ is then tasked with predicting the latent representations of the target blocks, conditioned on their positional embeddings p_j:

$$\hat{s}_{r,j} = g_\psi(z^{(c)}_r, p_j), \quad \forall j \in M_{t,k}$$

The predictive loss for the real modality minimizes the squared ℓ_2 distance between the predictor's output and the target encoder's actual representations $s_{r,j} = f^r_\theta(x_{r,j})$:

$$\mathcal{L}^r_{JEPA} = \frac{1}{K} \sum_{k=1}^{K} \sum_{j \in M_{t,k}} \left\| g_\psi(f^r_\phi(x^{(c)}_r), p_j) - f^r_\theta(x_{r,j}) \right\|_2^2$$

Crucially, following the standard JEPA framework, the target encoder parameters θ are not updated via gradient descent. Instead, they are updated via an exponential moving average (EMA) of the context encoder parameters ϕ, such that $\theta \leftarrow \alpha\theta + (1-\alpha)\phi$, where $\alpha \in [0, 1]$ is the momentum coefficient.

An identical but mathematically decoupled parallel process is applied to the synthetic images $x_s \in \mathcal{D}_s$. The context patches $x^{(c)}_s$ are processed by the LoRA-adapted encoder f^s_ϕ, yielding the synthetic predictive loss $\mathcal{L}^s_{JEPA}$. The total predictive objective is thus the expected sum of the modality-specific losses:

$$\mathcal{L}_{pred} = \mathbb{E}_{x_r \sim \mathcal{D}_r}[\mathcal{L}^r_{JEPA}] + \mathbb{E}_{x_s \sim \mathcal{D}_s}[\mathcal{L}^s_{JEPA}]$$

By decoupling the patch encoding and prediction steps through these low-rank adaptations, SJEPA prevents the shared predictor g_ψ from overfitting to the generative artifacts of $\mathcal{D}_s$. To formalize the necessity of this decoupled architecture, we show that a shared encoder leads to destructive gradient interference.

Proposition 1 (Gradient Interference Avoidance). Let W be a shared weight matrix optimized jointly on real and synthetic predictive losses, $\mathcal{L}^r_{JEPA}$ and $\mathcal{L}^s_{JEPA}$. If the synthetic data contains high-frequency generative artifacts ϵ_s, the gradients $g_r = \nabla_W \mathcal{L}^r_{JEPA}$ and $g_s = \nabla_W \mathcal{L}^s_{JEPA}$ can destructively interfere, i.e., $\langle g_r, g_s \rangle < 0$. By decoupling the parameter updates via LoRA, SJEPA guarantees monotonic minimization of the real objective.

Proof. Assume the synthetic distribution can be modeled as $x_s = x_r + \epsilon_s$. In a fully shared architecture, the parameter update is $\Delta W \propto -(g_r + g_s)$. If the variance of the artifacts ϵ_s is high, the predictive loss $\mathcal{L}^s_{JEPA}$ forces g_s to prioritize minimizing the prediction error of these artifacts, leading to a gradient direction that diverges from g_r. The change in the real loss is approximated by a first-order Taylor expansion:

$$\Delta \mathcal{L}^r_{JEPA} \approx \langle g_r, \Delta W \rangle \propto -\|g_r\|_2^2 - \langle g_r, g_s \rangle$$

If the interference is severe such that $\langle g_r, g_s \rangle < -\|g_r\|_2^2$, then $\Delta \mathcal{L}^r_{JEPA} > 0$, causing the real representations to degrade and collapse. By introducing the LoRA-adapted encoder $f^s_\phi(x; W_0, A, B)$ and freezing W_0 for synthetic updates, we isolate the optimization trajectories. The real updates solely modify W_0, yielding $\Delta W_0 \propto -g_r$, which guarantees $\Delta \mathcal{L}^r_{JEPA} \propto -\|g_r\|_2^2 \leq 0$. Meanwhile, the synthetic-specific gradients are entirely absorbed by the low-rank matrices: $\Delta A, \Delta B \propto -\nabla_{A,B} \mathcal{L}^s_{JEPA}$. This decoupled optimization actively prevents synthetic artifact interference on the real manifold. ■

3.3. Cross-Modality Alignment Objective

While the dual-modality predictive architecture successfully prevents the predictor network from collapsing into generative artifacts, the isolated training of modality-specific encoders introduces a new challenge: latent space divergence. Because the synthetic encoder f^s_ϕ and the real encoder f^r_ϕ are updated independently via their respective predictive losses, their learned manifolds may drift apart. To ensure that the representations learned from the synthetic distribution genuinely enhance the model's understanding of real-world semantics, we must explicitly harmonize these two latent spaces.

