BEYOND THE OBSERVED: KNOWLEDGE-INFORMED CONTROLLABLE AUGMENTATION FOR FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-Shot Class-Incremental Learning (FSCIL) addresses the practical need to continually acquire novel classes from limited labeled data while retaining previously learned knowledge. While previous studies primarily focus on model-side strategies for improving adaptation and knowledge retention, the role of data-side augmentation remains relatively underexplored. We argue that a key limitation of few-shot scenarios lies in the severely limited intra-class variations captured by sparse observations, and that external knowledge can provide useful priors for exploring plausible variations beyond the observed samples. Based on this insight, we propose **Knowledge-Informed Controllable Augmentation (KICA)** that derives plausible variation instructions from external class knowledge and realizes them through image-conditioned diffusion-based editing. By combining the visual evidence of few-shot observations with external variation knowledge, KICA preserves class-specific characteristics while enriching intra-class variations, achieving broader coverage of real intra-class variations as evidenced by distributional analyses. Extensive experiments across multiple FSCIL benchmarks and network architectures demonstrate consistent performance improvements, which validate the effectiveness of knowledge-informed augmentation and motivate a data-centric direction through expanded intra-class variation coverage.

1 INTRODUCTION

Real-world recognition systems face the challenge of continually acquiring novel classes from limited labeled data while maintaining previously learned knowledge. This challenge motivates the study of Class-Incremental Learning (CIL), which focuses on learning new classes over time while preserving knowledge acquired from previous classes. However, existing CIL settings typically assume sufficient labeled examples for each new class, which is often violated in real-world scenarios due to privacy constraints, annotation costs, and class rarity. Few-Shot Class-Incremental Learning (FSCIL) addresses this more challenging setting by enabling models to learn novel classes from only a few observations while retaining previously acquired knowledges (Tao et al., 2020).

Most FSCIL methods focus on model-side strategies to improve adaptation and knowledge retention. Prototype calibration, representation regularization, prompt learning, and pretrained vision-language models enhance transfer to novel classes, yet they still rely on limited visual observations to learn robust representations (Wang et al., 2023; Ahmed et al., 2024; Oh et al., 2024; Park et al., 2024; Jiang et al., 2025; Chen et al., 2025). As illustrated in Fig. 1, few-shot observations cover only a limited portion of the intra-class variation space, making it difficult to capture the diverse appearances of novel classes. Data augmentation provides a potential solution by enriching these limited observations. However, generating additional samples alone does not guarantee sufficient coverage of class-relevant variations. Existing knowledge-agnostic approaches either rely on observed samples to generate variations or use generic semantic descriptions without explicit guidance on which variations are plausible for a target class. For example, text-to-image generation conditioned solely on class names can produce visually plausible samples, but the coarse semantic information provides limited guidance for exploring diverse yet class-relevant variations. This motivates the need for a controllable augmentation strategy that leverages additional knowledge to discover meaningful class-relevant variations beyond limited observations.

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

To address this challenge, we propose **Knowledge-Informed Controllable Augmentation (KICA)**, a data-level framework that leverages class-specific knowledge to discover plausible intra-class variations and guide the editing of available support images. KICA integrates explicit knowledge from lexical definitions and visually grounded category attribute co-occurrences with semantic knowledge elicited from a pretrained large language model (LLM), forming a class-specific attribute base. The retrieved knowledge is converted into operation-specific positive and negative editing prompts, enabling controllable modifications of background, attributes, and viewpoints while preserving instance-level visual evidence. The generated samples are directly incorporated into the support set, allowing KICA to be applied to various FSCIL methods without modifying their architectures. Extensive experiments on CIFAR-100, miniImageNet, and CUB benchmarks demonstrate consistent improvements across diverse FSCIL architectures based on CNNs, ViTs, and CLIP, covering different adaptation paradigms including prototype-based classification and prompt tuning. Furthermore, distributional analyses show that KICA improves the coverage of intra-class variations, achieving favorable performance in *coverage* and *adjusted IRS*.

