Machine Learning (2025) 114:63
https://doi.org/10.1007/s10994-024-06665-1

Semantic-aware contrastive learning via multi-prompt alignment

Zhuoran Zhao[1] · Hao Qin[1] · Ming Kong[1,2] · Luyuan Chen[3] · Di Xie[2] · Jiang Zhu[2] · Qiang Zhu[1]

Received: 28 May 2024 / Revised: 13 August 2024 / Accepted: 13 December 2024 /
Published online: 6 February 2025

Abstract
The role of the sample generation mechanism in contrastive learning is pivotal. It not only determines the pairings of positive and negative samples but also enriches the diversity of the sample pool, thereby substantially affecting the quality of the learned representations. Yet, maintaining semantic consistency within positive sample pairs and amplifying sample diversity remain persistent hurdles. To address these challenges, this paper investigates the potential of synthesizing semantically consistent samples by leveraging multi-source and multi-modal prompts, guided by the capabilities of Large Multimodal Models. Through a concise and elegant design, we construct a framework capable of generating semantic-aware positive sample pairs. Based on this framework, we delve deeper into the crucial role of semantic consistency in representation learning through visualization and ablation experiments. Additionally, we systematically outline the fundamental principles and universal methods for generating synthetic samples in contrastive learning using large model techniques. Extensive experimental results prove the superior performance of our method and help us uncover related patterns. We will make all the code and generated datasets publicly available.

Keywords Contrastive learning · Semantic-aware · Multi-prompt alignment

1 Introduction

Contrastive learning has become an effective and widely used self-supervised learning (SSL) method, with representative implementations including SimCLR (Chen et al., 2020; MoCo He et al., 2020; VICReg Bardes et al. (2021), etc. The core idea of this approach is to use data augmentation techniques to generate positive samples from original images. During the training process, the model attempts to maximize the similarity between the positive samples and the difference between negative samples (Chen et al., 2020; Guo

Editors: Kee-Eung Kim, Shou-De Lin.

Zhuoran Zhao and Hao Qin have contributed equally to this work.

Extended author information available on the last page of the article

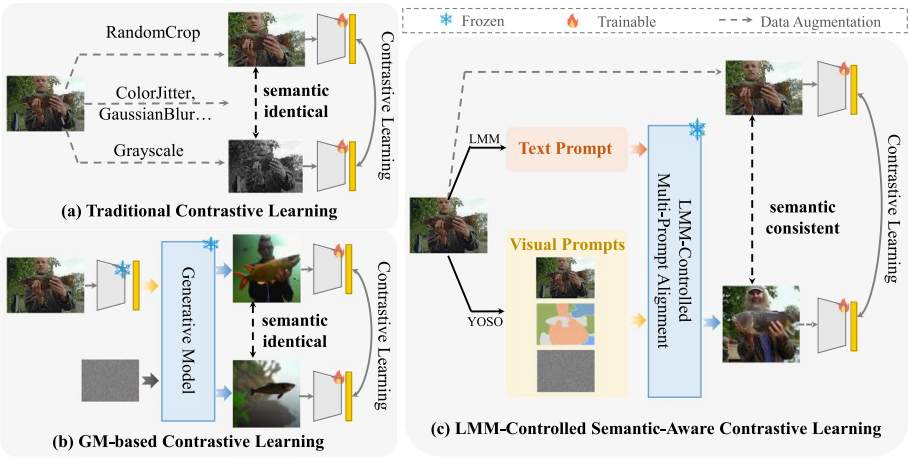

Fig. 1 Three distinct approaches for generating positive samples in contrastive learning: (**a**) Positive samples are created using hand-designed data augmentation methods, which are semantically identical. Visual encoders trained using this approach exhibit robustness primarily to simple geometric and affine transformations. (**b**) To preserve semantic information, the embedding obtained from the image encoder is used as a guiding vector, which is input into the generative model along with noise to generate synthetic images that are semantically consistent with the original image. (**c**) By combining the Text Prompt obtained by Large Multimodal Models (LMM) with various Visual Prompts (image, mask map obtained by YOSO, mask-image mixing, random noise),we achieve sample diversification while maintaining semantic consistency. This approach enables Semantic-Aware Contrastive Learning, resulting in the development of a robust and widely applicable visual encoder

et al., 2022). This dual objective aims to equip the learned visual encoder with robustness and generalization capabilities.

To ensure the rationality of positive samples, traditional data augmentation methods, as shown in Fig. 1a, typically generate positive samples through techniques like cropping, rotation, and flipping. These methods yield limited diversity among positive samples, and the sample space lacks smoothness, significantly restricting the performance of SSL. Therefore, researchers are exploring more effective methods to generate positive samples, such as using generative models to create novel samples (Jahanian et al., 2021; Tian et al., 2024). However, existing methods lack precise control over the sample generators, which can lead to domain shift in the synthetic samples. This, in turn, might compromise the semantic consistency between the positive pairs and affect the stability of the SSL process.

Therefore, we expect to discover a method to control the generator, allowing the two samples in a positive pair are semantically consistent and yet exhibit sufficient visual diversity, so as to improve the robustness and universality of the visual encoder. With the rapid development of Large Multimodal Models (LMM) and generative technologies Huang et al. (2023), the exploration of the impact of synthetic data on models in various fields has attracted increasingly widespread attention (Yang et al., 2024; Tian et al., 2024; Dunlap et al., 2024). However, there is still a lack of exploration into applying the superior cross-modal and generalization abilities of LMM to control the consistency and diversity among positive samples.

This paper proposes a method that leverages existing LMM techniques to enrich the positive samples required for contrastive learning. As shown in Fig. 1c, we construct a concise and elegant framework for generating positive samples: Initially, we input the original image into the LMM to extract its semantic information, forming the Text Prompt for the

generative model to ensure semantic consistency. Subsequently, we utilize various Visual Prompts (image, mask map, mask-image mixing, random noise) to further control the sample generation process, enabling the generator to produce positive samples that are semantically consistent with the original image. Finally, we combine the synthetic images with the original image to create positive sample pairs, which are then input into a contrastive learning framework. To rationally enrich the number of semantic-aware synthetic images, we also design expansion components for the text prompts and visual prompts respectively. It is worth noting that we do not make any task-specific designs but retain the task-agnostic characteristics of self-supervised learning.

Based on this framework, we conduct extensive experiments across a variety of mainstream contrastive learning methods, task scenarios, and data benchmarks, demonstrating the robustness and generalizability of our approach. In addition, more importantly, we thoroughly analyze the fundamental principles of semantic mining and sample generation based on our framework, exploring the optimal generation strategies for semantic-aware samples in contrastive learning. The analysis results reveal the exciting prospects of the ideas proposed in this paper. We believe that these analytical conclusions will offer certain inspirational significance for further research on contrastive learning sample generation based on LMM.

