GenView++: Unifying Adaptive View Generation and Quality-Driven Supervision for Contrastive Representation Learning

Xiaojie Li, Bei Wang, Jianlong Wu, *Member, IEEE,* Yue Yu, *Member, IEEE,*
Liqiang Nie, *Senior Member, IEEE,* Min Zhang

***Abstract*—The success of contrastive learning depends on the construction and utilization of high-quality positive pairs. However, current methods face critical limitations on two fronts: on the construction side, both handcrafted and generative augmentations often suffer from limited diversity and risk semantic corruption; on the learning side, the absence of a quality assessment mechanism leads to suboptimal supervision where all pairs are treated equally. To tackle these challenges, we propose GenView++, a unified framework that addresses both fronts by introducing two synergistic innovations. To improve pair construction, GenView++ introduces a multi-source adaptive view generation mechanism to synthesize diverse yet semantically coherent views by dynamically modulating generative parameters across image-conditioned, text-conditioned, and image-text-conditioned strategies. Second, a quality-driven contrastive learning mechanism assesses each pair's semantic alignment and diversity to dynamically reweight their training contribution, prioritizing high-quality pairs while suppressing redundant or misaligned pairs. Extensive experiments demonstrate the effectiveness of GenView++ across both vision and vision–language tasks. For vision representation learning, it improves MoCov2 by +2.5% on ImageNet linear classification. For vision–language learning, it raises the average zero-shot classification accuracy by +12.31% over CLIP and +5.31% over SLIP across ten datasets, and further improves Flickr30k text retrieval R@5 by +3.2%. The code is available at https://github.com/xiaojieli0903/GenViewPlusPlus.**

***Index Terms*—Contrastive Learning, Generative Augmentation, Generative Models, Vision Representation Learning, Vision-Language Representation Learning, Quality-Driven Supervision.**

I. Introduction

SELF-SUPERVISED learning (SSL) has emerged as a fundamental paradigm for acquiring robust and generalizable representations from large-scale unlabeled data. Among various SSL methods, contrastive learning (CL) [1], [2] has shown remarkable success in both *vision* [3], [4], [5] and *vision-language* [6], [7] representation learning. The core principle of CL is to learn invariant and discriminative representations by maximizing the similarity between different views of the same instance (positive pairs) while pushing them apart from other instances. Its effectiveness crucially depends on constructing and utilizing *high-quality positive pairs* that navigate a critical trade-off: they must be semantically consistent to ensure true invariance is learned, yet sufficiently diverse to foster robust generalization to unseen variations.

However, current methods often employ handcrafted augmentations (e.g., random cropping, color jittering) on the same instance to obtain positive views. These transformations offer *limited diversity*, as they only modify surface-level visual characteristics and fail to introduce new content or high-level variations (e.g., object viewpoints, textures, or variations within a semantic category), often yielding redundant views (Fig. 1 a-1, a-4). Overly aggressive augmentations are not always precise, potentially causing *false positive pairs* when essential content is cropped out (a-2, a-3). These issues are further exacerbated in vision–language pretraining: redundant views provide limited benefit for visual–text grounding, while distorted views cause semantic mismatches that disrupt image–caption alignment. The limitations of handcrafted augmentations highlight generative models such as Stable Diffusion [8] as promising alternatives, as their abundant prior knowledge learned from large-scale datasets enables the synthesis of high-quality, diverse images that enrich view contents. However, integrating pretrained generative models into contrastive learning is non-trivial. Despite their strong generative capability, these models *lack controllability over generation contents*, making it difficult to balance semantic fidelity with diversity and increasing the risk of producing views that deviate from the conditional inputs, thus creating false positive pairs. Moreover, most existing methods *rely on a single conditioning source* [9], [10], [11], conditioning on either the image or the text, but not both. This unimodal approach overlooks the powerful synergistic information in image-text pairs, missing the opportunity to synthesize views that are simultaneously consistent with the visual content and aligned with textual semantics.

Compounding all these **construction side** issues is a more fundamental limitation on the **learning side**: *the lack of pair-level quality assessment*. Current pipelines treat every training pair as equally valuable, assigning a uniform importance score regardless of its semantic alignment or diversity (as illustrated by the gray scores in Fig.1 (a)). This indiscriminate supervision forces the model to learn from a mix of high-quality, redundant, and false positive pairs from both aggressive augmentations and noisy web-crawled data [12], [13], ultimately limiting the quality of the final representations.

Xiaojie Li, Bei Wang, Jianlong Wu, Liqiang Nie, and Min Zhang are with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), 518055, China. E-mail: xiaojieli0903@gmail.com, wangbei010118@gmail.com, wujianlong@hit.edu.cn, nieliqiang@gmail.com, zhangmin2021@hit.edu.cn.

Xiaojie Li, Jianlong Wu and Yue Yu are with the Pengcheng Laboratory, Shenzhen, China. E-mail: yuy@pcl.ac.cn.

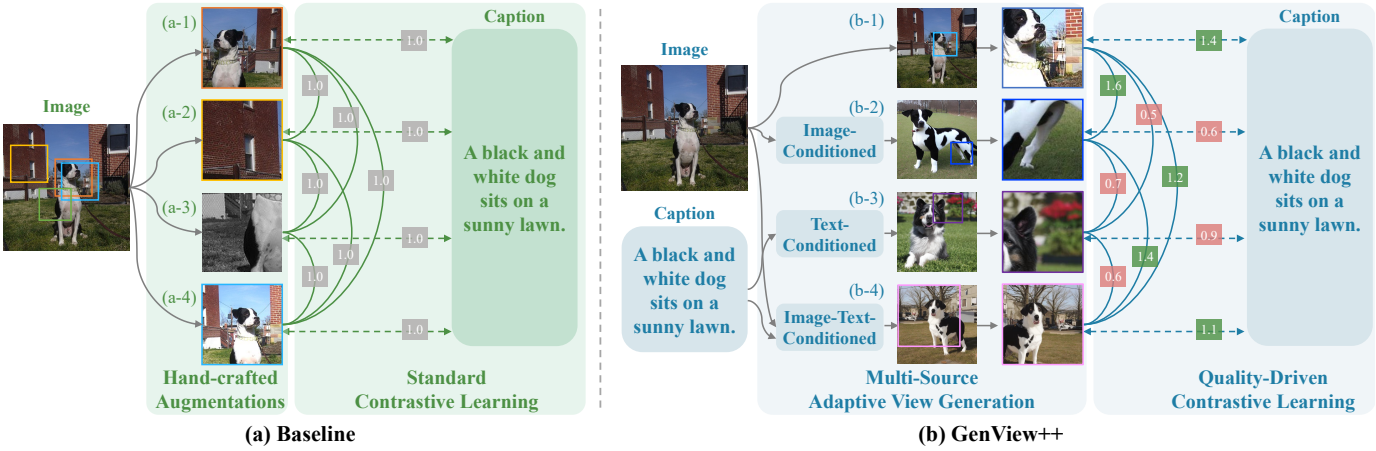


Fig. 1: **Motivation of GenView++. (a)** Standard contrastive learning methods rely on handcrafted augmentations, which provide limited diversity, risk semantic distortions, and lack pair-level quality control during training. **(b)** GenView++ addresses these limitations via (i) **Multi-Source Adaptive View Generation**, which synthesizes diverse and semantically aligned views, and (ii) **Quality-Driven Contrastive Learning**, which reweights supervision by assessing the quality of both image–image and image–text pairs.

To tackle these fundamental challenges in both construction and learning sides, we propose **GenView**, a controllable framework that enhances view quality for vision representation learning using the powerful pretrained generative model, and guides contrastive learning via quality assessment. Specifically, given an image as the source view, we construct its positive view using the synthetic image sampled from a pretrained generative model conditioned on this image. To optimally balance the trade-off between diversity and semantic fidelity, we design an **adaptive view generation** strategy to generate diverse yet semantically consistent views by adaptively perturbing the image embedding according to foreground saliency. Low-saliency subjects receive weaker perturbations to ensure the correct semantic content of the synthetic image, while high-saliency subjects receive stronger perturbations to create more variations for more diverse content and environments. Furthermore, we develop a **quality-driven contrastive learning** mechanism to mitigate the effect of potential false positive pairs that could mislead contrastive learning. We score each image-image pair considering both foreground similarity and background diversity. It prioritizes the positive pairs with high foreground similarity to ensure semantic coherence, while also favoring the pairs with low background similarity to promote diverse environments for learning invariant representations. We then recalibrate the contrastive loss function by reweighting each pair with its pair quality, which enhances the contributions of high-quality positive pairs, and simultaneously reduces the influence of low-quality and even false pairs.

Building upon our prior work, GenView [14], published in ECCV 2024 [14], which first introduced the principles of adaptive generation and quality-driven supervision for vision-only representation learning, we propose **GenView++**, a substantial extension that generalizes these principles into a unified framework for both vision and vision–language representation learning through two synergistic innovations.

First, to tackle the pair construction challenges such as limited diversity, lack of controllability, and single-modality conditioning, GenView++ introduces a **multi-source adaptive view generation** module. This offline mechanism leverages pretrained diffusion models to synthesize a rich spectrum of diverse yet semantically faithful views. It uniquely integrates three complementary and adaptively controlled conditioning strategies: **(1) Image-Conditioned** mode (Fig. 1 (b-2), inherited from GenView, enhances visual invariance by introducing saliency-aware variations that preserve the core semantics of the original image. **(2) Text-Conditioned** mode (Fig. 1 (b-3)) leverages captions to guide synthesis, adaptively tuning the guidance strength based on caption complexity: simple captions trigger higher guidance strength to maintain strict alignment, while complex captions allow relaxed guidance to foster creative transformations. **(3) Image–Text-Conditioned** mode (Fig. 1 (b-4)) leverages both image and text input for view generation, synergistically adapting both image noise and text guidance to produce views that are visually diverse, semantically consistent, and textually grounded. Collectively, these strategies form a unified and controllable generation mechanism that benefits both intra-modal (image–image) and cross-modal (image–text) contrastive learning. For vision-only pretraining (solid blue lines in Fig. 1 (b), the generated views enhance the robustness and generalization capability of visual encoders. For vision–language pretraining (dashed blue lines), they help bridge the modality gap by pairing captions with semantically aligned and diverse synthetic images.

Second, to address the lack of pair-level quality assessment on the learning side, we introduce an extensible **quality-driven contrastive learning** mechanism. Unlike the conventional "augment-and-trust" paradigm, it explicitly evaluates and reweights both unimodal (image–image) and, for the first time, cross-modal (image–text) pairs. For **image-image pairs**, we retain the saliency-based scoring from GenView, rewarding pairs with high foreground consistency and high background diversity to encourage object-level invariance. For **image-text pairs**, we score them by combining *cross-modal semantic alignment*, measured via CLIP similarity between a view

and its caption, which rewards views that accurately preserve the semantic elements described in the text while penalizing mismatches; and *visual diversity*, which quantifies background novelty relative to the original image, encouraging informative variations while reducing redundancy. The combined scores dynamically reweight the contrastive loss, amplifying the influence of high-quality informative pairs (green scores in Fig. 1 (b)) while suppressing redundant or misaligned ones (red scores). This ensures that the powerful views created by our generative module are utilized with maximum effectiveness, leading to more robust and fine-grained representations.

Our contributions are as follows:

- We propose GenView++, a unified and controllable framework that extends our prior vision-only method, GenView, to simultaneously enhance both visual and vision-language representation learning.
- We introduce a multi-source adaptive view generation module with three adaptive strategies (Image-Conditioned, Text-Conditioned, and Image-Text-Conditioned). This module adaptively modulates generative behavior based on input characteristics, enabling the synthesis of semantically consistent yet diverse positive pairs.
- We introduce a quality-driven contrastive loss applicable to both image–image and image–text pairs. By evaluating semantic alignment and diversity, it dynamically re-weights each training pair to prioritize high-quality pairs while suppressing redundant or misaligned ones.
- We conduct extensive experiments demonstrating that GenView++ consistently enhances mainstream contrastive methods: for vision tasks, it improves baselines such as MoCov2, SimSiam, and BYOL; for vision-language tasks, it surpasses state-of-the-art methods like CLIP and StableRep on tasks like cross-modal retrieval and zero-shot classification.

II. Related Work

A. Contrastive Representation Learning

1) ***Vision Representation Learning****:* Self-supervised learning aims to learn visual representations from unlabeled data. Early works relied on pretext tasks such as solving jigsaws [15], predicting rotations [16], or auto-encoding [17], [18]. More recent approaches fall into two dominant paradigms: *masked image modeling* (MIM) [5] and *contrastive learning* (CL) [1], [3], [19], [20], [21], [22], [23], [24]. MIM reconstructs missing pixels to enforce holistic understanding, while CL learns invariant features through an instance discrimination principle: they learn invariant representations by pulling different augmented views of the same image (positive pairs) together, while pushing views from different images (negative pairs) apart. This principle has been further refined by non-contrastive approaches [2], [25], [26], [27], [28], [4], [29], [30], which eliminate the need for explicit negative pairs through mechanisms like asymmetric encoders or stop-gradients to prevent representational collapse. CL has shown strong generalization ability but remains constrained by the quality of positive pairs, which are typically built from handcrafted augmentations that offer limited diversity and may cause semantic distortions.

2) ***Vision-Language Representation Learning****:* Contrastive pretraining has also driven breakthroughs in multimodal learning, with models such as CLIP [6], ALIGN [31], and BLIP [7] achieving state-of-the-art results across image–text tasks. These frameworks are trained on large-scale web-crawled datasets composed of image–caption pairs. Their performance, however, is highly sensitive to data quality: web-crawled datasets like CC12M [32], and LAION [13] often contain noisy or weakly aligned pairs. Moreover, naive augmentations often fail to enrich diversity without breaking cross-modal alignment [33], [34]. As a result, constructing positive pairs that are both semantically aligned and diverse remains a key challenge in vision and vision–language learning.