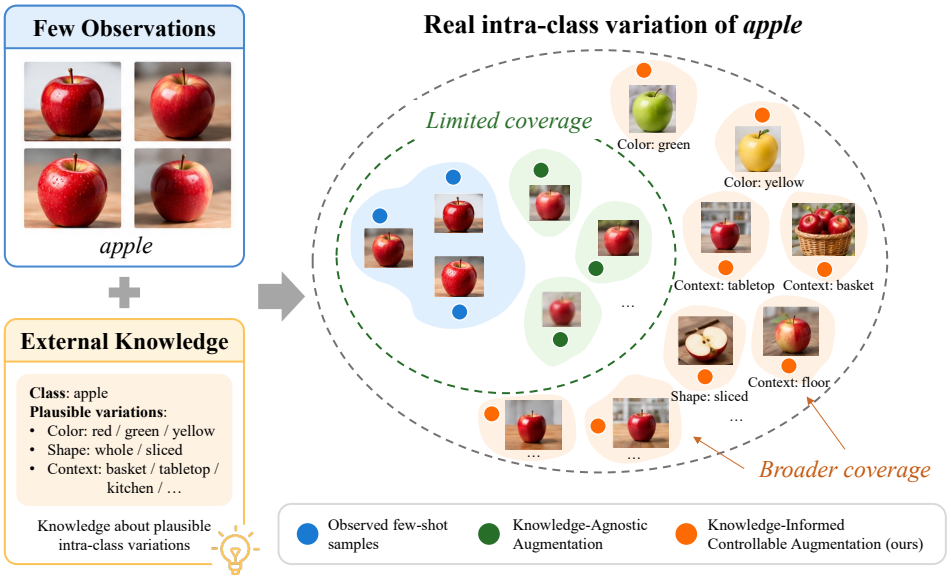


Figure 1: In FSCIL, few-shot observations cover only a limited portion of the intra-class variation space. Existing knowledge-agnostic augmentation strategies generate additional samples based on observed instances or coarse semantic descriptions. Knowledge-informed augmentation class-specific attribute knowledge to discover plausible variations beyond the observed samples, enabling broader and more targeted augmentation under severely limited observations.

2 Related Works

2.1 Few-Shot Class-Incremental Learning

Class-Incremental Learning requires a model to recognize all classes observed so far without task identities at inference time (Rebuffi et al., 2017). Few-Shot Class-Incremental Learning (FSCIL) imposes an additional constraint: after an initial base session with sufficient training data, the model receives only a few labeled examples of each novel class (Tao et al., 2020). FSCIL must therefore retain base-class knowledge, acquire novel classes, and avoid overfitting to limited observations.

Representation- and prototype-based methods improve adaptation by organizing the feature space or estimating more reliable class representations. TOPIC preserves feature-space topology, whereas TEEN calibrates novel-class prototypes using semantic relationships with base classes without additional training (Tao et al., 2020; Wang et al., 2023). OrCo introduces orthogonality constraints and contrastive learning, while CLOSER improves representation quality by explicitly considering inter-class separation (Ahmed et al., 2024; Oh et al., 2024). These approaches mainly alleviate the impact of limited observations through representation optimization or prototype adaptation.

Recent methods exploit large pretrained models through prompt adaptation, modality calibration, or feature transfer. PriViLege introduces knowledge-preserving tuning for pretrained vision-language transformers (Park et al., 2024). LGSP learns complementary local and global spatial prompts, while BiMC performs training-free bi-level modality calibration on CLIP (Jiang et al., 2025; Chen et al., 2025). Diffusion-FSCIL further explores diffusion models by using multi-scale Stable Diffusion features as a unified FSCIL representation and constructing replay features without storing synthetic images (Kim et al., 2025). Different from these model-side adaptation strategies, KICA operates at the data level by creating class-preserving image variations before incremental training, without modifying downstream FSCIL architectures.

2.2 Generative Models for Data Augmentation

Conventional augmentation applies geometric, photometric, or mixing transformations to observed samples: RandAugment automatically composes a set of predefined transformations without requiring task-specific policy search (Cubuk et al., 2020), while CutMix constructs a training image by replacing one region with a patch from another image (Yun et al., 2019). In FSCIL, the Tripartite Weight-Space Ensemble uses CutMix-augmented data for knowledge distillation during incremental learning (Lee et al., 2025). These transformations are inexpensive and broadly applicable, but they operate within the observed visual space and provide limited control over category-specific variations. Their diversity is particularly limited when a novel class has only one or a few images.