The main contributions of our work can be summarized as follows:

- In the field of contrastive learning, we introduce the semantic-aware positive sample generation mechanism for the first time. By employing a generative model-based approach, we can reasonably enhance sample diversity and optimize the smoothness of the sample space.
- We construct a concise and elegant framework for generating positive samples. By constraining with prompts from multiple sources and modes, we not only ensure the semantic consistency between positive sample pairs, but also make it easy for the generated samples to be seamlessly integrated into various mainstream contrastive learning frameworks and traditional hand-designed data augmentation methods.
- We explore semantic-aware positive samples in contrastive learning to uncover related patterns and find the optimal generation strategy.
- Extensive experiments and visualization results prove the effectiveness, robustness, and generalizability of our approach.

2 Related works

2.1 Contrastive learning

Contrastive learning, which brings positive sample pairs closer together in the representation space while pushing negative sample pairs apart, has been proven effective in visual representation learning. For instance, SimCLR (Chen et al., 2020) generates positive samples through data augmentation and treats other samples as negatives. To reduce memory overhead, MoCo (He et al., 2020) designs a Memory Bank to store negative samples. BYOL (Grill et al., 2020; Chen & He, 2021) avoid training collapse through network architecture design without using negative samples. Barlow Twins Zbontar et al. (2021) and VICReg Bardes et al. (2021), on the other hand, employ the concept of decorrelating features to design loss functions, preventing training collapse and enabling self-supervised

training even with symmetric structures. In addition, many studies now emphasize the crucial role of semantic consistency in contrastive learning (Li et al., 2022; Song et al., 2023). However, these methods focus on the design of model architectures but rely solely on hand-designed data augmentation methods to generate sample pairs, limiting the scope of knowledge that the encoder can learn. We attempt to find a reasonable method to guide the generator to produce richer positive samples.

2.2 Synthetic data augmentation

In the training process of neural networks, data plays a decisive role (Kumar et al., 2023). Efforts have been made to obtain infinite synthetic data from limited real data, and there are currently two main approaches: hand-designed methods (Gong et al., 2021) and modal-based methods (Sariyildiz et al., 2023). Many excellent hand-designed data augmentation methods have emerged in areas such as image segmentation (Bochkovskiy et al., 2020), object detection (Kisantal et al., 2019), and action recognition (Xu et al., 2022). However, hand-designed methods have inherent limitations, with restricted diversity in generated samples and insufficient smoothness in sample space. Based on this, many researchers attempt to train models using synthetic images. Jahanian et al. (2021) utilize Generative Adversarial Networks to generate positive sample pairs from neighboring latent spaces for unsupervised training. Tian et al. (2024) discover that inputting real captions into Stable Diffusion to generate multiple synthetic images, and then training self-supervised methods on these synthetic images, can match or even surpass the performance of methods trained on real images. Sariyildiz et al. (2023) explore the reasonable use of synthetic images in image classification tasks. Meanwhile, Tian et al. (2023) demonstrate that training with synthetic images can yield results surpassing those achieved with real images. These studies explore the potential of synthetic data, but the control over synthetic data is relatively weak. We aim to leverage Large Multimodal Models in conjunction with various prompts to achieve more reasonable control over synthetic data and enhance the effectiveness of visual representation learning.

2.3 Cross-modal content generation

Sample generation has always been a focal point of research. Initially, researchers used different models for the uni-modal generation of images, such as VAE (Kingma & Welling, 2013), CycleGAN (Zhu et al., 2017), and others. With the rapid rise of Large Language Models (LLM) (Touvron et al., 2023) and diffusion models (Ho et al., 2020), cross-modal generation tasks have gained increasing attention. With simple fine-tuning, LLM can have the ability to process cross-modal information (Li et al., 2023), and various variants of diffusion provide the ability to generate images using multi-modal information. Stable Diffusion (Rombach et al., 2022), building upon DDPM (Ho et al., 2020), adds conditional constraints, enabling the generation of 'text-conditioned images'. DALL·E (Reddy et al., 2021) inputs text into GPT during inference to autonomously generate images in an autoregressive manner. ControlNet (Zhang et al., 2023) is an extended model of Stable Diffusion, implementing more efficient multi-condition control channels on the basis of diffusion models. In this paper, we leverage the latest cross-modal understanding and generation techniques to enhance the rationality of contrastive learning.

3 Methods

3.1 Overview

The overall architecture of our method is shown in Fig. 2, which mainly includes LMM-Controlled Multi-Prompt Alignment (LMM-MPA) and Semantic-Aware Contrastive Learning (SA-CL). To obtain Semantic-Aware Synthetic Images (SASI), LMM-MPA first performs 'Semantic Extraction' through large model techniques to capture the representations of input images in the textual space, namely, the semantic information z^*. Then, z^* serves as the control signal for the generator, guiding it to produce SASI by integrating multiple visual conditions. We take SASI and the inputs as positive sample pairs for contrastive learning. MoCo (He et al., 2020) is taken as an example in Fig. 2. This method can be applied to any contrastive learning framework that requires positive sample pairs.

3.2 LMM-controlled multi-prompt alignment

LMM-MPA includes two stages: Semantic Extraction and Generation. Here we explain their technical routes and key details.

Semantic Extraction: Q-Former is an important module of the BLIP2 (Li et al., 2023). It is a lightweight transformer that uses a set of learnable query vectors to extract visual features from a frozen vision model. During this process, the interaction between the learnable query vectors and text is achieved through shared self-attention layers, thereby aligning the image and text modalities. Given the input image X, it is first subjected to a cross-attention operation with a Semantic Extraction Prompt *SEP* (e.g. *"Please describe the picture in detail."*) and multiple trained Queries Q (Li et al., 2023) to extract the information z from X that is meaningful to *SEP*:

$$z = QFormer(X, SEP, Q). \tag{1}$$

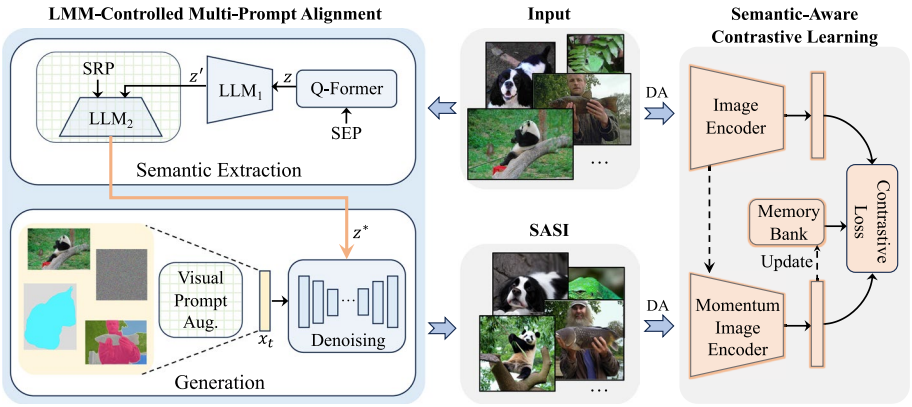

Fig. 2 The overall architecture of the proposed method. Our method mainly consists of LMM-Controlled Multi-Prompt Alignment (LMM-MPA) and Semantic-Aware Contrastive Learning (SA-CL). LMM-MPA provides training samples for SA-CL. Where 'SASI' stands for Semantic-Aware Synthetic Images, 'SEP' for Semantic Extraction Prompt, 'SRP' for Semantic Rewriting Prompt, and 'DA' for Data Augmentation. ▭ indicates optional component

Then, z is fed into an LLM to obtain its semantic representation z' in the text space. Finally, to enhance the diversity of the samples, we design a semantic augmentation phase: generating new semantic representations z^* under the guidance of a Semantic Rewriting Prompt *SRP* using the LLM. To ensure the consistency of semantics before and after augmentation, we provide multiple examples in the *SRP*, as follows:

You are a helpful, respectful and honest assistant, and you can rewrite sentences while retaining their original meaning. I will give you a few examples:

a green python curled up on a branch → A branch held a coiled green python.
two men posing with a fish on a boat → On a boat, two men are seen posing with a fish.
... ...
a man holding a large fish →

We prepare 100 examples and randomly select 5 examples each time to construct the *SRP*, thereby enhancing the richness of the samples. More details are available in the supplementary materials.