To achieve this, we introduce a Cross-Modality Alignment Objective that anchors the synthetic representations strictly to the real manifold. We construct training batches consisting of B semantically paired images, denoted as $\{(x^{(i)}_r, x^{(i)}_s)\}_{i=1}^B$. These pairs can be formed by matching a real image with a synthetic image generated using the same class label or textual caption.

For the alignment step, we process the unmasked full images through their respective target encoders to extract global semantic features. Given that the target encoders output a sequence of N patch embeddings, we apply Global Average Pooling (GAP) to aggregate these local features into a single, compact global representation vector for each image:

$$h^r_i = \frac{1}{N} \sum_{j=1}^{N} f^r_\theta(x^{(i)}_{r,j})$$

$$h^s_i = \frac{1}{N} \sum_{j=1}^{N} f^s_\theta(x^{(i)}_{s,j})$$

Here, $h^r_i \in \mathbb{R}^d$ and $h^s_i \in \mathbb{R}^d$ denote the global representations for the i-th real and synthetic image, respectively.

To bridge the modality gap, we enforce structural similarity between these embeddings using an InfoNCE-style contrastive loss. By treating the paired real and synthetic images as positive samples and all other real images in the batch as negative samples, the alignment loss pulls the synthetic embedding h^s_i toward its semantic real counterpart h^r_i, while pushing it away from unrelated real embeddings h^r_j:

$$\mathcal{L}_{align} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\text{sim}(h^s_i, h^r_i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(h^s_i, h^r_j)/\tau)}$$

where $\text{sim}(u, v) = \frac{u^\top v}{\|u\|_2 \|v\|_2}$ represents the cosine similarity function, and τ is a temperature hyperparameter that controls the sharpness of the distribution.

The overall training objective of SJEPA is formulated as a weighted combination of the decoupled predictive losses and the cross-modality alignment loss:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda \mathcal{L}_{align}$$

where λ is a scaling coefficient. We mathematically justify this alignment objective by demonstrating that it bounds the mutual information between the two modalities.

Proposition 2 (Mutual Information Maximization). Minimizing the cross-modality alignment loss $\mathcal{L}_{align}$ explicitly maximizes a lower bound on the mutual information $I(H^r; H^s)$ between the real and synthetic latent spaces, preventing latent space divergence.

Proof. Let the optimal similarity critic function be $f(h^r, h^s) = \frac{1}{\tau}\text{sim}(h^r, h^s)$. Following the InfoNCE formulation, we assume the critic optimally models the density ratio $f(h^r_i, h^s_i) \propto \log \frac{p(h^r_i, h^s_i)}{p(h^r_i)p(h^s_i)}$. The expected alignment loss over a batch of size B is:

$$\mathbb{E}[\mathcal{L}_{align}] = \mathbb{E}\left[-\log \frac{\exp(f(h^r_i, h^s_i))}{\frac{1}{B}\sum_{j=1}^{B} \exp(f(h^r_j, h^s_i))} \right]$$

As the batch size $B \to \infty$, the denominator converges to the expectation over the marginal distribution:

$$\lim_{B\to\infty} \frac{1}{B} \sum_{j=1}^{B} \exp(f(h^r_j, h^s_i)) = \mathbb{E}_{h^r_j \sim p(h^r)}[\exp(f(h^r_j, h^s_i))]$$

Table 1. **Image captioning.** We perform fine-tuning on pre-trained ClipCap, IFCap, LLaVA, and LLaMA-3 for image captioning on COCO. We report the standard metrics to evaluate the quality of generated captions. The best results are indicated in **bold**.