Diffusion models provide a more expressive mechanism for generating synthetic data. Latent Diffusion Models perform denoising in a compressed latent space and support high-quality generation under semantic conditioning (Rombach et al., 2022). Adding diffusion-generated images to real training data can improve ImageNet classification (Azizi et al., 2023). However, text-to-image generation conditioned only on class-level text is strongly influenced by the generator's pretrained distribution. With only a class name or textual description as supervision, generated images may correspond weakly to observed few-shot instances and may contain class-irrelevant changes.

Instruction-based image editing provides an alternative by conditioning image generation on both a source image and a textual instruction. Editing models, such as InstructPix2Pix, enable semantic modifications while preserving content unrelated to the requested change (Brooks et al., 2023). KICA adopts this editing paradigm for FSCIL augmentation, where the key challenge is not image editing itself but identifying meaningful class-specific variations under limited observations. It constructs a class attribute base from explicit and implicit knowledge, converts attributes into structured editing instructions, and modifies one variation dimension at a time: background, class attribute, or viewpoint. Unlike existing editing approaches that focus on general-purpose image manipulation, KICA uses knowledge-guided editing to expand support sets around observed instances.

3 Method

3.1 Problem Setup

Let $\{\mathcal{D}^{(t)}\}_{t=0}^{T}$ denote an FSCIL sequence with one base session and T incremental sessions. The base session $\mathcal{D}^{(0)}$ contains abundant training samples, whereas each incremental session $\mathcal{D}^{(t)}$, $t \geq 1$, contains K labeled images per novel class (i.e., a K-shot setting) in the class set $\mathcal{C}^{(t)}$. The class sets are disjoint across sessions, i.e., $\mathcal{C}^{(t)} \cap \mathcal{C}^{(s)} = \emptyset$ for $t \neq s$. After session t, the model is evaluated on all classes observed so far without a session identifier.

KICA augments only the novel-class support set before incremental training. It does not modify the downstream FSCIL objective or inference procedure. As shown in Fig. 2, KICA consists of three stages: **Attribute Base Construction** combines explicit knowledge from a lexical database and category attribute co-occurrences with implicit knowledge elicited from a pretrained LLM to build a class-specific attribute base. **Knowledge-Informed Prompt Generation** converts the collected attributes into structured editing instructions for each class. **Editing-based Augmentation** applies these instructions to the original few-shot images and adds the edited samples to the support set.

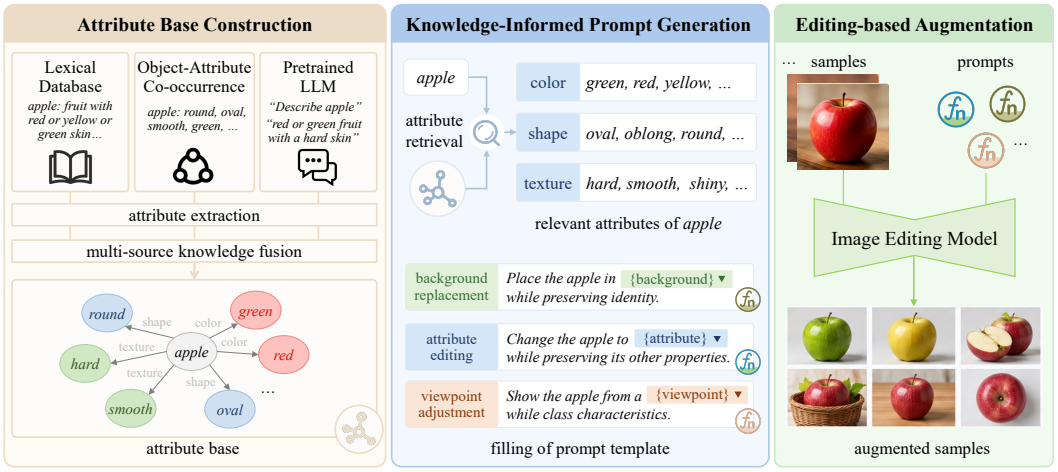


Figure 2: Overview of the KICA pipeline. Explicit and implicit knowledge sources are integrated to construct a class attribute base, providing knowledge cues for exploring plausible intra-class variations beyond limited observations. The extracted attributes are converted into knowledge-informed prompt templates for controllable image editing, producing augmented support samples for FSCIL.