Generation: z^* has strong interpretability and can naturally serve as the text prompt for generative models. To more flexibly control the type of SASI, we design various visual prompts (e.g. original image, random noise, mask map, or mask-image mixing) and utilize Stable Diffusion (Rombach et al., 2022) and ControlNet (Zhang et al., 2023) as generators. In the 'Semantic Extraction' stage, we augment semantic information. Similarly, in this stage, we design a visual prompt augmentation phase: generating new visual prompts by adding Gaussian noise to the visual prompts or randomly cropping them.

After the text prompt z^* and visual prompt x_t are given, the synthetic image x_0 can be obtained through DDPM (Ho et al., 2020):

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, t, z^*\right)\right) + \sigma_t n, \tag{2}$$

where ϵ_θ is a noise estimation network, $n \sim \mathcal{N}(0, I)$ and $t \in \{T, ..., 1\}$. More details are available in the supplementary materials.

3.3 Semantic-aware contrastive learning

Semantic is a highly abstract concept that is challenging to represent directly. To determine whether our method truly achieves 'semantic-aware' and to identify the specific semantic information it extracts, we observe and display the extracted semantics alongside the input and output images, as shown in Fig. 3.

To facilitate analysis, we explore representative concepts from various parts of speech for discussion. In Fig. 3, each row, from top to bottom, showcases concepts related to adjectives, prepositions, verbs, and numerals. A real-synthetic sample pair may respond to multiple semantics, and we only annotate the semantics of the group it belongs to for brevity. It can be observed that the positive samples generated by our method maintain consistency with the original images on various critical semantic levels. In fact, the definition of positive sample pairs in contrastive learning is not rigorous enough; at different discrimination scales, two images may transition from the positive to negative sample pair. For example, a lying dog

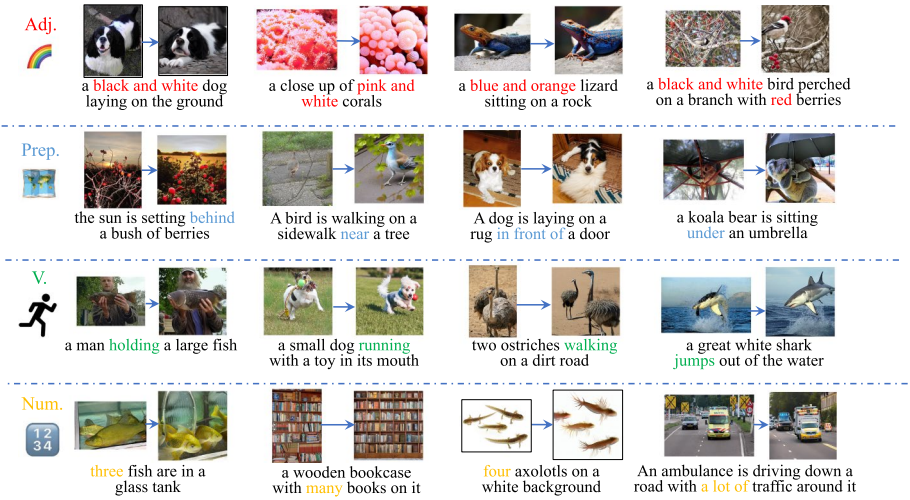

Fig. 3 Visualization of real-synthetic sample pairs and their semantic information. Each row of images represents additional attention to a specific part of speech, from top to bottom: adjectives, prepositions, verbs, and numerals. A real-synthetic sample pair may respond to multiple semantics, and we only additionally annotate the semantics of the group it belongs to

and a running dog may form a positive sample pair on the scale of species classification, but on the scale of action classification, they become a negative sample pair. Unfortunately, in the process of self-supervised learning, we aim to obtain a universal encoder without providing additional information about discrimination scales. Our semantic-aware approach inherently focuses on the most crucial semantic information in the images, such as color and action information in the first column of the first row, action and size information in the second column of the third row, as shown in Fig. 3. This minimizes the generation of inappropriate positive sample pairs during the image generation process. Further exploration and analysis will be conducted in Sect. 4.

4 Experiments

The LMM and LLM used in Semantic Extraction are BLIP2 (Li et al., 2023) and LLaMA (Touvron et al., 2023), respectively. For Generation, we use StableDiffusion-v2.1 and ControlNet-sd15-seg. The mask map is obtained by YOSO (Hu et al., 2023). For SA-CL, to ensure a fair comparison, we use the preset parameters from the public code repositories of the corresponding contrastive learning frameworks as much as possible. All experiments are conducted on a server with 2 A100 GPUs. The datasets we use include Cifar10, Cifar100, ImageNet, Tiny-ImageNet, DTD (Cimpoi et al., 2014), Flowers (Nilsback & Zisserman, 2008), Pets (Parkhi et al., 2012), Food101, STL10, Car196, and ImageNet-Sketch-1k Wang et al. (2019).

4.1 Evaluation on multi-scenarios

Verification and discussion on different SSL methods: We conduct experiments on several classical SSL frameworks on ImageNet-100 (Tian et al., (2020) to verify the universality of

Table 1 The accuracy of different SSL frameworks, where 'w/ SASI' represents training using semantic-aware positive sample pairs generated by LMM-MPA and MoCo-v2* denotes that ResNet101 is used as the backbone

Method		MoCo-v2 (Chen et al., 2020)	MoCo-v2* (Chen et al., 2020)	Simsiam (Chen & He, 2021)	Swav (Caron et al., 2020)	Adco (Hu et al., 2021)	MoCo-v3 (Xinlei et al., 2021)
Linear	Base	73.1	72.4	52.4	82.1	73.8	83.9
	w/ SASI	**78.4** $_{5.3\uparrow}$	**79.6** $_{7.2\uparrow}$	**61.5** $_{9.1\uparrow}$	**86.6** $_{4.5\uparrow}$	**81.2** $_{7.4\uparrow}$	**85.4** $_{1.5\uparrow}$
KNN	Base	67.8	73.8	46.2	75.0	69.0	79.3
	w/ SASI	**76.1** $_{8.3\uparrow}$	**77.5** $_{3.7\uparrow}$	**51.6** $_{5.4\uparrow}$	**80.0** $_{5.0\uparrow}$	**79.1** $_{10.1\uparrow}$	**81.4** $_{2.1\uparrow}$