B. Positive Pair Construction

The critical role of positive pairs has spurred extensive research into moving beyond simple handcrafted augmentations. These advanced efforts can be broadly categorized into two main approaches: recomposition-based and generation-based.

1) ***Recomposition-based Augmentation****:* These methods transform or recombine existing data to form positive views. *In vision-only learning*, techniques include task-aware transformations [35], saliency-preserving augmentations [36], clustering [25], [2], or retrieval of near-duplicates [37]. In the *vision–language domain*, approaches include performing semantic-preserving perturbations [38], [39] or matching samples from image or text [40]. However, these methods are fundamentally bounded by the semantic space of the original dataset: they cannot create novel visual concepts such as unseen poses, backgrounds, or styles. This constraint is especially limiting in vision–language learning, which demands varied visual contexts to establish robust cross-modal alignment.

2) ***Generation-based Augmentation****:* The advent of powerful generative models has unlocked a new paradigm for data augmentation: synthesizing entirely new views. Early efforts in this direction involved training generative models (e.g., GANs, DMs) from scratch on the target dataset itself [41], [42], [43], [44], [45]. However, this approach is inherently self-limiting. Since the generator is confined to the original data distribution, it can only interpolate within the seen semantics and cannot introduce truly novel, high-level variations like unseen object poses or backgrounds. A more powerful and now dominant approach is to *leverage large-scale pretrained diffusion models* like Stable Diffusion [8], which are trained on massive web datasets [13]. By injecting the vast visual knowledge encoded within these models, this strategy can synthesize photorealistic views with rich semantic variations far beyond the source data's scope, showing great promise for representation learning [11], [46], [47], [48], [49], [50], [51], [52], [53]. Despite this promise, a critical challenge persists: *a lack of fine-grained, input-aware controllability*. Existing methods govern the synthesis process using either fixed hyperparameters for all inputs [11], [46] or randomly sampled ones [47], [49]. These strategies fail to adapt to the complexity of each input, often producing views that are either semantically distorted or redundant.

Compounding these pair construction-side challenges is a more fundamental limitation in how the resulting pairs are

utilized: the *lack of pair-level quality assessment*. Standard contrastive learning pipelines operate on an "augment-and-trust" principle, treating every positive pair, whether handcrafted, recomposed, or generated, as equally valuable. This indiscriminate scheme is unable to distinguish between a highly informative, semantically aligned pair and a redundant or even misaligned one, limiting the quality of learned representations. This problem is especially severe with noisy web-crawled data and stochastic generative outputs.

GenView++ provides a two-stage solution to these layered challenges. It first addresses the construction limitations with a multi-source adaptive generation module that synthesizes a diverse set of controllable, high-fidelity views. It then tackles the supervision limitation with a quality-driven contrastive learning mechanism that explicitly assesses and re-weights every pair. To our knowledge, GenView++ is the first framework to systematically unify offline adaptive view creation with online quality-driven supervision, establishing a more robust foundation for contrastive representation learning.

III. Preliminaries

A. Vision Representation Learning

Given an input image $\mathbf{I}_i$, positive pairs $(\mathbf{P}_{i,1}, \mathbf{P}_{i,2})$ are created via random handcrafted augmentations:

$$\mathbf{P}_{i,1} = t_1(\mathbf{I}_i), \quad \mathbf{P}_{i,2} = t_2(\mathbf{I}_i), \tag{1}$$

where $t_1(\cdot)$ and $t_2(\cdot)$ are sampled from the same or different augmentation distributions. Both views are passed through a feature encoder $f(\cdot)$ and a projection head $g(\cdot)$ to obtain the final representation: $\mathbf{v}_{i,1} = g(f(\mathbf{P}_{i,1}))$, $\mathbf{v}_{i,2} = g(f(\mathbf{P}_{i,2}))$.

Methods like SimCLR [1] and MoCo [3] adopt the Noise Contrastive Estimation (NCE) objective, which pulls together embeddings of two augmented views of the same image while pushing apart embeddings from different images. Given two encoded views $\mathbf{v}_{i,1}, \mathbf{v}_{i,2} \in \mathbb{R}^d$, the NCE loss is defined as:

$$\mathcal{L}_{i,\text{NCE}}^{\text{img-img}} = -\log \frac{\exp\left(\mathbf{v}_{i,1} \cdot \mathbf{v}_{i,2}/\tau\right)}{\sum_{k=1}^{N} \exp\left(\mathbf{v}_{i,1} \cdot \mathbf{v}_{k}/\tau\right)}, \tag{2}$$

where τ is a temperature parameter (either learnable or fixed) controlling the similarity distribution sharpness, and the inner product $\cdot$ measures similarity of features.

BYOL [4] and SimSiam [28] remove the need for negative samples by introducing a non-linear prediction head $q(\cdot)$. One view is passed through $q(\cdot)$ to obtain a prediction vector $\mathbf{p}_{i,1} = q(\mathbf{v}_{i,1})$, which is then aligned with the representation of another view using cosine similarity:

$$\mathcal{L}_{i,\text{COS}}^{\text{img-img}} = -\frac{\mathbf{p}_{i,1}}{\|\mathbf{p}_{i,1}\|} \cdot \frac{\mathbf{v}_{i,2}}{\|\mathbf{v}_{i,2}\|}. \tag{3}$$

SwAV [2] introduces an online clustering approach that aligns different views of the same image with a shared prototype. Each view is mapped to a probability distribution over prototypes $\tilde{\mathbf{v}}_{i,1}, \tilde{\mathbf{v}}_{i,2}$, and the Sinkhorn-Knopp (SK) algorithm generates targets. Training adopts a swapped prediction objective, where one view predicts the prototype assignment of the other by minimizing the KL divergence between their distributions:

$$\mathcal{L}_{i,\text{KL}}^{\text{img-img}} = -D_{\text{KL}}\left(\tilde{\mathbf{v}}_{i,1} \| SK(\tilde{\mathbf{v}}_{i,2})\right). \tag{4}$$

B. Vision-Language Representation Learning

Vision–language pretraining extends contrastive learning to the multimodal setting by jointly training an image encoder f_{img} and a text encoder f_{text} to project images and their paired captions into a shared embedding space. Given a batch of N image-text pairs $(\mathbf{I}_1, \mathbf{T}_1), \ldots, (\mathbf{I}_N, \mathbf{T}_N)$, the encoders extract visual features $\mathbf{v}_i = f_{\text{img}}(\mathbf{I}_i)$ and textual features $\mathbf{t}_j = f_{\text{text}}(\mathbf{T}_j)$, where $i, j \in 1, \ldots, N$. Matched pairs $(\mathbf{v}_i, \mathbf{t}_i)$ serve as positives, and mismatched pairs $(\mathbf{v}_i, \mathbf{t}_j)$ for $j \neq i$ act as negatives. The model is trained to maximize similarity for positive pairs while minimizing it for negatives, typically using symmetric contrastive losses: Image-to-Text loss encourages each image feature $\mathbf{v}_i$ to be most similar to its paired text feature $\mathbf{c}_i^{\text{text}}$, compared against all text features in the batch:

$$\mathcal{L}_{i,\text{i2t}}^{\text{img-text}} = -\log \frac{\exp(\cos(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{N} \exp(\cos(\mathbf{v}_i, \mathbf{t}_j)/\tau)}. \tag{5}$$

where $\cos(\cdot)$ is cosine similarity between features. Text-to-Image loss encourages text feature $\mathbf{t}_i$ to align with its paired image feature $\mathbf{v}_i$ over all image features:

$$\mathcal{L}_{i,\text{t2i}}^{\text{img-text}} = -\log \frac{\exp(\cos(\mathbf{t}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N} \exp(\cos(\mathbf{t}_i, \mathbf{v}_j)/\tau)}. \tag{6}$$

IV. The GenView++ Framework

A. Overall Architecture

GenView++ is a unified framework for enhancing contrastive learning in both vision-only and vision–language scenarios. As shown in Fig. 2, it comprises two synergistic modules: **(a) Offline Multi-Source Adaptive View Generation** module, which synthesizes diverse and semantically consistent views base on input conditioning content's characteristics, and **(b) Online Quality-Driven Contrastive Learning** module, which dynamically reweights training pairs based on their semantic alignment and diversity to optimize the supervision. The framework operates in two distinct modes: a unimodal version for vision-only pretraining, **GenView++ (V)**, and a multimodal version for vision–language pretraining, **GenView++ (VL)**. The difference lies in their view generation strategies and contrastive objectives. We detail each mode below.

*1) **GenView++ (V) for Vision-Only Pretraining**:* In the vision-only setting, the framework improves vision representation learning by pairing each real image $\mathbf{I}_i$ with a synthetic view $\mathbf{I}_{i,\text{IC}}$ generated under the *Image-Conditioned* strategy. Unlike prior works that pair two synthetic views [41], [44], [43], this *real-to-synthetic pairing*, inherited from GenView [14], mitigates potential feature drift caused by domain gaps between the generative model's training data and the current pretraining dataset. Moreover, the real image helps stabilize representation quality when the synthetic view contains noise or artifacts. After applying standard augmentations to both images, the resulting pair $\{\mathbf{P}_{i,\text{ori}}, \mathbf{P}_{i,\text{IC}}\}$ is evaluated and optimized using our quality-driven image-image contrastive loss, $\mathcal{L}_i^{\text{img-img, QD}}$, which prioritizes pairs with high foreground consistency and background diversity.

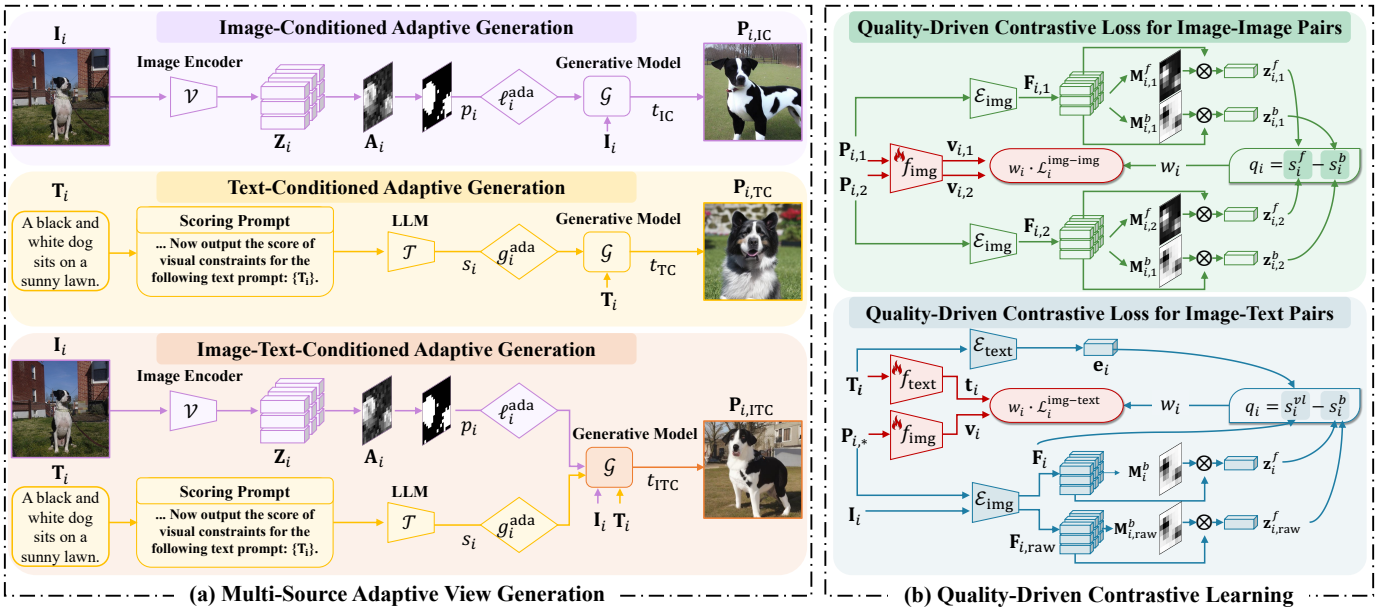


Fig. 2: **Overview of the GenView++ Framework.** The framework enhances contrastive learning via two synergistic modules: **(a) Offline Multi-Source Adaptive View Generation.** Diverse views are synthesized using three adaptive strategies: **Image-Conditioned** generation perturbs visual embeddings of the source image $\mathbf{I}_i$ using adaptive noise ℓ_i^{ada} based on foreground ratio p_i to produce $\mathbf{P}_{i,\text{IC}}$; **Text-Conditioned** generation adjusts guidance scale g_i^{ada} based on caption complexity s_i for $\mathbf{P}_{i,\text{TC}}$; **Image-Text-Conditioned** generation jointly leverages both modalities for $\mathbf{P}_{i,\text{ITC}}$. **(b) Online Quality-Driven Contrastive Learning.** A dynamic reweighting scheme adjusts contrastive losses based on quality scores. For intra-modal loss $\mathcal{L}_i^{\text{img-img}}$, q_i is derived from foreground consistency (s_i^f) and background diversity ($-s_i^b$). For cross-modal loss $\mathcal{L}_i^{\text{img-text}}$, q_i reflects cross-modal alignment (s_i^{vl}) and visual diversity ($-s_i^b$). The quality scores are normalized as reweighting factors w_i to prioritize high-quality pairs and suppress noisy pairs. Only the encoders (f_{img}, f_{text}) are trainable; all other modules are frozen.