Method	B@4 (↑)	METEOR(↑)	CIDEr (↑)	SPICE (↑)	ROUGE-L (↑)	WMD (↑)
ClipCap [19]	32.15	27.10	108.35	20.12	–	–
ClipCap + GMAIL [18]	38.12	31.67	119.53	23.75	56.27	62.16
ClipCap + SJEPA (ours)	**40.25**	**34.21**	**123.62**	**26.32**	**58.76**	**65.38**
IFCap [12]	33.25	28.60	115.27	21.58	51.35	56.72
IFCap + GAMIL [18]	39.32	32.07	127.86	23.98	59.83	63.51
IFCap + SJEPA (ours)	**42.12**	**34.56**	**130.57**	**26.23**	**62.25**	**65.83**
LLaVA [14]	39.67	32.38	134.29	24.17	61.36	65.78
LLaVA + GMAIL [18]	43.26	34.89	146.38	27.23	65.25	71.39
LLaVA + SJEPA (ours)	**46.21**	**36.23**	**152.25**	**28.56**	**67.65**	**73.78**
Llama3 [17]	47.36	35.21	158.13	28.35	68.32	75.13
Llama3 + GMAIL [18]	50.21	38.59	168.53	32.58	73.29	80.25
Llama3 + SJEPA (ours)	**55.67**	**42.15**	**185.36**	**38.76**	**78.65**	**85.16**

Table 2. **Zero-shot image retrieval on COCO.** We perform zero-shot retrieval on pre-trained SemiCLIP for image retrieval on the COCO benchmark. We report the image-to-text and text-to-image Recall@1,5,10 metrics to evaluate the quality of retrieved images.

Method	Image-to-Text			Text-to-Image		
	R@1 (↑)	R@5 (↑)	R@10 (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)
CLIP [21]	51.8	76.8	84.3	32.7	57.7	68.2
CLIP + GMAIL [18]	56.8	80.1	87.2	37.5	62.7	73.2
CLIP + SJEPA (ours)	**59.2**	**83.3**	**89.1**	**39.8**	**65.1**	**75.3**
Long-CLIP [33]	57.2	80.8	87.8	40.4	65.9	75.7
Long-CLIP + GMAIL [18]	62.3	84.1	91.2	45.6	69.8	79.5
Long-CLIP + SJEPA (ours)	**67.8**	**87.9**	**95.3**	**49.8**	**73.9**	**83.6**

Substituting the optimal density ratio into the expectation, we get:

$$\mathbb{E}_{h_j^r}\left[\frac{p(h_j^r, h_i^s)}{p(h_j^r)p(h_i^s)}\right] = \int p(h^r)\frac{p(h^r, h_i^s)}{p(h^r)p(h_i^s)}dh^r$$

$$= \frac{\int p(h^r, h_i^s)dh^r}{p(h_i^s)} = 1$$

Thus, the expected loss becomes:

$$\mathbb{E}[\mathcal{L}_{align}] = -\mathbb{E}_{p(h^r,h^s)}\left[\log\frac{p(h^r, h^s)}{p(h^r)p(h^s)}\right] + \log B$$

By definition, the first term is the negative mutual information $-I(H^r; H^s)$. Rearranging the inequality yields the bound:

$$I(H^r; H^s) \geq \log B - \mathcal{L}_{align}$$

Therefore, minimizing $\mathcal{L}_{align}$ strictly pushes the lower bound of the mutual information upward. This guarantees that the synthetic representations H^s encode maximal information about the real representations H^r, effectively preventing the two manifolds from diverging during the decoupled predictive training. ■

4. Experiments

In this section, we comprehensively evaluate the proposed Synthetic Joint Embedding Predictive Architecture (SJEPA). We detail our experimental setup, datasets, and implementation specifics. Subsequently, we compare SJEPA against state-of-the-art baselines across a diverse suite of vision-language and representation learning tasks, including image captioning, zero-shot image retrieval, long caption retrieval, and zero-shot image classification.

4.1. Experimental Setup

Datasets. Our experiments utilize a diverse collection of benchmarks to rigorously test SJEPA's generalization and cross-modal alignment on CC3M [24] and CC12M [4], following [18]. For image captioning and zero-shot image retrieval, we use the widely adopted COCO [13] and Flickr30k [31] datasets. To evaluate long-context comprehension, we employ the ShareGPT4V [5] dataset, which contains highly descriptive and complex captions. For zero-shot image classification, we follow standard evaluation protocols and test on eight diverse benchmarks: DTD [6], Stanford Cars [11], SUN397 [29, 30], Food 101 [3], Aircraft [16], Oxford Pets [20], Caltech 101 [8], and ImageNet 1K [7]. To construct our synthetic modality ($\mathcal{D}_s$), we leverage Stable Diffusion v2 to generate synthetic images conditioned on the textual captions from COCO train2014 split.