The parts highlighted in bold indicate the optimal results

Table 2 The experimental results of MoCo-v2 on multiple datasets

Dataset		Cifar10	Cifar100	ImageNet-1k	Tiny-ImageNet	Sketch-1k
Linear	Base	85.5	56.2	67.5	34.8	30.4
	w/ SASI	**88.4** $_{2.9\uparrow}$	**62.6** $_{6.4\uparrow}$	**68.5** $_{1.0\uparrow}$	**47.3** $_{12.5\uparrow}$	**50.3** $_{19.9\uparrow}$
KNN	Base	81.9	45.9	55.6	22.1	24.8
	w/ SASI	**87.5** $_{5.6\uparrow}$	**56.0** $_{10.1\uparrow}$	**62.2** $_{6.6\uparrow}$	**36.6** $_{14.5\uparrow}$	**36.1** $_{11.3\uparrow}$

The parts highlighted in bold indicate the optimal results

our approach. The experimental results, as shown in Table 1, indicate a significant improvement across all models. The models we use include MoCo-v2 (Chen et al., 2020), Simsiam (Chen & He, 2021), Swav (Caron et al., 2020), Adco (Hu et al., 2021), and MoCo-v3 (Xinlei et al., 2021), which contain many classical structures in contrastive learning, such as momentum updating, stop-gradient, multi-crop, asymmetric projection head, online clustering, adversarial contrast, and so on. General improvement of these models strongly proves the universality of our method and the rationality of SASI.

Verification and discussion in different domains: To investigate whether our method can effectively expand samples in different domains, we conduct experiments on Cifar10, Cifar100, ImageNet-1k, Tiny-ImageNet, and Sketch-1k based on MoCo-v2, as shown in Table 2. Meanwhile, we conduct additional experiments on Sketch-1k based on MoCo-v2. As shown in Fig. 4, during the experiments from 100 to 600 epochs on Sketch-1k, the linear classification accuracy is higher after applying our method. It is evident that SASI significantly improves model performance in both accuracy and convergence speed, and the effectiveness of SASI on Sketch-1k demonstrates **the significant potential of our approach in scenarios where it is difficult to collect samples**. Further, we explore the real-synthetic pairs in Sketch-1k, as shown in Fig. 5: 1) We find that even in the presence of some low-quality (even out-of-domain) synthetic images, the model performance still significantly improves. We speculate that this is because the synthetic images, even when out-of-domain, still maintain semantic consistency with the original images, which does not greatly affect the encoder's convergence process. As shown in Table 3 , we also conduct experiments on Sketch-1k using DINO, and after applying our method, the linear classification accuracy improved by 13.6%, further supporting our hypothesis. 2) With

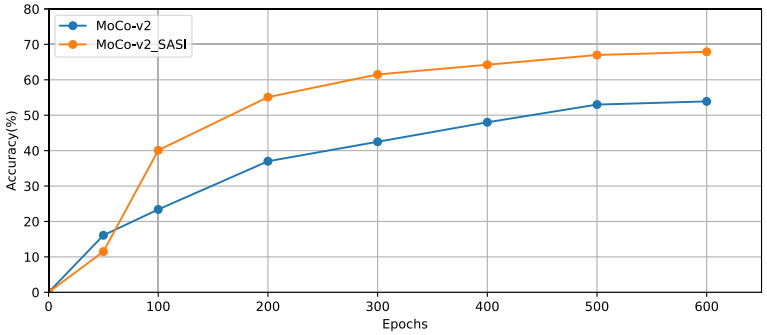

Fig. 4 The linear classification accuracy of MoCo-v2 on ImageNet-Sketch-1k across different epochs, with MoCo-v2_SASI representing our method

Table 3 Top-1 Accuracy on ImageNet-Sketch-1k with DINO (200 epochs) - 13.6% improvement using our method

Method		DINO (Caron et al., 2021)
Linear	Base	33.6
	w/ SASI	**47.2** $_{13.6\uparrow}$

The parts highlighted in bold indicate the optimal results

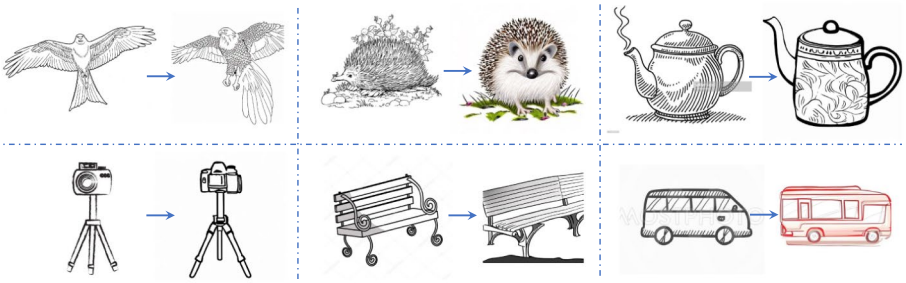

Fig. 5 Visualization of real-synthetic sample pairs in Sketch-1k

Table 4 Experimental results before and after disrupting the semantic-aware positive pairs on Sketch-1k

Method	Linear	KNN
w/ SASI	**50.3**	**36.1**
Mix	43.0 $_{7.3\downarrow}$	32.3 $_{3.8\downarrow}$

The parts highlighted in bold indicate the optimal results

only a few dozen training samples under each category in Sketch-1k, to verify whether the improvement in model performance is solely due to the increased sample diversity, we no longer pair SASI with the original images to form positive pairs but instead mix them directly for experimentation, as shown in Table 4. We find that model performance declines after disrupting the semantic-aware positive pairs, indicating that **both the increase in**

sample diversity and the semantic-aware positive pairs have a positive impact on model performance.

Next, we will utilize the trained encoders Adco and Adco-SASI from Table 1 to conduct experiments on a variety of downstream tasks.

Object detection: We adopt the protocol in Hu et al. (2021) to fine-tune the pre-trained backbone. And the detection network is fine-tuned with the VOC07+12 trainval dataset and tested on the VOC07 test set. The experimental results, as shown in Table 5, indicate that SASI not only enhances the semantic understanding capability of the model but also has a positive effect on the model's localization and detection capabilities.

Transfer learning: Compared with the samples obtained through hand-designed data augmentation, SASI has a richer and smoother sample space, and the encoders trained by SASI should have better generalization ability. To verify this, we conduct transfer learning experiments across multiple datasets, as shown in Table 6. SASI improves accuracy on all datasets, proving the generality of our method.

Semi-supervised learning: Similar to Guo et al. (2022), we evaluate on the task of classification with limited labels. Following the traditional settings, we randomly select 1% or 10% data from the train set to train the classification head of the model, and the experimental results are shown in Table 7. It can be seen that pre-training with SASI can greatly improve the model's performance even with a limited amount of labeled data, highlighting the importance of maintaining semantic consistency between positive samples for contrastive learning.