*2) **GenView++ (VL) for Vision-Language Pretraining**:* In the vision-language setting, the framework leverages the full power of its multimodal capabilities to learn a robust, aligned embedding space. Given an image–text pair $(\mathbf{I}_i, \mathbf{T}_i)$, it activates all three complementary generation strategies: *Image-Conditioned, Text-Conditioned, and Image-Text-Conditioned*. Together with the original image, these yield four views after standard augmentations:

$$\mathcal{P}_i = \{\mathbf{P}_{i,\text{ori}}, \mathbf{P}_{i,\text{IC}}, \mathbf{P}_{i,\text{TC}}, \mathbf{P}_{i,\text{ITC}}\}. \tag{7}$$

The training is guided by two complementary quality-driven objectives. For *intra-modal contrast*, all image pairs in $\mathcal{P}_i$ are optimized with the quality-weighted image–image loss $\mathcal{L}_i^{\text{img-img, QD}}$, guiding the model to learn invariant features that remain robust to variations in pose, style, and background. For *cross-modal contrast*, each generated view in $\mathcal{P}_i$ is aligned with its caption embedding $\mathbf{t}_i$ through the image–text loss $\mathcal{L}_i^{\text{img-text, QD}}$, ensuring that diverse visual variations remain semantically consistent with the textual description and thereby strengthening cross-modal alignment. By combining intra-modal contrast, which builds invariance to visual changes, with cross-modal contrast, which grounds these variations in textual semantics, the framework learns embeddings that are simultaneously robust and semantically aligned.

B. Multi-Source Adaptive View Generation

To move beyond the limitations of both handcrafted augmentations and existing generative methods, we propose a multi-source adaptive view generation framework, which is built on a core principle: *controllability*. Instead of using fixed or random generation parameters, we dynamically adapt the synthesis process for each input to optimally balance semantic fidelity and view diversity. As illustrated in Fig. 2 (a), this mechanism extends our prior work, GenView, by incorporating three adaptive generation modes: Image-Conditioned (IC), Text-Conditioned (TC), and Image-Text-Conditioned (ITC). These modes are powered by a shared, pretrained diffusion model (Stable Diffusion 2.1 [8]), but differ in how they leverage input data and how they adaptively control the generation process.

Importantly, the entire view generation process is performed *offline*, introducing no additional computational overhead during training. The generated high-quality views can be saved and reused across different models and pretraining frameworks.

*1) **Mechanisms for Controllable Synthesis**:* The generation process starts from a standard Gaussian latent vector $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I})$, followed by T iterative denoising steps using a pretrained diffusion model $G(\cdot)$, conditioned on image features, text embeddings, or both. To enable controllable synthesis, we adaptively modulate two key hyperparameters of the diffusion model, which govern the trade-off between fidelity and diversity. Below, we detail each parameter and its effect on generation.

Noise Level (ℓ): This parameter governs the extent of visual variation from the input image $\mathbf{I}_i$ and is applied in image-conditioned modes (IC and ITC). We adopt Stable unCLIP 2-1 as our generative model $\mathcal{G}$, which supports joint conditioning

on both image and text inputs. The image condition embedding is obtained via the model's image encoder $\mathcal{V}$: $\mathbf{c}_i^{\text{img}} = \mathcal{V}(\mathbf{I}_i)$. To introduce controlled diversity, we perturb this embedding via a forward diffusion process over ℓ steps. The noise level ℓ directly controls the deviation from the original semantics by injecting noise into the embedding as:

$$\hat{\mathbf{c}}_i^{\text{img}} = \text{noisy}(\mathbf{c}_i^{\text{img}}, \ell) = \sqrt{\bar{\alpha}_\ell}\mathbf{c}_i^{\text{img}} + \sqrt{1-\bar{\alpha}_\ell}\varepsilon, \tag{8}$$

where $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ denotes Gaussian noise, $\bar{\alpha}_\ell = \prod_{j=1}^{\ell} \alpha_j$ is the cumulative product of noise schedule variances as in standard DDPM [54]. The perturbed embedding $\hat{\mathbf{c}}_i^{\text{img}}$ is then used to guide the reverse diffusion process, generating a new image conditioned on the noised semantics. The choice of ℓ plays a pivotal role in generative augmentation. A static or randomly selected noise level often leads to suboptimal trade-offs. **Higher values of** ℓ promote greater *diversity* by introducing stronger perturbations, but may distort key semantic content, especially for images with small or indistinct subjects. Conversely, **lower values of** ℓ better preserve *fidelity*, yet risk generating near-duplicates or visually redundant samples when the original image already contains a salient, well-defined subject.

Guidance Scale (g): This parameter modulates the strength of textual conditioning during generation via Classifier-Free Guidance [55] using Stable Diffusion 2-1, and is activated in generation modes involving textual input (i.e., TC and ITC). To guide generation with the caption $\mathbf{T}_i$, we first encode it via the generator's text encoder $\mathcal{T}$ to obtain the embedding $\mathbf{c}_i^{\text{text}} = \mathcal{T}(\mathbf{T}_i)$. During each reverse diffusion step t, the noise estimate is computed as an interpolation between an unconditional prediction (based on an empty caption $\varnothing$) and a text-conditioned one, with the guidance scale g controlling the relative strength:

$$\epsilon_t = \mathcal{G}(\mathbf{z}_i, t, \varnothing) + g \cdot \left(\mathcal{G}(\mathbf{z}_i, t, \mathbf{c}_i^{\text{text}}) - \mathcal{G}(\mathbf{z}_i, t, \varnothing)\right). \tag{9}$$

The selection of g directly impacts the degree of semantic alignment with the textual description. **Lower values of** g weaken the textual constraint, fostering greater *diversity* and creative variation. While this benefits complex or descriptive captions, it may induce semantic drift for simple or ambiguous captions. In contrast, **higher values of** g enforce semantic *fidelity*, tightly aligning the generation with the caption. However, overly strong guidance may lead to repetitive outputs or limit variation, especially when the caption is already highly detailed.

Both ℓ and g significantly affect the fidelity-diversity trade-off in generative view synthesis. We propose a unified strategy to adaptively select noise perturbation level ℓ_i^{ada} and guidance scale g_i^{ada} for each sample $\mathbf{I}_i$ or $\mathbf{T}_i$, based on its visual and semantic characteristics. This enables controllable and context-aware generation that we will detail next.

*2) **Image-Conditioned Adaptive Generation**:* In IC mode, GenView++ generates diverse yet semantically coherent views by modulating the noise level ℓ_i^{ada} based solely on the source image. Following our prior work, GenView [14], the adaptive noise level is determined according to the visual saliency of the main subject. Specifically, salient subjects tolerate stronger perturbations to enrich diversity, while less prominent ones are preserved with lower perturbation to maintain semantic fidelity. The process involves the following steps:

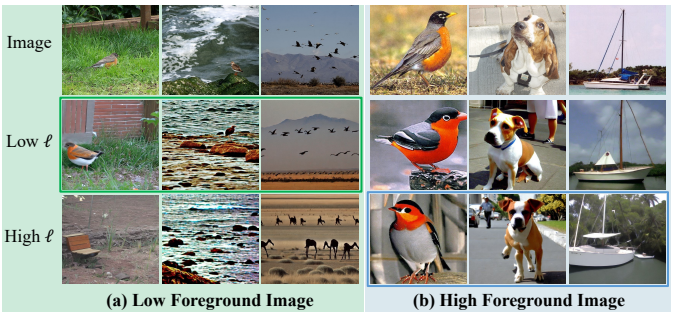


Fig. 3: **Image-Conditioned Adaptive Generation.** The noise level ℓ is adapted to the image's foreground proportion. (a) For low-foreground images, a low ℓ (green) avoids semantic drift, object loss, or distortion (Col 1–3). (b) For high-foreground images, a high ℓ (blue) enriches diversity, e.g., varying pose (Col 4), action (Col 5), and background (Col 6).

Step 1: Foreground Proportion Estimation. We first compute a foreground proportion score $p_i \in [0, 1]$ that quantifies the saliency of the subject in image $\mathbf{I}_i$. We extract dense visual features $\mathbf{F}_i = \mathcal{V}(\mathbf{I}_i) \in \mathbb{R}^{H\times W\times K}$ using the pretrained CLIP image encoder $\mathcal{V}$ of the generator $\mathcal{G}$. We perform Principal Component Analysis (PCA) over a subset of 10,000 training samples to derive the first principal component $\mathbf{w} \in \mathbb{R}^K$, which acts as a global foreground direction. The foreground activation map is then computed as: $\mathbf{A}_i = \mathbf{F}_i \cdot \mathbf{w}, \quad \mathbf{A}_i \in \mathbb{R}^{H\times W}$. This map is normalized to obtain a foreground probability map:

$$\hat{\mathbf{A}}_i = \frac{\mathbf{A}_i - \min(\mathbf{A}_i)}{\max(\mathbf{A}_i) - \min(\mathbf{A}_i)}, \quad \hat{\mathbf{A}}_i \in \mathbb{R}^{H\times W}. \tag{10}$$

The foreground proportion p_i is defined as the ratio of spatial locations whose activation exceeds a threshold α:

$$p_i = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbf{1}\left\{\hat{\mathbf{A}}_i[h, w] > \alpha\right\}, \tag{11}$$

where α is set so that approximately 40% of image tokens are considered foreground, and $\mathbf{1}\cdot$ denotes the indicator function, which equals 1 if the normalized activation value at spatial location (h, w) exceeds the threshold a, and 0 otherwise.

Step 2: Mapping to Adaptive Noise Level. The foreground proportion p_i is mapped to an adaptive noise level ℓ_i^{ada}. We discretize the continuous proportion p_i into five bins, each corresponding to a progressively higher noise level:

$$\ell_i^{\text{ada}} = 100 \cdot \left\lfloor \frac{p_i}{0.2} \right\rfloor, \ell_i^{\text{ada}} \in \{0, 100, 200, 300, 400\}, \tag{12}$$

where we cap ℓ_i^{ada} at 400 to prevent semantic collapse caused by excessive perturbing. As shown in Fig. 3, our adaptive strategy assigns a low ℓ to images with less prominent subjects (low p_i), which are vulnerable to corruption, to preserve crucial semantic details. Conversely, it assigns a high ℓ to images with dominant subjects (high p_i) to safely encourage diversity. Finally, the synthetic view is generated by conditioning the diffusion model $\mathcal{G}$ on image embedding perturbed by this adaptive noise level:

$$\mathbf{I}_{i,\text{IC}} = \mathcal{G}\left(\mathbf{z}_i, \text{noisy}(\mathbf{c}_i^{\text{img}}, \ell_i^{\text{ada}})\right). \tag{13}$$

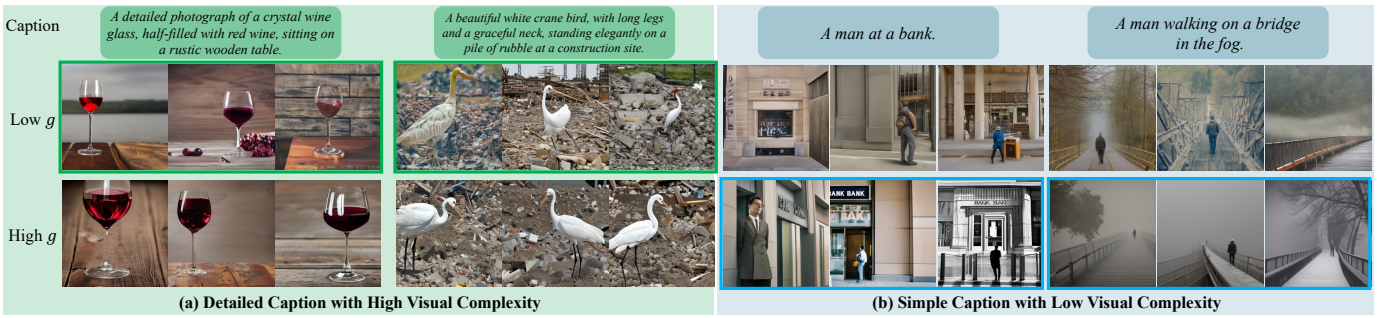


Fig. 4: **Text-Conditioned Adaptive Generation.** The guidance scale g is adjusted by caption complexity. (a) Detailed captions with high visual complexity: a low g (green boxes) encourages diverse generations while preserving fine semantics. In contrast, a high g over-constrains the output, yielding repetitive or rigid results. (b) Simple captions with low visual complexity: a high g (blue boxes) strengthens text-image alignment and ensures key concepts are retained.

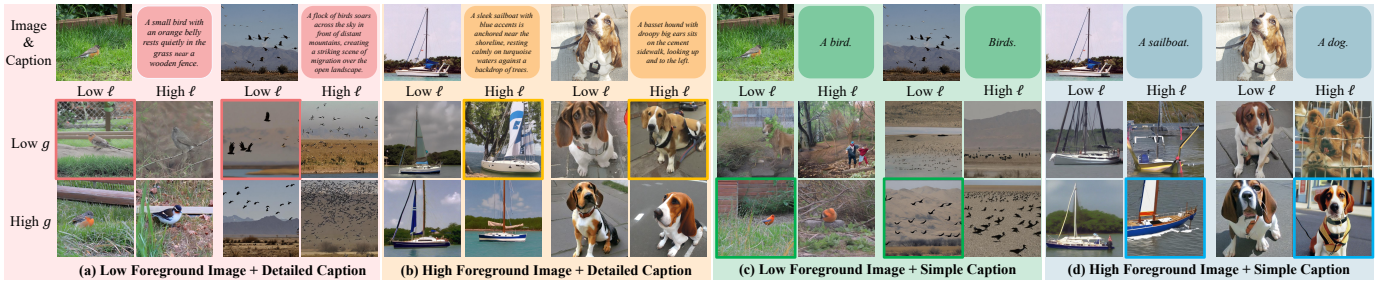


Fig. 5: **Image-Text-Conditioned Adaptive Generation.** This mode jointly adapts noise level ℓ and guidance scale g for fine-grained cross-modal control. (a) Low-foreground + detailed captions: low ℓ preserves small subjects and low g promotes diverse poses and environments (pink boxes); high ℓ or g leads to semantically mismatches (e.g., Row 2, Col 2) or reduced diversity (e.g., Row 3, Col 1). (b) High-foreground + detailed captions: high ℓ enriches variation and low g avoids over-constraining by text (yellow boxes); low ℓ or high g limits diversity. (c) Low-foreground + simple captions: low ℓ preserves subjects while high g enforces text grounding (green boxes); low g induces weak textual constraints, risking semantic drift from the intended concept (e.g., unrelated humans in Row 2, Col 10). (d) High-foreground + simple captions: jointly high ℓ and g produce diverse yet consistent generations grounded in strong visual anchors and minimal text.