3.2 Attribute Base Construction

We construct a class attribute base from three complementary sources. A *lexical database* and visually grounded *object-attribute co-occurrences* provide explicit knowledge, while a *pretrained LLM* provides knowledge elicited from its implicit semantic representations. The attributes describe visual dimensions of intra-class variations, including color, shape, and texture, providing class-specific cues for controllable image editing.

Lexical Database. We map each class name c to its WordNet (Miller, 1995) synset and retrieve the corresponding lexical definition. WordNet provides concise, standardized descriptions for many object classes. For example, it defines an apple as a "*fruit with red or yellow or green skin and sweet to tart crisp whitish flesh*." An LLM is then used as an attribute extractor to convert this textual definition into structured attribute-value pairs, such as *color*: {red, yellow, green} and *texture*: {crisp, smooth}. Let d denote an attribute dimension and E_d denote the dimension-specific LLM extractor. Given the WordNet definition s_c^{lex} of class c, the extractor returns

$$\mathcal{A}_{c,d}^{\text{lex}} = E_d(s_c^{\text{lex}}), \tag{1}$$

where $\mathcal{A}_{c,d}^{\text{lex}}$ denotes the attribute set extracted from the lexical source for dimension d. If no reliable synset is available, the lexical source contributes no values for that class, and the remaining sources provide complementary attribute knowledge. This limitation is particularly common for fine-grained categories, such as bird species in CUB.

Object-Attribute Co-Occurrence. Attributes frequently associated with an object class provide visual cues about its typical appearance. We construct object-attribute associations by aggregating the attribute annotations in Visual Genome (Krishna et al., 2017). These associations complement lexical definitions with visually grounded object properties observed in real images. They define an attribute set $\mathcal{A}_{c,d}^{\text{vg}}$ for each applicable visual dimension. The proposed framework is compatible with alternative visual knowledge resources that provide richer attribute annotations.

Pretrained Large Language Model (LLM). As a complementary knowledge source, we query a pretrained LLM for typical visual descriptions of class c. Unlike the concise WordNet definitions, these descriptions characterize diverse appearances of possible class instances. For example, "*A photo of an apple shows a round, red fruit with a green stem*" provides properties of shape, color, and fine-grained appearance. Applying the same attribute extractors to these descriptions yields the LLM-derived attribute set $\mathcal{A}_{c,d}^{\text{llm}}$.

The final attribute set is the union of the three sources:

$$\mathcal{A}_{c,d} = \mathcal{A}_{c,d}^{\text{lex}} \cup \mathcal{A}_{c,d}^{\text{vg}} \cup \mathcal{A}_{c,d}^{\text{llm}}, \tag{2}$$

where $\mathcal{A}_{c,d}$ denotes the combined attribute set of class c for dimension d. The union aggregates attributes from different sources while removing exact duplicate entries. The resulting attribute base describes properties of class c and supplies class-specific values to the prompt generation stage.

3.3 Knowledge-Informed Prompt Generation

To expand visual coverage while limiting unintended changes, KICA edits one variation dimension of a support image at a time. We consider three types of intra-class variations: background, foreground attributes, and foreground viewpoints, corresponding to *background replacement*, *attribute editing*, and *viewpoint adjustment*[1]. Each operation is implemented with a positive and a negative prompt template, where the positive prompt specifies the desired modification and the negative prompt constrains unrelated changes. We elaborate on the positive prompts first:

For *background replacement*, a class-compatible scene description suggests by an LLM is inserted into the background fields of the positive template. For *attribute editing*, an attribute dimension d and a target value $a \in \mathcal{A}_{c,d}$ are selected. The generated instruction modifies only the selected attribute while preserving other class characteristics. For example, selecting *green* from the color attributes of *apple* generates an instruction that changes the apple color to green while maintaining its identity. For *viewpoint adjustment*, KICA uses a manually defined set of viewpoint candidates shared across classes, including side, top-down, and 45-degree views etc. Therefore, semantic knowledge guides variations requiring class compatibility, whereas viewpoint candidates are class-independent.