Table 5 The experimental results of object detection using Faster R-CNN with a R50-C4 backbone on VOC

Method	AP50	AP	AP75
Base	76.7	50.3	54.7
w/ SASI	**78.5** $_{1.8\uparrow}$	**51.1** $_{0.8\uparrow}$	**56.2** $_{1.5\uparrow}$

The parts highlighted in bold indicate the optimal results

Table 6 Linear evaluation on image datasets from various domains

Method	Cifar10	Cifar100	DTD	Flowers	Pets	Food101	STL10	Car196	Avg
base	73.5	55.3	52.4	68.5	60.2	43.2	80.4	32.1	58.2
w/ SASI	**77.4**$_{3.9\uparrow}$	**56.6**$_{1.3\uparrow}$	**54.0**$_{1.6\uparrow}$	**73.7**$_{5.2\uparrow}$	**64.3**$_{4.1\uparrow}$	**44.2**$_{1.0\uparrow}$	**86.9**$_{6.5\uparrow}$	**37.7**$_{5.6\uparrow}$	**61.9**$_{3.7\uparrow}$

The parts highlighted in bold indicate the optimal results

Table 7 Performance on semi-supervised learning

Method	1%		10%	
	Top-1	Top-5	Top-1	Top-5
Base	53.6	79.2	67.6	87.6
w/ SASI	**67.6** $_{14.0\uparrow}$	**87.3** $_{8.1\uparrow}$	**77.8** $_{10.2\uparrow}$	**93.3** $_{5.7\uparrow}$

The parts highlighted in bold indicate the optimal results

Table 8 Comparison of direct generation and controlled generation

Method	Linear	KNN
Direct gen. (Tian et al., 2024)	44.7	31.3
Controlled gen	**50.3** $_{5.6\uparrow}$	**36.1** $_{4.8\uparrow}$

The parts highlighted in bold indicate the optimal results

Table 9 Ablation experiments for *SEP*. '– (base)' indicates not using LMM-MPA to generate positive samples, and relying solely on hand-designed data augmentation

Semantic Extraction Prompt	Linear	KNN
— (base)	73.1	67.8
Question: What does this picture describe? Answer	77.6	74.6
Please describe the picture in detail	**78.4**	**76.1**
What objects are included in this picture?	77.0	73.6

The parts highlighted in bold indicate the optimal results

4.2 Effectiveness of multi-prompt alignment

One existing work (Tian et al., 2024) that is most relevant to SASI uses Stable Diffusion to generate multiple images for contrastive learning, where multiple images generated from the same text under different random seeds form a set of positive samples. Compared to this, the main difference in SASI is the use of multiple prompts to collaboratively control the entire generation process. To compare the effects of direct generation versus controlled generation, we conduct experiments on Sketch-1k. To mimic the data generation process of Tian et al. (2024), we first used BLIP2 to extract captions from the images in the training set, then used these captions to generate images under different random seeds with Stable Diffusion, and finally paired the images with the same caption as positive sample pairs for training on MoCo-v2. As shown in Table 8, the results indicate that controlled generation with multiple prompts as control signals achieved better results compared to direct generation. We speculate that this is because the semantic consistency between positive sample pairs is better ensured under the collaborative work of multiple prompts.

4.3 Ablation studies and analysis

To further explore the underlying principles of SASI's impact on contrastive learning, we conduct ablation experiments on ImageNet-100 and based on MoCo-v2.

Ablation for Semantic Extraction Prompts (SEP): In LMM-MPA, the extraction of semantic information is regulated by *SEP*, and we conduct ablation experiments to explore general patterns. As shown in Table 9, we design three different *SEP*: (a) *"Question: What does this picture describe? Answer:"*, (b) *"Please describe the picture in detail."*, and (c) *"What objects are included in this picture?"*. It can be observed that (1) compared to the baseline without using SASI, regardless of which prompt is selected, there is a significant improvement in accuracy; (2) **providing a more detailed and logical description of the image is beneficial**.

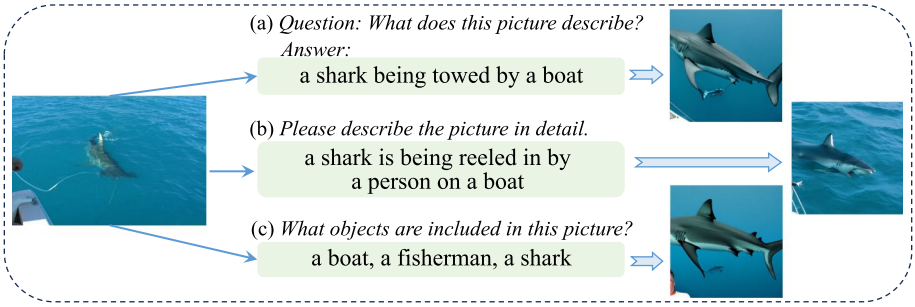

Fig. 6 Diagram of the impact of different *SEP* on LMM-MPA, which shows the detailed process of generating synthetic images from original images

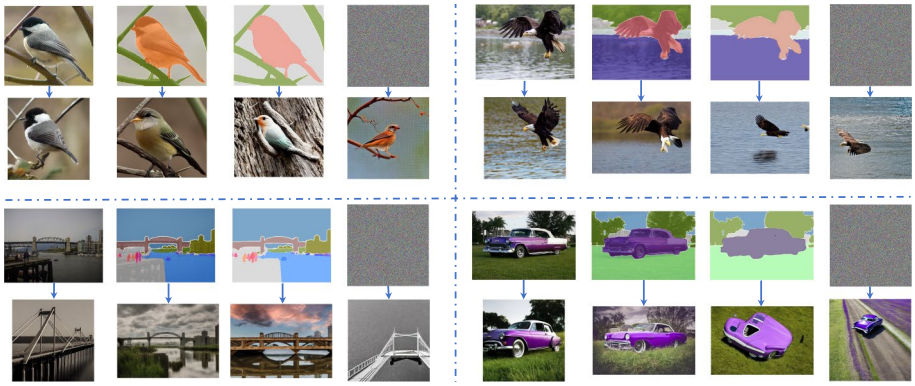

Fig. 7 Schematic of the SASI generated with different *VP*

Table 10 Ablation experiments for *VP*. 'Mask-Image Mixing' indicates that a real image is superimposed with its mask map as the *VP*

Visual Prompt	Linear	KNN
— (base)	73.1	67.8
Real Image	78.4	76.1
Mask Map	77.2	75.5
Mask-Image Mixing	**79.2**	**76.2**
Random Noise	74.7	70.0

The parts highlighted in bold indicate the optimal results

To further analyze the impact of *SEP*, we visualize the semantic information of real-synthetic sample pairs under various *SEP*, as shown in Fig. 6. We find that (1) providing a detailed description is crucial for retaining key semantic information in visually less prominent areas of the image; (2) simple descriptions may lead to the loss of some semantic information; and (3) merely listing objects in the image and neglecting their interactions may result in low-quality synthetic images. Take Fig. 6 as an example, the shark in the original image visually dominates, but the fisherman's hand in the lower left corner

provides crucial semantic information for the entire image. (a) If a general description is used, the synthetic image would lose the key semantic information of 'fisherman catching shark'. (b) In contrast, a more detailed description would preserve this semantic information (the fisherman's hand is retained in the lower left corner of the synthetic image). (c) By using only object information and ignoring their relationships, the synthetic image loses not only the original semantic content of 'fisherman catching shark' but also omits the expected object 'boat'.