*3) **Text-Conditioned Adaptive Generation**:* In TC mode, we generate diverse yet textually-aligned views by modulating the guidance scale g_i^{ada} based on the caption. The adaptive guidance scale is determined according to the caption's estimated visual complexity. Simple or abstract captions require stronger guidance to ensure semantic alignment, while complex, descriptive captions can tolerate weaker guidance to promote creative diversity. The process involves the following steps:

Step 1: Caption Visual Complexity Estimation. We first compute a visual complexity score $s_i \in \{1, 2, 3, 4\}$ that quantifies how richly the caption $\mathbf{T}_i$ describes concrete visual elements. We employ a pretrained language model, prompted with a rule-based scoring instruction $R(\mathbf{T}_i)$ (detailed in Appendix B-C) to perform this evaluation:

$$s_i = \mathcal{T}(R(\mathbf{T}_i)), s_i \in \{1, 2, 3, 4\}. \tag{14}$$

A higher score s_i indicates a caption with greater visual details.

Step 2: Mapping to Adaptive Guidance Scale. The visual complexity score s_i is then mapped to an adaptive guidance scale g_i^{ada}. We discretize the score into four levels, each corresponding to a progressively lower guidance scale:

$$g_i^{\text{ada}} = 10 - 2 \cdot s_i, g_i^{\text{ada}} \in \{2, 4, 6, 8\}. \tag{15}$$

As shown in Fig. 4, our adaptive strategy assigns low g to detailed captions with high visual complexity (high s_i) containing rich semantic constraints to safely encourage the model to explore a wider range of *diversity*. Conversely, it assigns a high g to a low-complexity caption (low s_i), which are vulnerable to semantic drift, to enforce strong text-image alignment. Finally, the synthetic view is generated by conditioning the diffusion $\mathcal{G}$ on the text embedding using this adaptive guidance scale:

$$\mathbf{I}_{i,\text{TC}} = \mathcal{G}(\mathbf{z}_i, \mathbf{c}_i^{\text{text}}, g_i^{\text{ada}}). \tag{16}$$

*4) **Image-Text-Conditioned Adaptive Generation**:* The ITC mode enables fine-grained cross-modal control by jointly adapting the noise level ℓ_i^{ada} and guidance scale g_i^{ada} based on the input image's saliency and the caption's complexity. The dual adaptation unifies the principles of IC and TC modes, with ℓ_i^{ada} and g_i^{ada} computed as in Eq.12 and Eq.15.

As illustrated in Fig. 5, this mechanism dynamically balances fidelity and diversity across four representative scenarios, consistently selecting optimal parameter combinations (highlighted boxes). Specifically, for semantically rich captions (scenarios a & b), a lower guidance scale encourages diverse generation, while simpler captions (scenarios c & d) benefit from a higher guidance scale to maintain semantic alignment. Meanwhile, low-saliency images (a & c) require a lower noise level to

preserve fine details, whereas high-saliency images (b & d) tolerate a higher noise level to enable greater visual variation.

The final synthetic view is generated by conditioning the diffusion model $\mathcal{G}$ on both the perturbed image embedding and the text embedding, guided by adaptively selected parameters:

$$\mathbf{I}_{i,\mathrm{ITC}} = \mathcal{G}(\mathbf{z}_i,\ \mathrm{noisy}(\mathbf{v}_i, \ell_i^{\mathrm{ada}}),\ \mathbf{T}_i,\ g_i^{\mathrm{ada}}). \tag{17}$$

To ensure seamless integration into standard contrastive learning pipelines, the synthesized views $\mathbf{I}_{i,\mathrm{IC}}$, $\mathbf{I}_{i,\mathrm{TC}}$, and $\mathbf{I}_{i,\mathrm{ITC}}$ are further augmented with a random transformation t, producing $\mathbf{P}_{i,\mathrm{IC}}$, $\mathbf{P}_{i,\mathrm{TC}}$, and $\mathbf{P}_{i,\mathrm{ITC}}$ during training. This design ensures plug-and-play compatibility with existing frameworks while combining the semantic richness of generative synthesis with the local invariance benefits of handcrafted augmentations, resulting in complementary multi-scale diversity that better supports robust representation learning.

C. Quality-Driven Contrastive Learning

While our adaptive generation framework creates a diverse set of high-quality source views, the final training pairs may still suffer from noise arising from the stochasticity of generative sampling, distortions from handcrafted augmentations, or weak alignments in web-crawled data. Conventional contrastive learning pipelines treat all positive pairs uniformly, which not only weakens the contribution of high-quality pairs but also exposes the model to misleading supervision.

To address this, we introduce quality-driven contrastive learning that replaces uniform supervision with dynamic supervision. As shown in Fig 2 (b), this online supervision module evaluates the quality of each image-image or image-text pair and reweights its loss contribution accordingly. Importantly, this strategy does not introduce additional trainable parameters. We leverage a frozen, pretrained CLIP model as a lightweight quality assessor, enabling on-the-fly evaluation of both image-image and image-text pair quality. This module complements the generative pipeline by amplifying high-quality, semantically aligned views and suppressing noisy or redundant ones

*1) **Quality-Driven Reweighting for Image-Image Pairs**:* A high-quality image-image pair should exhibit semantic consistency in the foreground while maintaining diversity in the background. This encourages the model to learn representations that are invariant to contextual variations. We quantify this quality by measuring foreground similarity and background dissimilarity, we adopt the assessment mechanism from our prior work, GenView [14], as illustrated in Fig. 2 (b, top right).

Step 1: Saliency-Guided Feature Decoupling. We first sample two augmented views $(\mathbf{P}_{i,1}, \mathbf{P}_{i,2})$ from the offline-generated view pool $\mathcal{P}_i$ (Eq. 7), and extract their dense visual representations using a frozen CLIP image encoder $\mathcal{E}_{\mathrm{img}}(\cdot)$:

$$\mathbf{F}_{i,1}, \mathbf{F}_{i,2} = \mathcal{E}_{\mathrm{img}}(\mathbf{P}_{i,1}), \mathcal{E}_{\mathrm{img}}(\mathbf{P}_{i,W}) \in \mathbb{R}^{H\times W\times K}. \tag{18}$$

where $H \times W$ denotes the spatial resolution and K is the feature dimension at each location. To estimate the spatial importance within each feature map, we apply PCA across the K-dimensional feature vectors at each spatial position. We retain only the first principal component to generate a coarse saliency heatmap, serving as a proxy for object-level attention. This produces unnormalized maps that are then min-max normalized into foreground attention maps: $\mathbf{M}_{i,1}^f, \mathbf{M}_{i,2}^f \in [0,1]^{H\times W}$. The background attention maps are computed as $\mathbf{M}_{i,1}^b = 1 - \mathbf{M}_{i,1}^f, \mathbf{M}_{i,2}^b = 1 - \mathbf{M}_{i,2}^f$. These attention maps guide the pooling process over the dense feature maps, enabling a saliency-aware decomposition into foreground and background features for each view.

$$\begin{aligned} \mathbf{z}_{i,1}^f &= \mathbf{M}_{i,1}^f \otimes \mathbf{F}_{i,1}, \quad \mathbf{z}_{i,2}^f = \mathbf{M}_{i,2}^f \otimes \mathbf{F}_{i,2}, \\ \mathbf{z}_{i,1}^b &= \mathbf{M}_{i,1}^b \otimes \mathbf{F}_{i,1}, \quad \mathbf{z}_{i,2}^b = \mathbf{M}_{i,2}^b \otimes \mathbf{F}_{i,2}, \end{aligned} \tag{19}$$

where the pooling operation $\otimes$ is defined as: $\mathbf{z} = \mathbf{M} \otimes \mathbf{F} = \sum_{h=1}^{H}\sum_{w=1}^{W} \mathbf{M}_{h,w}\mathbf{F}_{[h,w,:]}, \mathbf{z} \in \mathbb{R}^K$. As a result, each input pair yields four decoupled vectors: $\mathbf{z}_{i,1}^f$ and $\mathbf{z}_{i,2}^f$ for foreground semantics, and $\mathbf{z}_{i,1}^b$ and $\mathbf{z}_{i,2}^b$ for background context.

Step 2: Quality Score Computation. The quality score for the pair is computed based on foreground-foreground similarity and background-background similarity as follows:

$$s_i^f = \cos(\mathbf{z}_{i,1}^f, \mathbf{z}_{i,2}^f), \quad s_i^b = \cos(\mathbf{z}_{i,1}^b, \mathbf{z}_{i,2}^b), \tag{20}$$

where s_i^f reflects the semantic similarity between foreground object regions, while s_i^b captures the visual redundancy of backgrounds. The quality score is defined to reward high foreground similarity and penalize contextual redundancy with high background similarity:

$$q_i^{\text{img-text}} = s_i^f - s_i^b. \tag{21}$$

A high $q_i^{\text{img-text}}$ implies the two views contain semantically consistent objects in distinct visual contexts, which is a desirable property for learning robust and invariant representations. Conversely, low scores may arise from excessive background similarity (high s_i^b) or mismatched object semantics (low s_i^f), which can lead to overfitting or misleading supervision.

Step 3: Quality-Driven Loss Reweighting. We normalize the scores via softmax to obtain weights:

$$w_i^{\text{img-text}} = \frac{\exp(q_i^{\text{img-text}})}{\sum_{j=1}^{N}\exp(q_j^{\text{img-text}})}. \tag{22}$$

These quality-aware weights serve as adaptive coefficients for the image-image contrastive loss:

$$\mathcal{L}_i^{\text{img-img,QD}} = w_i \cdot \mathcal{L}_i^{\text{img-img}}, \tag{23}$$

where $\mathcal{L}_i^{\text{img-img}}$ is the selected contrastive loss function from $\{\mathcal{L}_{i,\mathrm{NCE}}^{\text{img-img}}, \mathcal{L}_{i,\mathrm{COS}}^{\text{img-img}}, \mathcal{L}_{i,\mathrm{KL}}^{\text{img-img}}\}$ as defined in Eqs. (2)–(4).

*2) **Quality-Driven Reweighting for Image-Text Pairs**:* Extending the quality-driven supervision principle to the multimodal domain, we introduce a novel scoring mechanism for image-text pairs. The high-quality image-text pair should exhibit strong cross-modal semantic alignment and high visual novelty compared to the original image. We formalize this by measuring image-text consistency and background dissimilarity as shown in Fig. 2 (b, bottom, right).

Step 1: Multimodal Feature Extraction. Using the frozen CLIP encoders, we first extract the text embedding $\mathbf{e}_i$ for the caption $\mathbf{T}_i$, the dense feature maps $\mathbf{F}_{i,\mathrm{raw}}$ for the original,

unaugmented image $\mathbf{I}_i$, and $\mathbf{F}_{i,*}$ for the augmented counterparts $\mathbf{P}_{i,*}$ from the offline-generated view set (as defined in Eq. 7):

$$\begin{aligned} \mathbf{e}_i &= \mathcal{E}_{\text{text}}(\mathbf{T}_i), \quad \mathbf{e}_i \in \mathbb{R}^K, \\ \mathbf{F}_{i,\text{raw}} &= \mathcal{E}_{\text{img}}(\mathbf{I}_i), \quad \mathbf{F}_{i,\text{raw}} \in \mathbb{R}^{H\times W\times K}, \\ \mathbf{F}_{i,*} &= \mathcal{E}_{\text{img}}(\mathbf{P}_{i,*}), \quad \mathbf{F}_{i,*} \in \mathbb{R}^{H\times W\times K}. \end{aligned} \tag{24}$$

These visual features are processed as in Sec. IV-C1 to generate normalized foreground attention maps $\mathbf{M}^f_{i,raw}$ and $\mathbf{M}^f_i$ via PCA and min-max normalization. Background attention maps are:

$$\mathbf{M}^b_{i,\text{raw}} = 1 - \mathbf{M}^f_{i,\text{raw}}, \quad \mathbf{M}^b_{i,*} = 1 - \mathbf{M}^f_{i,*}. \tag{25}$$

Finally, we obtain background-sensitive features by applying spatial weighting between the background masks and their corresponding image feature maps:

$$\mathbf{z}^b_{i,\text{raw}} = \mathbf{M}^b_{i,\text{raw}} \otimes \mathbf{F}_{i,\text{raw}}, \quad \mathbf{z}^b_{i,*} = \mathbf{M}^b_{i,*} \otimes \mathbf{F}_{i,*}. \tag{26}$$

Step 2: Quality Score Computation. We first compute the alignment between the generated image $\mathbf{P}_{i,*}$ and its corresponding caption $\mathbf{T}_i$. We obtain this cross-modal alignment score $s^{vl}_{i,*}$ by calculating the cosine similarity between the text embedding $\mathbf{e}_i$ and the global image representation, obtained via average pooling over the spatial image features:

$$s^{vl}_{i,*} = \cos(\mathbf{e}_i, \text{AvgPool}(\mathbf{F}_{i,*})). \tag{27}$$

A higher $s^{vl}_{i,*}$ indicates better semantic consistency between the visual and textual modalities, which is critical for ensuring that the generated view correctly reflects the original caption.