To construct negative prompts, an LLM identifies class-relevant attributes that should remain unchanged and inserts them into the negative template together with general constraints on deformation, cropping, low resolution, and other artifacts. For *attribute editing*, the modified dimension is excluded from the preservation constraints to avoid conflicting instructions. The resulting positive and negative prompts are provided separately to the image editing model, which performs the requested edits while preserving content unrelated to the target variation.

3.4 Editing-based Augmentation

Image augmentation expands the support set without changing the downstream FSCIL method. Let $\mathcal{I}_c = \{(x_i, c)\}_{i=1}^{K}$ denote the K-shot support set of a novel class c. Given input image x_i and a selected editing operation o, prompt generation produces a positive prompt $p_{i,o}^{+}$ and a negative prompt $p_{i,o}^{-}$. The image editing model is denoted by G and produces

$$\widetilde{x}_{i,o} = G(x_i, p_{i,o}^{+}, p_{i,o}^{-}). \tag{3}$$

In Eq. (3), $\widetilde{x}_{i,o}$ denotes the edited image generated by applying operation o to source image x_i.

Each edited image inherits its source label, and all edited images are retained without additional selection. Let $\widetilde{\mathcal{I}}_c$ denote the edited set. The support set used for incremental learning is

$$\mathcal{I}_c^{\text{aug}} = \mathcal{I}_c \cup \widetilde{\mathcal{I}}_c. \tag{4}$$

The augmented support set is passed directly to the downstream FSCIL method without changing its architecture or learning objective.

[1]The detailed prompt templates are provided in the Appendix.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Datasets and Incremental Protocols. We follow the standard FSCIL protocol (Tao et al., 2020) on CIFAR-100 (Krizhevsky et al., 2009), miniImageNet (Vinyals et al., 2016), and CUB (Wah et al., 2011). CIFAR-100 and miniImageNet use 60 classes in the base session and distribute the remaining 40 classes over eight 5-way incremental sessions. CUB uses 100 base classes and distributes the remaining 100 classes over ten 10-way incremental sessions.

Compared Methods and Baselines. We evaluate augmentation methods with three representative FSCIL methods spanning different architectures: *CLOSER* (Oh et al., 2024) (CNN-based), *LGSP* (Jiang et al., 2025) (ViT-based), and *BiMC* (Chen et al., 2025) (CLIP-based).

For each downstream FSCIL method, we compare the unaugmented baseline (*Base*) with six augmentation strategies. Each strategy adds 20 images per novel class. The compared strategies include *RandAugment* (Cubuk et al., 2020), *CutMix* (Yun et al., 2019), *T2I Synthesis*, *T2I Synthesis (w/ knowledge)*, *I2I Editing*, and *I2I Editing (w/ knowledge, KICA)*. The two T2I variants use Qwen-Image, with the former prompted only by the class name and the latter using prompts derived from KICA knowledge. The two I2I variants use Qwen-Image-Edit (Wu et al., 2025), with the latter additionally incorporating class knowledge into the editing process.

Evaluation Metrics. Following standard FSCIL evaluation (Tao et al., 2020), we report Top-1 accuracy using two summary metrics: Last-Session Accuracy (A_{last}), which measures the final-session accuracy over all observed classes; and Average Accuracy (A_{avg}), which averages the accuracy across the base session and all incremental sessions.

Implementation Details. Following the augmentation protocol in Sec. 3.4, each novel class receives 10 *background replacements*, 5 *attribute edits*, and 5 *viewpoint adjustments*. Edited images are processed using the same resizing and cropping procedures as the original images. For *CLOSER*, *LGSP*, and *BiMC*, we follow the original training protocols and hyperparameters and only modify the incremental support sets while keeping the architectures and objectives unchanged. Results are averaged over repeated runs with paired random seeds across augmentation variants. Offline image editing is performed on an NVIDIA A800 GPU, while downstream training and evaluation are implemented in PyTorch on NVIDIA RTX 3090 GPUs.