Evaluation Metrics. We employ task-specific standard metrics. For image captioning, we report BLEU@4 (B@4), METEOR, CIDEr, SPICE, ROUGE-L, and Word Mover's Distance (WMD) to assess semantic accuracy. For zero-shot image retrieval (both image-to-text and text-to-image), we evaluate using Recall@1, Recall@5, and Recall@10. For zero-shot image classification, we report the top-1 accuracy to measure the robustness of the learned representations on unseen categories.

Table 3. **Zero-shot image retrieval on Flickr30k.** We perform zero-shot retrieval on pre-trained SemiCLIP for image retrieval on the Flickr30k benchmark. We report the image-to-text and text-to-image Recall@1,5,10 metrics to evaluate the quality of retrieved images.

Method	Image-to-Text			Text-to-Image		
	R@1 (↑)	R@5 (↑)	R@10 (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)
CLIP [21]	44.1	68.2	77.0	24.7	45.1	54.6
CLIP + GMAIL [18]	47.1	71.2	79.6	30.2	50.3	60.5
CLIP + SJEPA (ours)	**48.3**	**73.1**	**81.5**	**33.8**	**56.2**	**65.3**
Long-CLIP [33]	47.2	71.5	80.0	33.1	55.6	64.9
Long-CLIP + GMAIL [18]	51.6	75.3	83.6	39.3	61.5	71.8
Long-CLIP + SJEPA (ours)	**55.6**	**78.5**	**88.3**	**45.2**	**66.7**	**76.2**

Table 4. **Long caption retrieval on ShareGPT4V.** We report the image-to-text and text-to-image Recall@1 to evaluate the quality of retrieved images. The best results are indicated in **bold**.

Method	Image-to-Text	Text-to-Image
CLIP [21]	78.2	79.6
CLIP + GMAIL [18]	85.2	86.7
CLIP + SJEPA (ours)	**87.5**	**88.6**
Long-CLIP [33]	94.6	93.3
Long-CLIP + GMAIL [18]	97.2	96.1
Long-CLIP + SJEPA (ours)	**98.3**	**97.5**

Implementation. The base architecture for our framework is a Vision Transformer (ViT) optimized using the I-JEPA masking strategy. The real encoders (f^r_ϕ, f^r_θ) initialize from pre-trained weights, while the synthetic encoders incorporate Low-Rank Adaptation (LoRA) modules with a rank of $r = 4$ to maintain computational efficiency. We optimize the model using AdamW with a base learning rate of 1×10^{-4} and a weight decay of 0.01, modulated by a cosine annealing schedule. The target encoder is updated via an exponential moving average (EMA) with a momentum coefficient α starting at 0.996. The Cross-Modality Alignment Objective utilizes an InfoNCE temperature of $\tau = 0.07$.

4.2. Comparison to prior work

Image Captioning. We compare SJEPA's performance on the COCO dataset against strong vision-language baselines, including ClipCap, IFCap, LLaVA, and Llama3, as well as the recent GMAIL alignment framework. As shown in Table 1, SJEPA achieves state-of-the-art results across all architectures and metrics. Notably, when integrated with Llama3, SJEPA achieves a striking 185.36 CIDEr and 55.67 B@4, outperforming the Llama3 + GMAIL baseline by 16.83 CIDEr and 5.46 B@4. These substantial gains validate that our decoupled predictive architecture successfully extracts high-level semantics from synthetic data without overfitting to generative noise, leading to richer, more accurate textual generations.

Zero-Shot Image Retrieval. We evaluate cross-modal alignment via zero-shot retrieval on both the COCO (Table 2) and Flickr30k (Table 3) benchmarks. Compared to the base CLIP and Long-CLIP models, SJEPA significantly boosts recall rates across the board. On COCO, Long-CLIP + SJEPA yields an impressive 67.8 R@1 for Image-to-Text and 49.8 R@1 for Text-to-Image, completely outperforming the prior GMAIL alignment strategy. By ensuring that our InfoNCE alignment loss maximizes the mutual information between the synthetic and real manifolds, SJEPA provides a vastly superior shared latent space for retrieval tasks.