To assess the visual novelty of the generated view, we compute the cosine similarity between its background feature $\mathbf{z}^b_{i,*}$ and that of the original image $\mathbf{z}^b_{i,\text{raw}}$:

$$s^b_{i,*} = \cos(\mathbf{z}^b_{i,\text{raw}}, \mathbf{z}^b_{i,*}). \tag{28}$$

A higher $s^b_{i,*}$ indicates greater background similarity, implying lower contextual diversity. We define the final image-text quality score by jointly considering cross-modal alignment and background redundancy:

$$q^{\text{img-text}}_{i,*} = s^{vl}_{i,*} - s^b_{i,*}. \tag{29}$$

Step 3: Quality-Driven Loss Reweighting. Similarly, these scores are normalized across all image-text pairs in the batch via softmax to produce weights $w^{\text{img-text}}_{i,*}$:

$$w^{\text{img-text}}_{i,*} = \frac{\exp(q_{i,*})}{\sum_{j=1}^{N}\sum_{*\in\{\text{ori,IC,TC,ITC}\}}\exp(q^{\text{img-text}}_{j,*})}. \tag{30}$$

These weights are used to modulate the standard image-text contrastive loss $\mathcal{L}^{\text{img-img}}_i$:

$$\mathcal{L}^{\text{img-text,QD}}_i = w^{\text{img-text}}_{i,*}\mathcal{L}^{\text{img-img}}_i, \tag{31}$$

where $\mathcal{L}^{\text{img-text}}_i$ is the selected image-text contrastive loss from $\{\mathcal{L}^{\text{img-text}}_{i,\text{i2t}}, \mathcal{L}^{\text{img-text}}_{i,\text{t2i}}\}$ as defined in Eq.(5) and Eq.(6).

In summary, by systematically addressing the core limitations in both how positive pairs are created and how they are utilized during training, GenView++ establishes a more robust and effective foundation to learn more robust and generalizable representations for both vision and vision-language tasks.

TABLE I: **Linear classification accuracy on ImageNet-1K.** *: our reproduction.

Method	Backbone	Epochs	Top-1
InstDisc [19]	ResNet-50	200	56.5
SimCLR [1]	ResNet-50	200	66.8
PCL [27]	ResNet-50	200	67.6
Adco [56]	ResNet-50	200	68.6
InfoMin [35]	ResNet-50	200	70.1
NNCLR [37]	ResNet-50	200	70.7
LEVEL [57]	ResNet-50	200	72.8
Barlow Twins [29]	ResNet-50	300	71.4
CLIP [6]	ResNet-50	-	74.3
MoCov2 [3]	ResNet-50	200	67.5
MoCov2 + C-Crop [58]	ResNet-50	200	67.8
MoCov2 + GenView++ (V)	ResNet-50	200	**70.0**
SwAV [2]*	ResNet-50	200	70.5
SwAV + GenView++ (V)	ResNet-50	200	**71.7**
SimSiam [28]	ResNet-50	200	70.0
SimSiam + GenView++ (V)	ResNet-50	200	**72.2**
BYOL [4]*	ResNet-50	200	71.8
BYOL + GenView++ (V)	ResNet-50	200	**73.2**
MoCov3 [59]	ResNet-50	100	68.9
MoCov3 + GenView++ (V)	ResNet-50	100	**72.7**
MoCov3 [59]	ResNet-50	300	72.8
MoCov3 + GenView++ (V)	ResNet-50	300	**74.8**
MAE [5]	ViT-B	1600	68.0
I-JEPA [24]	ViT-B	600	72.9
VICReg [23]	CNX-S	400	76.2
DINO [60]*	ViT-S	50	70.3
DINO + GenView++ (V)	ViT-S	50	**71.9**
MoCov3 [59]	ViT-S	300	73.2
MoCov3 + GenView++ (V)	ViT-S	300	**74.5**
MoCov3 [59]	ViT-B	300	76.7
MoCov3 + GenView++ (V)	ViT-B	300	**77.8**

TABLE II: **Semi-supervised classification results on ImageNet-1K**. *: our reproduction.

Method	Epochs	1% Labels		10% Labels	
		Top-1	Top-5	Top-1	Top-5
UDA [61]	-	68.8	-	88.5	-
FixMatch [62]	-	-	71.5	89.1	-
PCL [27]	200	-	75.6	-	86.2
SwAV [2]	800	53.9	78.5	70.2	89.9
SimCLR [1]	1000	48.3	75.5	65.6	87.8
Barlow Twins [29]	1000	55.0	79.2	69.7	89.3
NNCLR [37]	1000	56.4	80.7	69.8	89.3
MoCov3 [59]*	100	50.4	76.6	66.8	88.4
MoCov3 + GenView++(V)	100	**51.9**	**78.5**	**68.4**	**89.4**
MoCov2 [3]*	200	42.1	70.9	60.9	84.2
MoCov2 + GenView++ (V)	200	**50.6**	**78.3**	**63.1**	**86.0**
BYOL [4]*	200	53.2	78.8	68.2	89.0
BYOL + GenView++ (V)	200	**55.6**	**81.3**	**68.6**	**89.5**
MoCov3 [59]*	300	56.2	80.7	69.4	89.7
MoCov3 + GenView++ (V)	300	**58.1**	**82.5**	**70.6**	**90.4**

V. Main Results

We conduct extensive experiments to evaluate GenView++. In vision-only pretraining, we integrate it into multiple self-supervised frameworks to verify its effectiveness (Sec. V-A). In vision–language pretraining, we further demonstrate its strong generalization to multimodal tasks (Sec. V-B).

TABLE III: **Transfer learning performance on MS-COCO for object detection and instance segmentation.** *: our reproduction.

Method	Object Det.			Instance Seg.		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
ReSim [63]	39.8	60.2	43.5	36.0	57.1	38.6
DenseCL [64]	40.3	59.9	44.3	36.4	57.0	39.2
SimSiam [28]*	38.5	57.8	42.3	34.7	54.9	37.1
SimSiam + GenView++ (V)	**39.1**	**58.5**	**43.0**	**35.2**	**55.9**	**37.7**
MoCov2 [22]*	39.7	59.4	43.6	35.8	56.5	38.4
MoCov2 + FreeATM [50]	40.1	-	-	-	-	-
MoCov2 + GenView++ (V)	**40.5**	**60.0**	**44.3**	**36.3**	**57.1**	**38.9**
BYOL [4]*	40.6	60.9	44.5	36.7	58.0	39.4
BYOL + GenView++ (V)	**41.2**	**61.5**	**44.9**	**37.0**	**58.4**	**39.7**

TABLE IV: **Comparison of linear classification performance on ImageNet-1K with data expansion methods. Baseline (IN1K):** Uses only the original 1.28M ImageNet-1K images. "+ Laion400M / ImageNet-21K ": Expands the training set by adding 0.3M additional real images from these datasets. "+ Synthetic": Introduces 0.3M diffusion-generated images without adaptive guidance. "+ GenView++ (V)": Leverages only 0.15M synthetic views generated via our image-conditioned mode and a real-to-synthetic pairing strategy.

Dataset	Images	Top-1	Top-5
Baseline (IN1K)	1.28M	62.39	84.57
+ Laion400M	1.28M + 0.3M	63.31	85.53
+ IN21K	1.28M + 0.3M	64.10	85.86
+ Synthetic	1.28M + 0.3M	63.36	85.14
+ GenView++ (V)	1.28M + 0.15M	**65.62**	**87.25**

A. Comparison with Vision Pretraining Methods

We apply GenView++ (V) as a plug-and-play module to six representative vision-only self-supervised learning methods: MoCov2 [3], BYOL [4], DINO [60], SwAV [2], SimSiam [28], and MoCov3 [59]. Experiments are conducted with both convolutional (ResNet-18, ResNet-50 [65]) and transformer (ViT-S, ViT-B [66]) backbones, with ResNet-50 as the default and ViTs used for MoCov3. All models are pretrained on ImageNet-1K [67] (IN1K) and evaluated across standard downstream tasks, including linear classification, semi-supervised learning, and transfer learning for object detection and instance segmentation. To ensure fair comparisons, we strictly follow the original training configurations of each baseline. Pretraining details are provided in Appendix A-C1.

1) ***Linear Classification****:* The primary results for linear classification on ImageNet-1K are presented in Tab. I with implementation details provided in Appendix A-C2. GenView++ (V) exhibits consistent improvements across a wide spectrum of SSL methods, including both contrastive (e.g., MoCov2, MoCov3) and non-contrastive (e.g., DINO, SimSiam, BYOL, SwAV) paradigms, and is effective under both ResNet and ViT backbones. These results demonstrate the plug-and-play nature and broad applicability of GenView++. Its advantages remain evident under varying pretraining durations, consistently outperforming MoCov3 baselines at both 100 and 300 epochs.

TABLE V: **Linear classification results of view construction methods on CIFAR-10, CIFAR-100, and TinyImageNet.**

Method	CF10	CF100	TinyIN
Variance within instance			
MoCov2 + C-Crop [58]	88.78	57.65	47.98
BYOL + C-Crop [58]	92.54	64.62	47.23
Variance within pretraining datasets			
SimCLR + ViewMaker [41]	86.30	-	-
SimCLR + NTN [44]	86.90	-	-
MoCov2 + LMA [43]	92.02	64.89	-
SimSiam + LMA [43]	92.46	65.70	-
Simsiam + DiffAug [45]	87.30	60.10	45.30
Variance beyond pretraining datasets			
W-perturb [68]	92.90	-	51.05
MoCov2 + GenView++ (V)	93.00	67.49	**56.76**
BYOL + GenView++ (V)	**93.56**	**67.53**	54.79

Compared to C-Crop [58], which also improves view diversity through advanced handcrafted transformations, GenView++ yields larger gains by leveraging controlled diffusion-based generative augmentation that is both controllable and content-aware. Notably, when combined with MoCov3 (300 epochs), GenView++ achieves 74.8% top-1 accuracy on IN1K, outperforming CLIP (74.3%) trained on 400M image-text pairs, underscoring the superior data efficiency and high quality of the positive pairs constructed by our framework.

2) ***Semi-Supervised Classification****:* We follow the standard protocol from SimCLR [1] using 1% and 10% labeled subsets of IN1K. Pretrained models are fine-tuned for 20 epochs with a cosine-annealed learning rate schedule. For 1% labels, the classifier learning rate is 1.0 and the backbone 0.00001; for 10%, they are 0.2 and 0.02. As shown in Tab II, GenView++ (V) consistently improves top-1 and top-5 accuracy across different baselines and training durations. With only 1% labeled data, MoCov2 + GenView++ (200 epochs) achieves a +8.5% top-1 accuracy gain; with 1% labels, MoCov3 + GenView++ (300 epochs) still outperforms the baseline by +1.9%.

3) ***Transfer to Dense Prediction Tasks****:* To test generalization to pixel-level tasks, we transfer the pretrained models to MS-COCO [69] object detection and instance segmentation using the Mask R-CNN [70] framework with ResNet-50-FPN. All models are pretrained on IN1K for 200 epochs and fine-tuned on COCO train2017, evaluated on val2017 using the Detectron2 1× schedule [71] (90k iterations, learning rate decay at 60k and 80k). As shown in Tab. III, GenView++ consistently improves both box AP and mask AP across multiple baselines (SimSiam, MoCov2, BYOL), highlighting its ability to produce strong and generalizable visual representations. Compared to FreeATM [50], which also generates additional views via prompt augmentation, GenView++ achieves higher accuracy in detection tasks even without using text, demonstrating the strength of its content-aware, generative augmentation strategy.

4) ***Comparison with Dataset Expansion Methods****:* We assess the effectiveness of GenView++ (V) against several dataset expansion strategies, including incorporating web-

TABLE VI: **Linear classification accuracy on visual benchmarks.** "I-I" and "I-T" indicate the number of image-image and image-text pairs used. StableRep results are reproduced by us with official weights.

Method	Venue	I-I	I-T	IN1K	CF10	CF100	Aircraft	DTD	Flowers	Pets	SUN397	Caltech-101	Food-101	Avg.
CLIP [6]	ICML'21	0	1	53.30	81.80	62.70	34.70	57.30	84.10	60.50	54.30	75.60	58.70	62.30
GenView++ (VL)	-	0	1	**60.49**	**92.24**	**76.08**	**35.01**	**64.47**	**86.18**	**69.96**	**62.41**	**92.45**	**68.24**	**70.75**
StableRep [11]	NeurIPS'23	6	0	63.86	93.03	76.02	29.19	**69.84**	83.53	70.51	**65.53**	90.67	70.91	71.31
GenView++ (VL)	-	6	0	**64.38**	**93.53**	**77.33**	**32.64**	69.15	**86.86**	**73.04**	65.77	**91.47**	**72.39**	**72.66**
SLIP [75]	ECCV'22	1	1	65.40	-	-	-	-	-	-	-	-	-	-
GenView++ (VL)	-	1	1	**66.44**	94.53	79.23	38.46	69.47	87.06	75.88	67.43	91.99	74.13	74.46
GenView++ (VL)	-	6	4	**68.17**	**95.08**	**80.26**	**40.23**	**71.17**	**90.00**	**77.46**	**69.03**	**92.22**	**75.96**	**75.96**

TABLE VII: **Zero-shot classification results on various downstream datasets.** Results for CLIP, SLIP, and TripletCLIP are reproduced by us using their official pretrained weights.