Ablation for Visual Prompts (VP): In the generation stage, visual prompts control information such as the layout and color tone of the images. To explore the impact of different *VP*, we conduct ablation experiments as shown in Table 10. The highest accuracy is achieved when using a combination of mask map and image as the *VP*. In contrast, using random noise as the *VP* introduces unreasonable positive samples due to the loss of fundamental information about the original image, leading to a decrease in accuracy.

As shown in Fig. 7, we also visualize different *VP* and their corresponding SASI. Through observation, we find that (1) regardless of the layout, perspective, or color tone, using both a mask map and the original image as *VP* usually results in SASI most similar to the original image; (2) directly using the image as *VP* may alter the original layout or generate a new perspective; (3) using only the mask map fails to retain color information in the details of the original image; and (4) the image generated using random noise as *VP* is only semantically related to the text prompt, exhibiting the maximum deviation from the original image. Combining the findings from Table 10, it is evident that **introducing more refined control conditions during the generation process often leads to improved performance**.

The significance of semantic and visual prompt augmentation: In the aforementioned experiment, we generate only one SASI for each real image. Here, we explore the impact of generating multiple SASI on the model and demonstrate the validity of semantic and visual prompt augmentation. As shown in Table 11, we augment the samples through these two means respectively. During training, one SASI from the multiple SASI belonging to the same image is randomly selected in each iteration to ensure a constant computational load. With the increase in the number of SASI, the model accuracy is further improved, indicating the great potential of our method. Employing either semantic augmentation or visual prompt augmentation alone can enhance the accuracy, proving the rationality of our augmentation strategy.

How to use SASI effectively? In order to explore how SASI can be used better, we conduct ablation experiments as shown in Fig. 8, with results present in Table 12. By comparing (a), (b), and (c), we find that hand-designed data augmentation is still indispensable and has a significant impact on model performance, while **our method can be well compatible with existing data augmentation methods**. Compared to (a) and (b), (c) has a much lower accuracy, which we hypothesize is due to the encoder overfitting caused by the fixed

Table 11 Ablation experiments for semantic augmentation and visual prompt augmentation. '× 1' and '× 2' represent doubling and tripling the number of SASI, respectively

Semantic Aug	*VP* Aug	Linear	KNN
–	–	78.4	76.1
× 1	–	80.1	77.0
–	× 1	80.4	77.5
× 1	× 1	81.2	78.4
× 2	× 2	**81.9**	**79.4**

The parts highlighted in bold indicate the optimal results

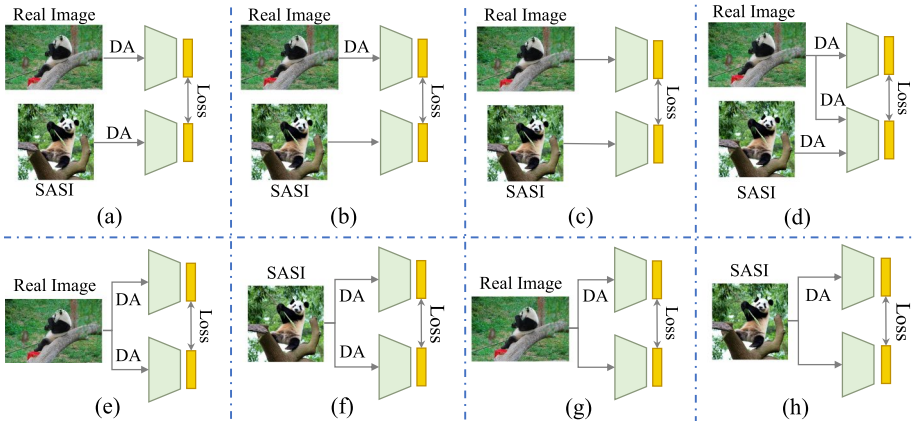

Fig. 8 Schematic of different ways to use SASI

Table 12 The accuracy of various methods in Fig. 8

Accuracy	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Linear	78.4	77.5	49.2	**80.4**	73.1	71.6	71.5	68.9
KNN	76.1	73.4	38.8	**78.3**	67.8	63.7	64.9	60.2

The parts highlighted in bold indicate the optimal results

nature of the samples that the encoder can receive. The method shown in (d) achieves the highest accuracy, indicating that **SASI and real images almost do not introduce domain conflicts during the training process**. Considering that the three-flow structure introduces a 1/2 increase in computational cost, to ensure the rigor of the comparison, the accuracies shown in other tables are all based on (a). By comparing (f) and (h) with (e) and (g) respectively, it can be observed that when synthetic images are used alone as a complete substitute for real images, there is a decrease in accuracy. This may be due to the presence of a certain domain gap between the synthetic images and the images in the test set, and some flaws exist in the details of the synthetic images.

4.4 Discussion and prospect

Through experiments, we demonstrate the effectiveness of our approach and uncover the potential of synthetic data in contrastive learning. During the visualization process, we find that some synthetic samples are of low quality, as shown in Fig. 9. Although these samples may lose some detailed information compared to the original images, they maintain semantic consistency with the original images. The presence of these samples makes our model more robust. To more clearly observe the impact of SASI on the encoder's feature extraction ability, ten categories are randomly selected from ImageNet for feature visualization using t-SNE (Maaten & Hinton, (2008), as shown in Fig. 10. It can be observed that after the introduction of SASI, the feature space becomes more reasonable, with smaller intra-class spacing and larger inter-class spacing.