Long Caption Retrieval. In Table 4, we test SJEPA on the ShareGPT4V benchmark, assessing its capability to associate images with highly detailed, long-form textual descriptions. SJEPA pushes the performance of Long-CLIP to 98.3 R@1 for Image-to-Text and 97.5 R@1 for Text-to-Image. This confirms that SJEPA's synthetic integration does not dilute complex contextual relationships, but rather reinforces the model's capacity to handle intricate descriptive prompts.

Zero-Shot Image Classification. Finally, to assess the pure visual representation quality, we evaluate zero-shot classification across eight challenging benchmarks (Table 5). Integrating SJEPA with SynCLR pushes the boundaries of synthetic data utility, consistently outperforming standard configurations. For instance, SynCLR + SJEPA achieves 89.85 on Aircraft (a +3.10 gain over SynCLR + GMAIL) and 87.25 on DTD (a +3.58 gain over SynCLR + GMAIL). These results decisively demonstrate that our dual-modality architecture prevents the corruption of fine-grained real-world visual features, allowing the model to leverage synthetic data to achieve robust classification generalizability.

4.3. Experimental Analysis

In this section, we conduct extensive ablation studies to isolate the contributions of our proposed components. We also analyze the scaling trends of SJEPA across varying sizes of synthetic training data.

Effect of Cross-Modality Alignment. To quantify the impact of our Cross-Modality Alignment Objective ($\mathcal{L}_{align}$), we compare SJEPA against a baseline configuration where the alignment loss is ablated ($\lambda = 0$). In this ablated setup, the real and synthetic encoders are trained exclusively on their respective predictive losses ($\mathcal{L}^r_{JEPA}$ and $\mathcal{L}^s_{JEPA}$). As shown in Table 6, removing the alignment objective results in a noticeable performance drop across all image captioning metrics on COCO. Specifically, without explicit harmonization, the CIDEr score drops from 123.62 to 118.45, and

Table 5. **Zero-shot image classification.** We perform a zero-shot evaluation on pre-trained image encoder for image classification on eight benchmarks. We report the top-1 accuracy to evaluate the quality of learned representations. The best results are indicated in **bold**.

Method	DTD	Stanford Cars	SUN397	Food 101	Aircraft	Oxford Pets	Caltech 101	ImageNet (ft)
CLIP [21]	55.20	77.53	69.31	93.08	32.88	93.33	93.24	75.54
CLIP + GMAIL [18]	65.26	81.32	75.53	95.21	37.85	95.23	95.57	77.68
CLIP + SJEPA (ours)	**67.37**	**83.56**	**77.82**	**96.58**	**38.93**	**96.87**	**97.25**	**79.35**
SynCLR [27]	79.90	93.80	76.20	91.60	81.70	93.60	95.30	85.80 (ft)
SynCLR + GMAIL [18]	83.67	96.56	81.25	96.38	86.75	95.70	98.35	87.95
SynCLR + SJEPA (ours)	**87.25**	**98.37**	**85.78**	**98.23**	**89.85**	**98.38**	**99.26**	**89.78**

Table 6. **Ablation on Cross-Modality Alignment.** We evaluate the impact of the alignment objective on COCO image captioning using the ClipCap base architecture.

Configuration	B@4 (↑)	METEOR (↑)	CIDEr (↑)	SPICE (↑)
SJEPA w/o $\mathcal{L}_{align}$	37.80	31.05	118.45	23.10
SJEPA (Full)	**40.25**	**34.21**	**123.62**	**26.32**

Table 7. **Scaling trend of SJEPA.** We report zero-shot retrieval on Flickr30k using models trained on synthetic data derived from COCO, CC3M, and CC12M.

Train Data	Image-to-Text			Text-to-Image		
	R@1 (↑)	R@5 (↑)	R@10 (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)
COCO	48.3	73.1	81.5	33.8	56.2	65.3
CC3M	50.8	76.4	84.8	36.5	59.1	67.8
CC12M	**53.4**	**78.9**	**87.2**	**38.6**	**61.5**	**70.4**

B@4 drops from 40.25 to 37.80. This empirically validates our theoretical proposition: while the dual-modality architecture prevents predictive mode collapse, the latent spaces of the two modalities will naturally drift apart without the InfoNCE-based alignment, reducing the utility of synthetic features for real-world downstream tasks.