Method	Venue	IN1K	CF10	CF100	Aircraft	DTD	Flowers	Pets	SUN397	Caltech-101	Food-101	Avg.
CLIP [6]	ICML'21	16.34	41.10	17.42	1.11	11.49	10.25	11.37	32.17	49.66	10.73	20.16
SLIP [75]	ECCV'22	23.00	64.38	33.61	1.23	13.19	13.58	**16.68**	33.00	58.16	**14.77**	27.16
LaCLIP [76]	NeurIPS'23	21.50	57.10	27.50	**1.60**	16.60	**14.70**	15.60	35.10	52.70	14.20	25.66
TripletCLIP [52]	NeurIPS'24	7.00	25.26	12.51	1.29	10.69	6.34	5.51	14.78	31.54	4.84	5.09
GenView++ (VL)	-	**23.16**	**87.82**	**50.64**	1.59	**19.10**	12.50	12.73	**42.44**	**60.63**	14.07	**32.47**

crawled image-text pairs (Laion400M [13]), adding external real images (ImageNet-21K [72], abbreviated as IN21K), and using synthetic views without adaptive generation. All models adopt MoCo v3 with ResNet-50 and are pretrained for 50 epochs on IN1K combined with 0.3M external samples, except GenView++(V), which uses only 0.15M generated views. More details are provided in Appendix A-D. We report linear probing results on IN1K in Tab.IV. Expanding IN1K with Laion400M or synthetic images yields a slight improvement in top-1 accuracy, suggesting a limited contribution when directly incorporating images with a domain gap. Extending IN1K with IN21K brings moderate improvements, indicating the benefits from more training data in a similar domain. The best performance is achieved by GenView++(V) using only 0.15 million generated images, yielding a significant 3.2% top-1 accuracy gain. This highlights that the effectiveness of GenView++ mainly stems from better pair construction and adaptive generation, instead of introducing more training data. Remarkably, it outperforms ImageNet-21K expansion using half the number of synthetic samples, while avoiding the significant manual effort of large-scale data collection. Since our diffusion-based view generation is offline, reusable, and automation-friendly, its one-time cost is both efficient and worthwhile.

*5) **Comparison with View Construction Methods**:* To evaluate GenView++ (V)'s effectiveness in enhancing vision representation learning models compared to existing positive view construction methods, we conduct pretraining and linear evaluation on CIFAR-10 [73] (CF10), CIFAR-100 [73] (CF100), and Tiny ImageNet [74] (TinyIN). Implementation details are in Appendix A-E. As shown in Tab. V, GenView++ (V), when combined with MoCov2, consistently outperforms the other data augmentation methods, demonstrating its effectiveness in borrowing rich knowledge from large-scale datasets to construct high-quality positive views.

B. Comparison with Vision-Language Pretraining Methods

We evaluate the full GenView+ (VL) on a suite of vision-language benchmarks, including linear probing on 10 datasets, zero-shot classification, and cross-modal retrieval. Following established protocols from StableRep [11] and the `CLIP_benchmark` evaluation pipeline, we use ViT-B/16 as the default backbone and compare GenView++ against strong baselines including CLIP [6], SLIP [75], TripletCLIP [52], and SynthCLIP [77]. We pretrain the model on the CC3M dataset [12], which is partially enhanced by our adaptive view generation module. Details are provided in Appendix A-F1.

*1) **Linear Classification**:* We evaluate the visual representations learned by GenView++ (VL) using the standard linear probing protocol across 10 benchmarks. Implementation details are provided in Appendix A-F3. Results are shown in Tab. VI, from which we make the following observations:

Vision-Language Setting (0 I-I, 1 I-T): Using one image-text pair per sample, we match CLIP's setup by sampling one view from the offline pool $\mathcal{P}$, constructing a image-text pair. Despite using only cross-modal supervision, GenView++ achieves a significant average accuracy of 70.75%, outperforming CLIP (62.30%) by +8.45%, highlighting the benefits of multi-source adaptive generation and quality-driven reweighting for learning robust visual features.

Vision-only setting (6 I-I, 0 I-T): All six pairs are image-image, constructed from the original view and three generated views using our multi-view contrastive strategy without image-text supervision. Compared to StableRep, which also uses four views generated from text prompts, GenView++ performs better on 9 of 10 datasets, especially on fine-grained categories like Aircraft (+3.45%). This confirms that multi-source image-conditioned generation provides richer intra-modal variation than text-only generation.

Mixed Setting (1 I-I, 1 I-T; 6 I-I, 4 I-T): Combining intra-modal and cross-modal supervision yields the best results. Under the 1 I-I, 1 I-T setup, GenView++ surpasses SLIP on the IN1K dataset, confirming the high quality of our generated views. In the full setting (6 I-I, 4 I-T), GenView++ achieves state-of-the-art results, with 68.17% on IN1K and 75.96% on Food-101. These gains demonstrate the synergistic benefits of aligning both intra-modal and cross-modal representations, made possible

TABLE VIII: **Zero-shot cross-modal retrieval performance on MS COCO, Flickr30k, and Flickr8k.** We report Recall@1 (R@1) and Recall@5 (R@5) for Text-to-Image and Image-to-Text retrieval. CLIP and SLIP results are reproduced by us.

Method	Venue	MS Coco				Flickr30k				Flickr8k	
		Text→Image		Image→Text		Text→Image		Image→Text		Text→Image	Image→Text
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@1
CLIP [6]	ICML'21	9.80	25.78	13.40	31.00	19.12	41.32	25.80	54.50	20.00	29.00
SLIP [75]	ECCV'22	15.25	34.87	**20.60**	43.88	27.74	51.48	36.90	63.80	28.36	40.10
TripletCLIP [52]	NeurIPS'24	-	11.28	-	10.38	-	22.00	-	22.00	-	-
GenView++ (VL)	-	**16.12**	**36.80**	20.12	**45.02**	**28.16**	**54.82**	**38.20**	**67.00**	**29.08**	**41.10**

by our unified framework's ability to generate and prioritize diverse, semantically rich training pairs.

*2) **Zero-Shot Classification**:* To evaluate the cross-modal generalization capability of GenView++, we perform zero-shot classification on ten diverse benchmarks after pretraining on CC3M, without any task-specific fine-tuning. We report top-1 accuracy across all datasets. As shown in Tab. VII, GenView++ achieves an average accuracy of **32.47%**, surpassing CLIP by **+12.31%**, SLIP by **+5.31%**, and LaCLIP by **+6.81%**. Notably, our model achieves strong results on general-purpose datasets such as CIFAR-10 (87.82%) and CIFAR-100 (50.64%), demonstrating the robustness and transferability of the learned representations. Unlike methods such as TripletCLIP [52], which focus primarily on hard negative synthesis, GenView++ improves contrastive learning by generating semantically consistent yet diverse positive views. These high-quality positives not only enhance the alignment signal but also implicitly *enrich the pool of negative samples*, enabling the model to learn more discriminative and transferable features. Coupled with our quality-driven contrastive learning strategy, which prioritizes informative pairs, this design leads to more stable cross-modal alignment and superior downstream generalization.

*3) **Zero-Shot Cross-Modal Retrieval**:* To further assess the learned cross-modal alignment, we conduct zero-shot text-to-image and image-to-text retrieval on MS COCO [69], Flickr30k [78], and Flickr8k [79]. Retrieval is based on cosine similarity between image and text embeddings obtained from the pretrained model. As shown in Tab. VIII, GenView++ consistently outperforms prior approaches across all benchmarks. On MS COCO, it achieves improvements over SLIP in both text-to-image (R@5: +1.93%) and image-to-text retrieval (R@5: +1.14%). On Flickr30k, GenView++ delivers a substantial improvement in image-to-text R@5, outperforming SLIP by **+3.20%** and CLIP by **+12.50%**. Similar trends are observed on Flickr8k. These results highlight GenView++'s ability to construct a semantically aligned and retrieval-effective shared embedding space, facilitating precise cross-modal matching.

VI. Ablation Studies

We conduct comprehensive ablation studies to dissect the contribution of each component in our GenView++ framework in Sec VI-A and Sec. VI-B.

A. Analysis of Unimodal Components

We examine the unimodal components of GenView++ in a purely visual representation learning setting. Experiments are conducted with ResNet-18 models pretrained on ImageNet-100 [21] (IN100) for 100 epochs using MoCov3 as the baseline. Full training details are provided in Appendix A-G.

TABLE IX: **Effect of core components in GenView++ (V) under linear evaluation on IN100.** "Our framework" denotes real-to-synthetic view construction without adaptive noise control or quality-driven loss. "AdaGen" refers to adaptive image-conditioned view generation. "Qual.Driv.Cont" indicates the use of quality-driven contrastive loss.

Our framework	AdaGen	Qual.Driv.Cont	Top-1
×	×	×	65.52
×	×	✓	66.97 (↑ 1.45)
✓	×	×	71.50 (↑ 5.98)
✓	✓	×	73.96 (↑ 8.44)
✓	×	✓	74.88 (↑ 9.36)
✓	✓	✓	**75.40 (↑ 9.88)**

TABLE X: **Comparison of noise level selection strategies.**

Method	CS(0)	CS(100)	CS(200)	CS(300)	CS(400)	RS	AS
Top-1	71.80	72.14	71.50	71.76	72.08	72.96	**73.96**
Top-5	92.19	92.34	91.88	92.02	92.36	92.78	**93.22**

TABLE XI: **Influence of generative augmentation probability under linear evaluation on IN1K.**

α	0	0.1	0.3	0.5	0.8	1.0
Top-1	62.39	65.86	68.38	69.04	69.47	**70.55**
Top-5	84.57	87.10	89.02	89.29	89.49	**90.34**

*1) **Influence of Each Component**:* As shown in Tab. IX, introducing our generative framework alone yields a notable **+5.98%** gain over the baseline. Adding adaptive view generation (AdaGen) further improves performance to **+8.44%**, while quality-driven contrastive learning (Qual.Driv.Cont) also provides significant benefits. The full model, which integrates all components, achieves the best result with a **+9.88%** overall improvement. These findings highlight the strong synergy between controllable generation and quality-aware supervision.

*2) **Impact of Noise Level Selection Strategy**:* To investigate the effectiveness of our adaptive noise selection strategy, we compare three strategies: fixed Constant Selection (CS), Random Selection (RS), and our Adaptive Selection (AS). Results in Tab. X show that **AS (73.96% Top-1)** consistently outperforms both static and random choices. This confirms that dynamically tailoring noise to image content is essential for balancing diversity and fidelity, thereby avoiding redundant or misleading pairs resulting from static or random noise levels.

TABLE XII: **Effectiveness of multi-source adaptive generation.** We compare different generation sources (IC, TC, ITC) and the effect of our adaptive strategy (AdaGen).

Generation Setting				500k Aug.		100k Aug.	
IC	TC	ITC	AdaGen	CF10	IN100	CF10	IN100
×	×	×	×	83.41	66.76	83.41	66.76
✓	×	×	×	86.98	72.96	85.17	68.98
✓	×	×	✓	88.05	76.50	86.02	71.54
×	✓	×	×	87.55	72.60	84.84	69.14
×	✓	×	✓	88.99	75.26	85.62	70.86
×	×	✓	×	87.39	72.86	84.40	68.92
×	×	✓	✓	88.60	76.38	85.45	71.40
✓	✓	✓	✓	90.69	78.38	87.61	74.28

*3) **Impact of Generative Augmentation Probability**:* We further assess the effect of applying generative augmentation with probability α. As shown in Tab. XI, higher α values consistently yield better results, with Top-1 accuracy rising from 62.39% (no augmentation) to **70.55%** at $\alpha = 1.0$. This steady improvement highlights the scalability of our framework and the clear benefits of fully leveraging high-quality synthetic views. We set $\alpha = 1$ to maximize this benefit in our main experiments.

B. Analysis of Multimodal Components

We analyze the core components of GenView++ in the vision-language setting. All models, using a ViT-B/16 backbone, are pretrained on a 1M subset of the CC3M dataset, which is partially enhanced (100k or 500k) by our adaptive view generation module. We evaluate performance using linear probe top-1 accuracy on CIFAR-10 and ImageNet-100. Full training details can be found in Appendix A-H.

*1) **Effectiveness of Multi-Source Adaptive Generation**:* Tab. XII examines how different generation sources and the adaptive generative strategy affect performance:

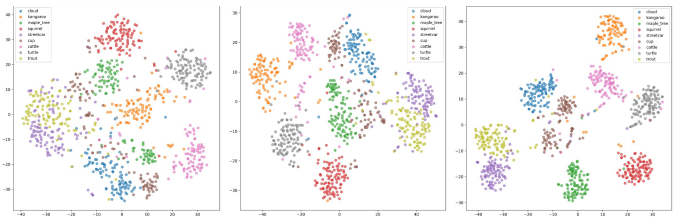
Foundational gains from generated views: Replacing part of the training data with views from a single generation mode (IC, TC, ITC in Rows 2, 4, 6) already improves over the no-generation baseline (Row 1). For instance, with IC-only at the 500k setting, IN100 rises from 66.76% to 72.96% (+6.20). These results demonstrate that enriching training with generative diversity consistently enhances representation quality.

Importance of the adaptive generation: Enabling AdaGen consistently boosts accuracy for all sources. For IC at the 500k setting, AdaGen further improves IN100 from 72.96% to 76.50% (+3.54). Similar trends hold for TC (+2.66 on IN100) and ITC (+3.52 on IN100). At the 100k setting, AdaGen also brings steady gains (e.g., IC: +2.56 on IN100). This underscores the effectiveness of our controllable generation strategy in producing more informative and reliable views.

Synergy of multi-source generation: Combining IC, TC, and ITC *with* AdaGen achieves the best results, reaching 90.69% (CF10) and 78.38% (IN100) at 500k, and 87.61% / 74.28% at 100k. This indicates that multi-source views, when adaptively controlled, yield a broader range of variations while preserving semantic consistency, surpassing single-source alternatives.

TABLE XIII: **Effectiveness of quality-driven contrastive learning.** "Qual.Driv.Cont" denotes quality-driven loss, and "Lang.Sup" denotes the image-text contrastive loss.

Qual.Driv.Cont	Lang.Sup	500k Aug.		100k Aug.	
		CF10	IN100	CF10	IN100
×	×	90.69	78.38	87.61	74.28
✓	×	91.61	79.58	90.77	78.10
×	✓	92.03	80.04	91.98	79.64
✓	✓	92.85	81.38	92.72	80.56



(a) BaseLine (b) Adaptive View Generation (c) Adaptive View Generation + Quality-Driven Contrastive Learning

Fig. 6: **T-SNE visualization of image features from 9 randomly selected CIFAR-100 classes**. (a) Baseline with standard augmentations; (b) model with multi-source adaptive view generation; (c) full GenView++ with the additional quality-driven contrastive loss.