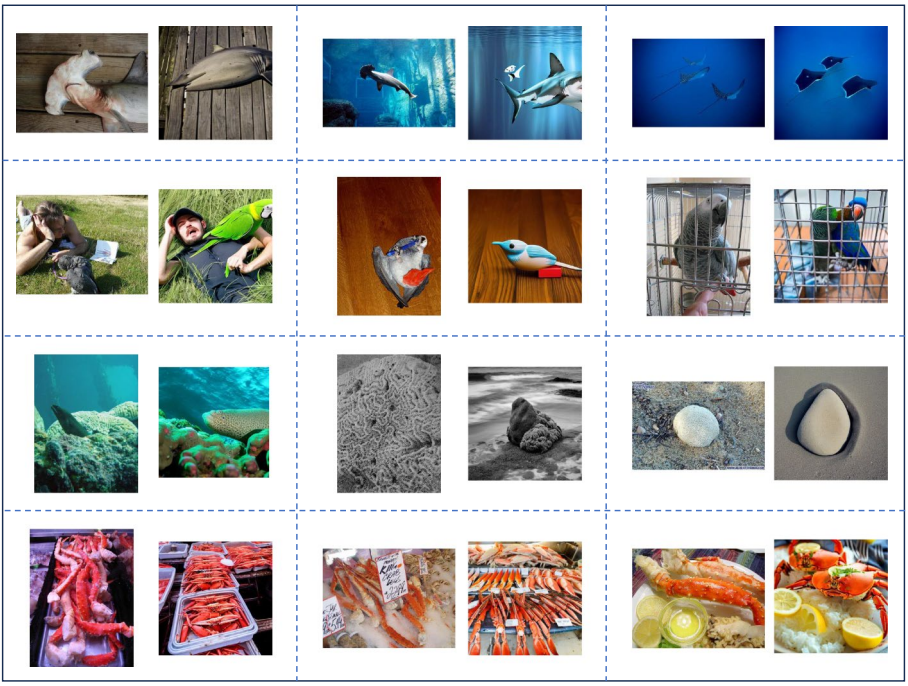

Fig. 9 More visualizations of “bad" samples on ImageNet-100

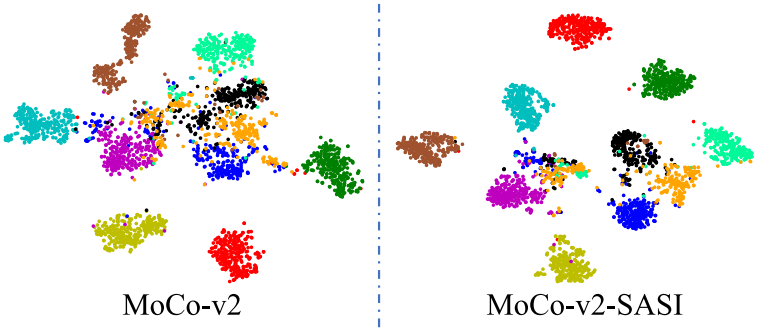

Fig. 10 The t-SNE visualization results of features extracted by MoCo-v2 and MoCo-v2-SASI

Table 13 The GPU hours (A100) required for different stages in our method

Stage	GPU Hours
Training	28
Semantic Extraction	3
Generation	37
Total	68

In addition to this, we compile statistics on the GPU hours (A100) required for various stages in our method during experiments on ImageNet-100 based on MoCo-v2, as shown in Table 13. Compared to the training process, the semantic extraction stage consumes significantly less time (about 10%), while the generation stage leads to a 1.3-fold increase in time consumption. This increase is entirely acceptable because: (1) Data generation is a one-time process; once SASI is initially obtained, many subsequent experiments can bypass the semantic extraction and generation stages and directly reuse existing data. (2) SASI brings substantial improvements with less than a 1.5-fold increase in time consumption, demonstrating its broad application prospects.

However, there are still some directions worth exploring with this method. Firstly, we adopt a multi-stage strategy, separating sample generation from training. Although this strategy can enhance the interpretability of the method, designing a more complex end-to-end strategy may further improve the performance gains brought by SASI. Secondly, for the rigor of experimental data comparison, we employ the default data augmentation methods from public code to process SASI. Nevertheless, exploring data augmentation methods more suitable for synthetic images is meaningful. If the online generation of SASI is achievable, perhaps hand-designed data augmentation will become unnecessary. Finally, synthetic data may conceal certain social biases and errors, and when extended to real-world applications, additional regulatory measures need to be provided.

5 Conclusion

Hand-designed data augmentation methods struggle to maintain semantic consistency while trying to enhance sample diversity, posing a significant challenge to the development of contrastive learning. This is largely because they focus on geometric transformations, leading to a less smooth sample space. To address this, we explored how to generate semantic-aware positive samples and their significance to contrastive learning. Through a concise and elegant framework, we validated the effectiveness, robustness, and generalizability of our method in generating SASI by leveraging large model techniques combined with multiple prompts and uncovered related patterns. Extensive experiments and visualization results helped us explore and summarize the fundamental principles of using large model techniques for sample generation in contrastive learning and demonstrated the broad prospects of our method.

Appendix A: More discussion about fishy SASI

In Fig. 5 of the main text, we observe some out-of-domain synthetic samples. Similarly, in other datasets, there may also be in-domain synthetic samples that do not conform to human cognition. We collectively refer to these synthetic samples, which are traditionally considered dissimilar or unreal compared to the original images, as Fishy Synthetic Samples. We find that the presence of some Fishy Synthetic Samples does not lead to the failure of SASI, which may be due to (1) Fishy Synthetic Samples not having a significant semantic shift compared to the original images, and (2) contrastive learning having a higher tolerance for Fishy Synthetic Samples. To further observe the overall distribution of SASI, we utilize CLIP-ViT-L Radford et al. (2021) to calculate the cosine similarity

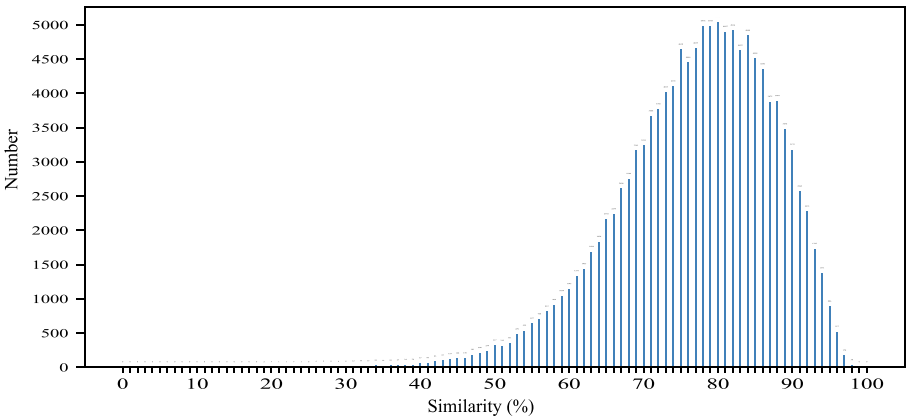

Fig. 11 The distribution of similarity between SASI and the original images

Table 14 Experimental results with different *SRP*

Example Type	Linear	KNN
w /SASI	78.4	76.1
No examples	77.3	75.7
Expansion examples	78.8	76.2
Rewriting examples	**80.1**	**77.0**

The parts highlighted in bold indicate the optimal results

between SASI and the original images, as shown in Fig. 11. It can be seen that the similarity distribution basically conforms to a normal Gaussian distribution, and there are few samples with low similarity. When we replace SASI with the original images for those having a low similarity (less than 0.6), the experimental results do not show improvement (78.4→78.28), indicating that SASI, which are traditionally considered dissimilar or unreal compared to the original images, do not significantly interfere with the training process of contrastive learning. Meanwhile, finding a suitable "bad sample metric" for contrastive learning may be an interesting research direction in the future.