Scaling Trend of Synthetic Data. One of the primary motivations for SJEPA is to unlock the limitless scale of synthetic data. To evaluate our framework's scalability, we generated synthetic training sets corresponding to the captions of COCO, CC3M, and CC12M. We trained SJEPA on these increasingly large datasets and evaluated the zero-shot image retrieval performance on the Flickr30k benchmark. The results in Table 7 reveal a strong, positive scaling trend. As the volume of synthetic data increases from COCO to CC12M, SJEPA's Image-to-Text R@1 improves from 48.3 to 53.4, and Text-to-Image R@1 improves from 33.8 to 38.6. This confirms that SJEPA does not saturate or collapse when exposed to massive amounts of generative data, but instead successfully leverages the expanded diversity to build richer, more generalizable representations.

Ablation on LoRA Rank and Efficiency. To adapt the real base encoder to the synthetic modality without catastrophic forgetting, SJEPA employs Low-Rank Adaptation (LoRA). We conduct an ablation study to determine the optimal rank r for the trainable matrices A and B. As detailed in Table 8, increasing the rank from $r = 2$ to $r = 4$ yields a notable improvement in captioning performance (CIDEr increases from 120.45 to 123.62). However, further increasing the rank to $r = 6$ or employing full fine-tuning of the synthetic encoder leads to marginal degradations. This indicates that a strictly constrained low-rank bottleneck ($r = 4$) is optimal; it provides enough capacity to model the modality shift while aggressively filtering out the high-frequency generative noise that would otherwise cause overfitting. Furthermore, utilizing LoRA updates significantly fewer parameters compared to full fine-tuning, preserving computational efficiency during the dual-modality training phase.

Table 8. **Ablation on LoRA rank.** We compare different LoRA ranks and full fine-tuning on the synthetic branch for COCO image captioning.

Method	B@4 (↑)	CIDEr (↑)	SPICE (↑)
LoRA ($r = 2$)	38.45	120.45	24.12
LoRA ($r = 4$)	**40.25**	**123.62**	**26.32**
LoRA ($r = 6$)	39.90	122.80	25.85
Full fine-tuning	39.10	121.55	25.10

5. Conclusion

In this work, we present SJEPA, a novel framework designed to address the profound modality gap between generated and real images. To prevent predictive models from exploiting high-frequency generative artifacts as learning shortcuts, SJEPA explicitly treats synthetic data as a distinct modality, processing it through computationally efficient, decoupled low-rank adapters. Furthermore, we proposed a cross-modality alignment objective that harmonizes the representations of semantically paired real and synthetic images within a shared latent space. By anchoring synthetic features to the real-world manifold, SJEPA ensures that the extracted representations are robust, semantic, and highly generalizable. Extensive experiments demonstrate the efficacy of our approach across a wide array of vision-language and representation learning tasks. Our framework significantly improves performance in zero-shot image classification, zero-shot image retrieval, and image captioning, establishing notable enhancements when integrated with advanced multimodal architectures like LLaVA and LLaMA-3. Crucially, the positive data scaling trends observed with massive synthetic datasets, such as CC12M, underscore the robustness and scalability of SJEPA. By effectively mitigating mode collapse and modality divergence, SJEPA paves the way for the limitless, cost-effective use of generative data in training robust, general-purpose vision models.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 1, 2

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Jiahui Zhuang, Joyce Lee, Yu Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, and A. Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 6

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 6

[6] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6

[8] 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 *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004. 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 1, 3

[11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 6

[12] Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. Ifcap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning. *arXiv preprint arXiv:2409.18046*, 2024. 6

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, 2023. 6

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 3

[16] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[17] AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*, 2024. 6

[18] Shentong Mo and Sukmin Yun. GMAIL: Generative modality alignment for generated image learning. In *Forty-second International Conference on Machine Learning (ICML)*, 2025. 6, 7, 8

[19] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 6

[20] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 6, 7, 8

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021. 3

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3

[24] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[25] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631: 755–759, 2024. 1, 3

[26] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Ananya Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[27] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15887–15898, 2024. 8

[28] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15887–15898, 2024. 1, 3

[29] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 6

[30] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119:3–22, 2014. 6

[31] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[32] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325, 2024. 3

[33] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024. 6, 7