*2) **Effectiveness of Quality-Driven Contrastive Learning**:* Tab. XIII reports the effect of our quality-driven contrastive loss:

Vision-only pretraining: When training without language supervision (image-to-image contrast only), incorporating our quality-driven loss yields substantial gains. With 100k augmented samples, IN100 accuracy improves from 74.28% to 78.10% (**+3.82%**, Row 2 vs. Row 1), confirming that quality-based reweighting is highly effective even in a unimodal setting.

Vision-language pretraining: The advantage persists when language supervision is enabled. Under the 500k setting, our quality-driven loss improves IN100 accuracy from 80.04% to 81.38% (**+1.34%**, Row 4 vs. Row 3). This improvement demonstrates that prioritizing high-quality pairs refines the training signal, enhancing both unimodal and multimodal learning.

*3) **Impact of Generative Augmentation Ratio**:* We further examine the effect of varying the proportion of synthetic samples. Comparing 100k and 500k settings across Tabs XII and XIII, we observe a clear positive trend. For example, in Table XIII (Row 4), increasing synthetic views from 100k to 500k boosts IN100 accuracy from 80.56% to 81.38% (**+0.82%**). These results indicate that larger amounts of high-quality generative data consistently enhance downstream performance, validating our framework as both scalable and effective.

C. Qualitative Analysis

To qualitatively evaluate feature representations, we visualize t-SNE embeddings of ViT-B/16 models pretrained on a 1M subset of CC3M for 40 epochs, with results shown on 9 randomly selected CIFAR-100 classes (Fig. 6). **(a) Baseline:** Standard augmentations produce diffuse and overlapping clusters, indicating limited inter-class separability and noisy features. **(b) +Adaptive views:** Incorporating multi-source adaptive generation yields more compact and clearly separated clusters,

showing that semantically diverse yet coherent views encourage more discriminative representations. **(c) Full GenView++:** Adding the quality-driven contrastive loss further sharpens decision boundaries and improves intra-class compactness, evidencing the synergy between controlled view generation and quality-aware reweighting. These visualizations provide strong qualitative evidence that GenView++ enhances both the discriminability and semantic alignment of learned features.

VII. Conclusion

In this work, we introduced GenView++, a unified framework that advances contrastive learning for both vision and vision–language representation learning. GenView++ integrates two complementary components: (1) a multi-source adaptive view generation module that leverages image-, text-, and joint-conditioned synthesis to produce diverse yet semantically aligned views; and (2) a quality-driven contrastive learning mechanism that conducts online quality assessment and dynamic reweighting, amplifying informative pairs while suppressing noise. Both modules are model-agnostic and seamlessly integrate into existing contrastive learning pipelines. The generation process is performed offline and reusable, while the supervision reweighting is lightweight and efficient, making GenView++ practical for large-scale pretraining. Extensive experiments verify its effectiveness and generality, consistently improving strong baselines across vision-only and vision–language tasks. Ablation studies further highlight the synergy between controllable view creation and quality-driven supervision as the key to its success. In summary, GenView++ provides a robust and scalable pipeline that jointly optimizes the construction and utilization of positive pairs, laying a solid foundation for future research on controllable data generation and quality-driven supervision in representation learning.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[2] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *NeurIPS*, 2020.

[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[4] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," in *NeurIPS*, 2020.

[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[7] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[9] Y. Zhang, D. Zhou, B. Hooi, K. Wang, and J. Feng, "Expanding small-scale datasets with guided imagination," in *NeurIPS*, 2023.

[10] M. Ye-Bin, N. Hyeon-Woo, W. Choi, N. Kim, S. Kwak, and T.-H. Oh, "Exploiting synthetic data for data imbalance problems: baselines from a data perspective," *arXiv:2308.00994*, 2023.

[11] Y. Tian, L. Fan, P. Isola, H. Chang, and D. Krishnan, "Stablerep: Synthetic images from text-to-image models make strong visual representation learners," in *NeurIPS*, 2023.

[12] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018.

[13] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *NeurIPS*, 2022.

[14] X. Li, Y. Yang, X. Li, J. Wu, Y. Yu, B. Ghanem, and M. Zhang, "Genview: Enhancing view quality with pretrained generative model for self-supervised learning," in *ECCV*, 2024.

[15] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.

[16] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.

[17] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008.

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.

[19] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018.

[20] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv:1807.03748*, 2018.

[21] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *ECCV*, 2020.

[22] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," in *arXiv:2003.04297*, 2020.

[23] A. Bardes, J. Ponce, and Y. LeCun, "Vicregl: Self-supervised learning of local visual features," *NeurIPS*, 2022.

[24] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *CVPR*, 2023.

[25] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *ECCV*, 2018.

[26] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *ICLR*, 2020.

[27] J. Li, P. Zhou, C. Xiong, R. Socher, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," in *ICLR*, 2020.

[28] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, 2021.

[29] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *ICML*, 2021.

[30] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in *ICML*, 2021.

[31] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[32] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021.

[33] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, 2019.

[34] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," in *EMNLP*, 2019.

[35] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *NeurIPS*, 2020.

[36] R. R. Selvaraju, K. Desai, J. Johnson, and N. Naik, "Casting your model: Learning to localize improves self-supervised representations," in *CVPR*, 2021.

[37] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *ICCV*, 2021.

[38] R. Tang, C. Ma, W. E. Zhang, Q. Wu, and X. Yang, "Semantic equivalent adversarial data augmentation for visual question answering," in *ECCV*, 2020.

[39] X. Hao, Y. Zhu, S. Appalaraju, A. Zhang, W. Zhang, B. Li, and M. Li, "Mixgen: A new multi-modal data augmentation," in *WACV*, 2023.

[40] N. Xu, W. Mao, P. Wei, and D. Zeng, "Mda: Multimodal data augmentation framework for boosting performance on sentiment/emotion classification tasks," *IEEE Intell. Syst.*, 2020.

[41] A. Tamkin, M. Wu, and N. Goodman, "Viewmaker networks: Learning views for unsupervised representation learning," in *ICLR*, 2020.

[42] P. Astolfi, A. Casanova, J. Verbeek, P. Vincent, A. Romero-Soriano, and M. Drozdzal, "Instance-conditioned gan data augmentation for representation learning," *arXiv:2303.09677*, 2023.

[43] Y. Yang, W. Y. Cheung, C. Liu, and X. Ji, "Local manifold augmentation for multiview semantic consistency," *arXiv:2211.02798*, 2022.

[44] T. Kim, D. Das, S. Choi, M. Jeong, S. Yang, S. Yun, and C. Kim, "Neural transformation network to generate diverse views for contrastive learning," in *CVPR*, 2023.

[45] Z. Zang, H. Luo, K. Wang, P. Zhang, F. Wang, S. Li, Y. You *et al.*, "Boosting unsupervised contrastive learning using diffusion-based data augmentation from scratch," *arXiv:2309.07909*, 2023.

[46] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi, "Is synthetic data from generative models ready for image recognition?" in *ICLR*, 2022.

[47] J. Shipard, A. Wiliem, K. N. Thanh, W. Xiang, and C. Fookes, "Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion," in *CVPR*, 2023.

[48] A. Jahanian, X. Puig, Y. Tian, and P. Isola, "Generative models as a data source for multiview representation learning," in *ICLR*, 2021.

[49] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," in *ICLR*, 2023.

[50] D. J. Zhang, M. Xu, C. Xue, W. Zhang, X. Han, S. Bai, and M. Z. Shou, "Free-atm: Exploring unsupervised learning on diffusion-generated images with free attention masks," *arXiv:2308.06739*, 2023.

[51] Z. Wang, L. Wei, T. Wang, H. Chen, Y. Hao, X. Wang, X. He, and Q. Tian, "Enhance image classification via inter-class image mixup with diffusion model," in *CVPR*, 2024.

[52] M. Patel, N. S. A. Kusumba, S. Cheng, C. Kim, T. Gokhale, C. Baral *et al.*, "Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives," in *NeurIPS*, 2024.

[53] Y. Fu, C. Chen, Y. Qiao, and Y. Yu, "Dreamda: Generative data augmentation with diffusion models," *arXiv:2403.12803*, 2024.

[54] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[55] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS*, 2022.

[56] G.-J. Qi, L. Zhang, F. Lin, and X. Wang, "Learning generalized transformation equivariant representations via autoencoding transformations," *TPAMI*, 2020.

[57] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Learning where to learn in cross-view self-supervised learning," in *CVPR*, 2022.

[58] X. Peng, K. Wang, Z. Zhu, M. Wang, and Y. You, "Crafting better contrastive views for siamese representation learning," in *CVPR*, 2022.

[59] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *ICCV*, 2021.

[60] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.

[61] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *NeurIPS*, 2020.

[62] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020.

[63] T. Xiao, C. J. Reed, X. Wang, K. Keutzer, and T. Darrell, "Region similarity representation learning," in *ICCV*, 2021.

[64] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *CVPR*, 2021.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," in *ICLR*, 2020.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[68] L. Han, S. Han, S. Sudalairaj, C. Loh, R. Dangovski, F. Deng, P. Agrawal, D. Metaxas, L. Karlinsky, T.-W. Weng *et al.*, "Constructive assimilation: Boosting contrastive learning performance through view generation strategies," *arXiv:2304.00601*, 2023.

[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: common objects in context," in *ECCV*, 2014.

[70] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.

[71] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019.

[72] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv:2104.10972*, 2021.

[73] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[74] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," in *CS 231N*, 2015.

[75] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," in *ECCV*, 2022.

[76] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023.

[77] H. A. A. K. Hammoud, H. Itani, F. Pizzati, P. Torr, A. Bibi, and B. Ghanem, "Synthclip: Are we ready for a fully synthetic clip training?" in *CVPR Workshop*, 2024.

[78] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, 2014.

[79] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *JAIR*, 2013.

[80] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[81] R. Beaumont, "Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them." GitHub, 2022.

[82] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv:1306.5151*, 2013.

[83] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *CV[R]*, 2014.

[84] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *ICVGIP*, 2008.

[85] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *CVPR*, 2012.

[86] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010.

[87] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *CVPR Workshop*, 2004.

[88] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*, 2014.

Appendix A
Implementation Details

A. Multi-Source Adaptive View Generation Details

*1) **Generative Models**:* For view synthesis, we employ Stable Diffusion v2-1-unCLIP in the image-conditioned and image–text-conditioned modes, and Stable Diffusion v2-1 in the text-conditioned mode.

*2) **Foreground Proportion Estimation**:* To estimate the foreground proportion in Step 1 of Sec. IV-B2, we adopt the pretrained CLIP ViT-H/14 backbone[1], which also serves as the conditional image encoder $\mathcal{V}$ in Stable UnCLIP v2-1. Given a 224×224 input, this backbone produces 256 tokens of dimension 1280. For PCA-based feature extraction, we randomly sample 10,000 images from the training set. The threshold α in Eq. 11 is chosen such that foreground tokens constitute roughly 40% of the total, ensuring a clear separation between foreground and background, as illustrated in Fig. 2.

*3) **Adaptive View Generation Details**:* **For IN1K, CF10, CF100, and TinyIN**, we generate one additional view per training image using the image-conditioned adaptive view generation introduced in Sec. IV-B2. The number of denoising steps T is set to 20 for computational efficiency, and the classifier-free guidance scale [55] is fixed to 10 to preserve image fidelity. The diversity of synthesized views is modulated by applying different noise perturbations to the image embeddings. To align with the input resolution of each dataset, the generated images are downsampled from their native 768^2 resolution to 512^2 (IN1K), 32^2 (CF10 and CF100), and 64^2 (TinyIN). **For the Conceptual Captions 3M (CC3M)** dataset [12], we adopt a hybrid data strategy: a portion of the data is enhanced with our multi-source adaptive view generation. Each enhanced image-text pair is expanded into a positive set (Eq. 7) using the three adaptive generation modes introduced in Secs. IV-B2, IV-B3, and IV-B4. All images in these positive sets, including both original and generated views, are further processed by a standard augmentation pipeline. In the *comparison with multimodal methods* setting (Sec. V-B), we train on the full CC3M dataset (~3M image-text pairs), where a 1M subset is enhanced by the adaptive generation strategy. For the *ablations on multimodal components* (Sec. VI-B), we train on a 1M subset of CC3M, within which only a portion (e.g., 100k or 500k samples) is enhanced, while the rest remain unmodified.

B. Quality-Driven Contrastive Learning Details

We use the pretrained CLIP ConvNeXt-Base (wide embedding) backbone[2] as the encoders $(\mathcal{E}_{\text{img}}, \mathcal{E}_{\text{text}})$ to extract feature maps from augmented positive views for pairwise quality evaluation. For an input of size 224×224, the resulting feature maps have a spatial resolution of 7×7.

C. Comparison with Vision Pretraining Methods

*1) **Pretraining Details**:* For fair comparison with existing vision representation learning approaches, all models are pretrained on IN1K using pretraining hyperparameters summarized in Tab. XV. Our generative augmentation is applied to the entire IN1K pretraining set. Baseline results of MoCov3 reported in Tab. I are reproduced using the official implementation[3].

TABLE XIV: Pretraining hyperparameters used for comparison with vision representation learning methods on IN1K.