Appendix B: More experiments about *SRP*

In *SRP*, we prepare many examples, which significantly influence the effect of Semantic Augmentation. Here, we demonstrate the model accuracy under different types of examples, as shown in Table 14. First, we attempt to augment semantics without using any examples, relying solely on the LLM's capability, which is referred to as "No examples":

You are a helpful, respectful and honest assistant. You can expand on anything I say to you to make it more expressive. You can imagine appropriately, and you only need to reply with two to three sentences:

a white shark swimming in an aquarium →

In this *SRP*, LLM tends to transform the original semantics into more abstract and artistic descriptions:

a white shark swimming in an aquarium → The white shark glides effortlessly through the crystal-clear waters of the aquarium, its sleek body cutting through the waves with grace and precision. Its pure white skin glistens in the light, giving it an otherworldly appearance as it effortlessly navigates its enclosure. With each flick of its tail, the shark propels itself forward, its powerful muscles rippling beneath its silky skin.

Then, we try to provide some examples in *SRP*, adding some imaginary extra objects to the examples to enhance the contrastive learning's ability to capture key semantics, referred to as "Expansion examples":

You are a helpful, respectful and honest assistant. You can expand sentences to make them more expressive, and you can imagine appropriately. I will give you a few examples:

a green python curled up on a branch → a green python curled up on a branch is about to attack a small bird
two men posing with a fish on a boat → the two pose with a fish on a boat with an island looming behind them
A man in a candy store holding a bowl of candy → A man with the bowl of sweets in the sweet shop, and next to him another little kid laughing his head off
the sydney harbour bridge → sydney Harbour Bridge, under which many ships pass
a person's hand holding a stopwatch on a white background → a person's hand holding a stopwatch on a white background, and he has a nervous look on his face
... ...

a white shark swimming in an aquarium →

This type of *SRP* allows LLM to add some previously nonexistent semantic information while retaining the original semantics. As can be seen from Table 14, the accuracy can be slightly improved in the case of semantic expansion, but the improvement is limited. We speculate that this is because simple semantic changes can cause huge

deviations in the synthesized images, thereby interfering with the contrastive learning. Therefore, in the end, we opt to use rewriting as a means of Semantic Augmentation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10994-024-06665-1.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064.

Author's contribution ZZ and HQ conceived the ideas in this paper, conducted the experiments, and wrote the manuscript. MK and LC assisted with the theoretical analysis. DX and JZ revised the manuscript. QZ was responsible for the overall direction and planning. All authors discussed the results and contributed to the final manuscript.

Funding This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064.

Data availability The source data used in this work are all public.

Code availability The source code is provided in the Supplementary Materials and it will be released after publishing.

Declarations

Conflict of interest None.

References

Bardes, A., Ponce, J., & LeCun, Y. (2021). *Vicreg: Variance-invariance-covariance regularization for self-supervised learning*. arXiv preprint arXiv:2105.04906

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *Yolov4: Optimal speed and accuracy of object detection*. arXiv preprint arXiv:2004.10934

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems, 33*, 9912–9924.

Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758.

Chen, X., Fan, H., Girshick, R., & He, K. (2020). *Improved baselines with momentum contrastive learning*. arXiv preprint arXiv:2003.04297

Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613.

Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J. E., & Darrell, T. (2024). Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in Neural Information Processing Systems, 36*.

Gong, C., Wang, D., Li, M., Chandra, V., & Liu, Q. (2021). Keepaugment: A simple information-preserving data augmentation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1055–1064.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems, 33*, 21271–21284.

Guo, Y., Xu, M., Li, J., Ni, B., Zhu, X., Sun, Z., & Xu, Y. (2022). Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9706–9715

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840–6851.

Hu, J., Huang, L., Ren, T., Zhang, S., Ji, R., & Cao, L. (2023). You only segment once: Towards real-time panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17819–17829.

Hu, Q., Wang, X., Hu, W., & Qi, G.-J. (2021). Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1074–1083.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al. (2023). *Language is not all you need: Aligning perception with language models*. arXiv preprint arXiv:2302.14045

Jahanian, A., Puig, X., Tian, Y., & Isola, P. (2021). *Generative models as a data source for multiview representation learning*. arXiv preprint arXiv:2106.05258

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational bayes*. arXiv preprint arXiv:1312.6114

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., & Cho, K. (2019). *Augmentation for small object detection*. arXiv preprint arXiv:1902.07296

Kumar, T., Turab, M., Raj, K., Mileo, A., Brennan, R., & Bendechache, M. (2023). *Advanced data augmentation approaches: A comprehensive survey and future directions*. arXiv preprint arXiv:2301.02830.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*. arXiv preprint arXiv:2301.12597

Li, Z., Zhu, Y., Yang, F., Li, W., Zhao, C., Chen, Y., Chen, Z., Xie, J., Wu, L., Zhao, R., et al. (2022). Univip: A unified framework for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14627–14636.

Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research, 9*(11).

Nilsback, M.-E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE.

Parkhi, O. M., Vedaldi, A., Zisserman, A., & Jawahar, C. (2012). Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505. IEEE.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR.

Reddy, M. D. M., Basha, M. S. M., Hari, M. M. C., & Penchalaiah, M. N. (2021). Dall-e: Creating images from text. *UGC Care Group I Journal, 8*(14), 71–75.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.

Sariyildiz, M. B., Alahari, K., Larlus, D., & Kalantidis, Y. (2023). Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Song, K., Zhang, S., Luo, Z., Wang, T., & Xie, J. (2023). Semantics-consistent feature search for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16099–16108.

Tian, Y., Fan, L., Isola, P., Chang, H., & Krishnan, D. (2023). *Stablerep: Synthetic images from text-to-image models make strong visual representation learners*. arXiv preprint arXiv:2306.00984

Tian, Y., Fan, L., Isola, P., Chang, H., & Krishnan, D. (2024). Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems, 36*.

Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK*, August 23–28, 2020, Proceedings, Part XI 16, pp. 776–794. Springer.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). *Llama: Open and efficient foundation language models*. arXiv preprint arXiv:2302.13971

Wang, H., Ge, S., Lipton, Z., & Xing, E. P. (2019). Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518.

Xinlei, C., Saining, X., & Kaiming, H. (2021). *An empirical study of training self-supervised visual transformers*. arXiv preprint arXiv:2104.020578, 7.

Xu, K., Ye, F., Zhong, Q., & Xie, D. (2022). Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 2866–2874.

Yang, L., Xu, X., Kang, B., Shi, Y., & Zhao, H. (2024). Freemask: Synthetic images with dense annotations make stronger segmentation models. *Advances in Neural Information Processing Systems, 36*.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR.

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232.

Authors and Affiliations

Zhuoran Zhao[1] · Hao Qin[1] · Ming Kong[1,2] · Luyuan Chen[3] · Di Xie[2] · Jiang Zhu[2] · Qiang Zhu[1]

✉ Qiang Zhu
zhuq@zju.edu.cn

Zhuoran Zhao
zhuoranzhao@zju.edu.cn

Hao Qin
haoqin@zju.edu.cn

Ming Kong
zjukongming@zju.edu.cn

Luyuan Chen
chenly@bistu.edu.cn

Di Xie
xiedi@hikvision.com

Jiang Zhu
zhujiang@hikvision.com

[1] School of Computer Science and Technology, Zhejiang University, HangZhou 310013, China

[2] Hikvision Research Institute, Hangzhou 310051, China

[3] Computer School, Beijing Information Science and Technology University, Beijing 100101, State, China