	MoCov2	SwAV	SimSiam	BYOL	MoCov3	MoCov3
Optimizer	SGD	LARS	SGD	LARS	LARS	AdamW
Learning Rate	0.03	0.6	0.05	4.8	1.2/9.6/4.8	2.4e-3
Weight Decay	1e-4	1e-4	1e-4	1e-6	1e-6	0.1
Momentum	0.9	0.9	0.9	0.9	0.9	-
Cosine Decay	✓	✓	✓	✓	✓	✓
Batch Size	256	256	256	4096	512/4096/4096	4096/4096
Loss	$\mathcal{L}_{\text{NCE}}^{\text{img-img}}$	$\mathcal{L}_{\text{KL}}^{\text{img-img}}$	$\mathcal{L}_{\text{COS}}^{\text{img-img}}$	$\mathcal{L}_{\text{COS}}^{\text{img-img}}$	$\mathcal{L}_{\text{NCE}}^{\text{img-img}}$	$\mathcal{L}_{\text{NCE}}^{\text{img-img}}$
Epochs	200	200	200	200	100/300	300
Backbone	ResNet50	ResNet50	ResNet50	ResNet50	ResNet50	VIT-S/ViT-B
Embedding Dim	2048	2048	2048	2048	2048	384/768
Projection Dim	128	128	2048	256	256	256

*2) **Linear Classification Details**:* After pretraining, we train a linear classifier on top of the frozen encoder using IN1K. To ensure fairness, we adopt identical training hyperparameters across all baselines: 90 epochs, a batch size of 1,024, and the SGD optimizer with momentum 0.9 and no weight decay. A cosine learning rate schedule [80] is applied, with an initial learning rate of 0.4 for ResNet-based models and 12.0 for ViT-based models.

D. Comparison with Dataset Expansion

To expand IN1K with additional training data without introducing new classes, we employ a retrieval-based technique [81] for expanding IN1K with Laion400M. We query the entire Laion400M dataset with 0.3 million randomly sampled IN1K images and select the most similar image for each query image. For expanding IN1K with IN21K, we randomly sample 0.3 million non-repeating images with labels matching those in the IN1K dataset. For GenView, positive views are generated for 0.15 million randomly sampled IN1K images. For the experiments, the ResNet-50 models are pretrained on the expanded dataset using a batch size of 512. We apply a cosine decay learning rate schedule and use the LARS optimizer with a learning rate of 1.2, weight decay of 1e-6, and momentum of 0.9. Linear evaluation settings align with those in Tab. I.

E. Comparison with View Construction methods

We compare our approach with several baseline methods, including ContrastiveCrop [58] (C-Crop), ViewMaker [41], Neural Transform Network [44] (NTN), Local Manifold Augmentation method [43] (LMA), Diffusion-based augmentation from scratch [45] (DiffAug), and $\mathcal{W}$-perturb [68]. To facilitate a clear comparison, we categorize existing methods by the type of data variance they introduce during augmentation: *Within-instance variance*, such as random cropping or color jitter, models minor appearance changes in the same image. *Within-dataset variance* uses generative or mixup-style methods to synthesize new views from existing samples in the same dataset. *Beyond-dataset variance* introduces additional semantic diversity via external data or models, such as pretrained generators or perturbation functions. For our experiments, we use three datasets: CF10, CF100, and TinyIN. We employ SGD as the

[1] https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

[2] https://huggingface.co/laion/CLIP-convnext_base_w-laion2B-s13B-b82K-augreg

[3] https://github.com/facebookresearch/moco-v3

optimizer with a learning rate of 0.5, weight decay of 5e-4, and momentum of 0.9. The learning rate follows a linear warm-up for 10 epochs and then switches to the cosine decay scheduler. Batch sizes are set to 512 for CF10 and CF100 and 256 for TinyIN. The momentum coefficient for the momentum-updated encoder and memory buffer size is set to 0.99/0.99/0.996 and 4096/4096/16384 for CF10/CF100/TinyIN, respectively. We use the $\mathcal{L}_{\text{SSL, NCE}}$ loss with a temperature of 0.2 and train for 500 epochs on CF10 and CF100 and 200 epochs on TinyIN. The backbone architecture used is ResNet18 with an embedding dimension of 512 and a projection dimension of 128. We replace the first 7x7 Conv of stride 2 with a 3x3 Conv of stride 1 and remove the first max-pooling operation. For data augmentations, we use random resized crops (the lower bound of random crop ratio is set to 0.2), color distortion (strength=0.4) with a probability of 0.8, and Gaussian blur with a probability of 0.5. The images from the CF10/CF100 and TinyIN datasets are resized to 32x32 and 64x64 resolution. Linear evaluation is conducted using a standard protocol: the classifier is trained for 100 epochs with SGD (momentum 0.9, learning rate 0.2, cosine annealing, no weight decay).

F. Comparison with Vision-Language Pretraining Methods

1) ***Pretraining Details****:* Pretraining is conducted on the CC3M dataset [12], with 1M data enhanced by our multi-source adaptive view generation mechanism. We primarily adopt the ViT-B/16 backbone, and all reported results are based on ViT-B/16 unless otherwise noted (TripletCLIP uses ViT-B/32). Unless specified, vision–language models are pretrained for 40 epochs on 8 GPUs with a batch size of 64 per GPU. The AdamW optimizer is used with $\beta_1 = 0.9$, $\beta_2 = 0.98$, a base learning rate of 2×10^{-4}, weight decay of 0.1, and a 5-epoch warmup schedule.

2) ***Downstream Datasets****:* The evaluation is performed on zero-shot image-text retrieval (MS COCO [69], Flickr30k [78], and Flickr8k [79]) and zero-shot/linear classification across ten diverse image classification datasets: ImageNet [67], CIFAR-10 [73], CIFAR-100 [73], Aircraft [82], DTD [83], Flowers [84], Pets [85], SUN397 [86], Caltech-101 [87] and Food-101 [88].

3) ***Linear Classification Details****:* For downstream image classification, we pretrain the vision encoder on CC3M, freeze its weights, and train a linear classifier on top. The classifier is trained for 90 epochs using SGD with a momentum of 0.9. Following standard practice, the base learning rate is selected from a sweep over $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$. Additional hyperparameters for the ten classification datasets used in the linear probe are summarized in Tab. XV.

TABLE XV: Hyperparameters for linear classification.

	IN1K	CF10	CF100	Aircraft	DTD	Flowers	Pets	SUN397	Caltech-101	Food-101
GPUs	8	2	2	1	1	1	1	1	1	2
Batch/GPU	256	256	256	128	64	64	128	256	128	256
Drop last	True	True	True	False	False	False	False	True	False	True
Classes	1000	10	100	1000	47	102	37	397	101	101

G. Ablations of Unimodal Components

We analyze the contribution of the two core components of GenView++: the generative framework and the quality-driven loss. ResNet-18 models are pretrained on IN100 for 100 epochs using MoCov3 as the baseline, with a batch size of 512. IN100 is a subset of IN1K selected by [21]. For conditioning the generation of positive views with GenView, we employ 50,000 randomly selected class-balanced images from IN100. We use a cosine decay learning rate schedule and employ the LARS optimizer with a learning rate of 1.2, weight decay of 1e-6, and momentum of 0.9. Linear evaluation settings are consistent with those detailed in Tab. I, with a training duration of 50 epochs.

H. Ablations of Multimodal Components

We further conduct a detailed ablation study to assess the impact of each component in GenView++ (VL). All models are based on a CLIP (ViT-B/16) backbone and pretrained for 25 epochs with a 2-epoch warmup on a 1M subset of CC3M, where part of the real data is replaced by synthetic samples (500k in the main setting, 100k for ablations). Pretraining uses the AdamW optimizer with weight decay of 0.1, learning rate 0.0004 (base learning rate 0.0002), β_1=0.9, β_2=0.98, batch size 64, and 4 GPUs. Evaluation is conducted via linear probe accuracy on CF10 and IN100. A linear classifier is trained for 90 epochs on top of the frozen backbone using SGD with momentum 0.9, no weight decay, and a batch size of 256. Following StableRep [11], we select the best base learning rate from a sweep in the range $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$.

Appendix B
More Results

A. Visualization and Limitation Analysis

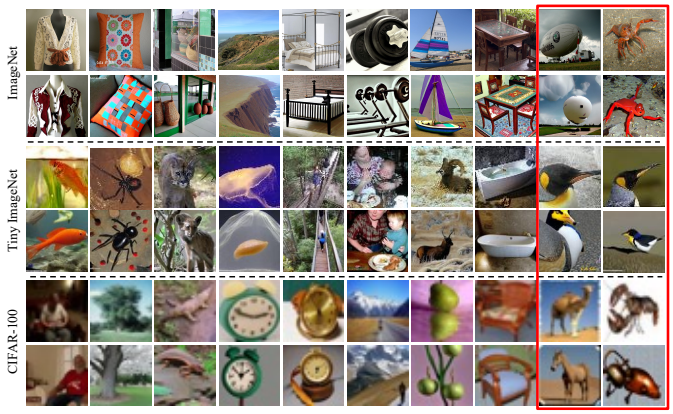


Fig. 7: Visualization of positive pairs generated by GenView++ (image-conditioned mode) across different datasets. Successful variations are shown alongside failure cases (highlighted in red).

Fig. 7 illustrates positive pairs generated by GenView++ across multiple datasets. The top rows display original images, and the bottom rows show images generated by the IC mode of GenView++. The results highlight GenView++'s ability to introduce variations in background, pose, and view angle while preserving the main semantics, which is crucial for learning invariant representations. We also reveal some of its limitations.

- Capacity constraints in the CLIP conditional encoder or generator can challenge the accurate representation of long-tailed categories or complex scenes, resulting in less realistic

generations, such as the generated airships (2nd row of the 9th column) and lobsters (2nd row of the 10th column).
- The granularity of conditional images is crucial, as lower-resolution images can lead to a loss of detail and misclassification of the generated images. For instance, conditioning on a camel image in the 5th row of the 9th column with a resolution of 32^2 produces a generated image resembling a horse, losing the camel's distinctive features.
- Partially visible objects, like the head of the king penguin in the 3rd row of the 9th column, may result in generation errors, yielding images resembling ducks (4th row of the 9th column) or birds (4th row of the 10th column).

Overall, despite these limitations, GenView++'s adaptive view generation produces synthetic samples that largely preserve the attributes of the conditioning images, providing valuable signals for self-supervised training. Moreover, the proposed quality-driven contrastive loss mitigates the effect of occasional semantic inconsistencies. Future directions include improving diffusion-based generation to further enhance view quality and mitigate the identified failure limatations.

B. Evaluation of Positive Views

To ensure effective contrastive learning, it is crucial that synthetic views preserve semantic alignment with the original image. In this experiment, we evaluate the semantic similarity between original images and their associated positive views generated by different strategies. Specifically, we compare three methods: (1) **Retrieval**, which selects the most similar image from the Laion400M dataset; (2) **RS**, a random noise selection strategy from our augmentation pipeline; and (3) **AS**, our proposed adaptive selection strategy. We randomly sample 50,000 images from ImageNet-1K and compute the average cosine similarity between the CLIP image embeddings of each image and its positive view.

TABLE XVI: Average cosine similarity between positive views and original images. **Retrieval** denotes positive pairs formed by retrieving the most similar image from Laion400M. **RS** and **AS** correspond to the noise selection strategies described in Tab. X.

Method	Retrieval	RS	AS (ours)
Cosine Similarity	0.674	0.729	**0.743**

As shown in Tab. XVI, the adaptive selection (AS) strategy achieves the highest semantic similarity (0.743), outperforming both RS (0.729) and Retrieval-based augmentation (0.674). These results demonstrate the effectiveness of our adaptive view selection mechanism in producing semantically consistent views, which is critical for constructing reliable positive pairs and minimizing semantic drift in representation learning.

C. Evaluation of Visual Complexity Estimation Precision

To enable adaptive control over generation quality, our text-conditioned generation module dynamically adjusts the guidance scale based on the visual complexity of input captions. Accurately assessing such complexity is thus essential for effective modulation. We adopt an automatic scoring strategy

TABLE XVII: Rule-based scoring instruction prompt R for evaluating visual complexity of caption inputs.

{"role": "system", "content": "You are tasked with evaluating the visual complexity of a text prompt intended for image generation. Your goal is to assign a score from 1 to 4 that reflects how richly the prompt describes concrete visual elements. The more detailed and diverse the visual information, the higher the score should be. Visual complexity is determined by identifying constraints that directly affect how an image would be generated. These include the specificity of objects mentioned (such as type, quantity, or material), descriptive visual features (such as color, texture, or lighting), spatial relationships (such as layout or perspective), dynamic elements (such as actions or interactions), and stylistic cues (such as artistic genres or cultural elements). A prompt that contains a wide range of such features is considered more complex than one that simply names general categories. A prompt that only refers to broad object types without any additional visual information should be scored as 1. If it includes a small number of visual constraints—typically one to three—it should be scored as 2. A score of 3 is appropriate when the prompt contains a moderate level of detail, roughly four to six constraints. A prompt that includes more than six distinct and specific visual elements—such as combinations of materials, textures, color, spatial arrangement, and style—should be scored as 4. When evaluating, do not include abstract or emotional language that does not translate directly into visual features. Focus only on concrete and visualizable information. After analyzing the prompt, return a score from 1 to 4. Return the result in the following format: [score: ?]."}
{"role": "user", "content": f"Now output the score of visual constraints for the following text prompt:{ }."}

using the DeepSeek-2-Chat-Lite model, which evaluates the richness and concreteness of visual information in text prompts. The scoring criteria, summarized in Tab. XVII, reflect the number and specificity of visual constraints mentioned in a prompt—including object type, quantity, material, texture, spatial relations, dynamics, and style. Prompts with more than six distinct visual attributes are rated as 4 (high complexity), while minimal visual detail yields a score of 1. To assess the accuracy of this automated scoring method, we construct a reference dataset by randomly sampling a subset of captions and annotating their visual complexity using GPT-4o, followed by manual verification. We then compare the LLM-based scores to this ground-truth reference. Experimental results show that our automatic scorer achieves a 93.8% accuracy within a tolerance of ±1. This indicates strong consistency with human perception and validates its use as a lightweight yet effective proxy for complexity-aware generation control.