4.2 BENCHMARK COMPARISON

Overall Performance Analysis. Tab. 1 reports the last-session and average accuracy of representative FSCIL methods. KICA achieves the best performance in 17 of the 18 combinations of datasets, FSCIL methods, and evaluation metrics, with the remaining case showing only a marginal gap compared with the best-performing method. These results demonstrate the generality of KICA across different FSCIL paradigms, including prototype learning (Oh et al., 2024), prompt tuning (Jiang et al., 2025), and vision-language model calibration Chen et al. (2025). The largest gains are observed on CUB, which may be attributed to its fine-grained nature, where additional class-specific variations provide more informative supervision under limited observations.

Ablation of Knowledge Guidance and Editing Strategies. We further investigate the contributions of knowledge guidance and editing-based augmentation through a controlled 2×2 comparison. As shown in Tab. 1, the performance generally follows the trend:

$$\boxed{\text{T2I} \rightarrow \text{T2I+Knowledge} \rightarrow \text{I2I Editing} \rightarrow \textbf{I2I Editing+Knowledge (KICA)}}$$

This trend highlights three complementary factors: knowledge-guided variation expansion, image-conditioned editing, and their joint effect.

First, incorporating class-aware knowledge improves both T2I synthesis and I2I editing, indicating that semantic knowledge provides additional guidance for exploring plausible intra-class variations beyond limited observations. Since FSCIL offers only a few examples for novel classes, such knowledge-guided expansion provides additional supervision under scarce visual observations.

Table 1: Last-session accuracy (A_{last}) and average accuracy (A_{avg}) under the 5-shot FSCIL protocol. Best results within each block of downstream FSCIL methods are shown in **bold**.

	Method	CIFAR-100		miniImageNet		CUB	
		A_{last}	A_{avg}	A_{last}	A_{avg}	A_{last}	A_{avg}
CNN -based	CLOSER (Oh et al., 2024)	52.49	62.55	53.44	63.28	63.10	68.02
	+ CutMix	50.48	61.63	52.60	62.13	60.65	67.75
	+ RandAugment	52.22	62.70	53.44	63.28	62.48	68.62
	+ T2I Synthesis	49.49	60.63	50.87	61.58	44.60	58.65
	+ T2I Synthesis (w/ knowledge)	50.75	61.83	52.66	62.58	55.02	64.39
	+ I2I Editing	53.25	63.35	54.82	63.85	64.24	69.61
	+ I2I Editing (w/ knowledge, KICA)	**53.96**	**63.78**	**55.56**	**64.21**	**64.95**	**70.85**
ViT -based	LGSP (Jiang et al., 2025)	76.85	80.90	81.18	85.90	76.77	78.94
	+ CutMix	77.88	82.43	87.61	88.83	77.06	79.69
	+ RandAugment	75.59	80.49	87.91	89.55	76.41	78.72
	+ T2I Synthesis	75.52	79.55	87.85	89.65	60.93	71.06
	+ T2I Synthesis (w/ knowledge)	79.33	82.50	87.96	**90.23**	72.21	75.73
	+ I2I Editing	79.09	82.37	87.90	89.77	76.42	78.78
	+ I2I Editing (w/ knowledge, KICA)	**79.82**	**83.52**	**88.73**	90.07	**80.15**	**81.19**
CLIP -based	BiMC (Chen et al., 2025)	72.00	72.50	93.20	91.93	78.00	71.99
	+ CutMix	68.80	71.45	89.40	91.31	69.00	67.19
	+ RandAugment	65.00	71.79	93.20	91.93	78.00	71.45
	+ T2I Synthesis	55.00	69.54	90.00	89.56	56.33	59.80
	+ T2I Synthesis (w/ knowledge)	57.40	70.82	91.00	90.93	59.67	64.08
	+ I2I Editing	66.60	72.00	93.80	91.88	75.67	72.85
	+ I2I Editing (w/ knowledge, KICA)	**73.26**	**76.20**	**94.00**	**92.47**	**79.67**	**76.72**

Second, I2I editing outperforms T2I synthesis, demonstrating the advantage of modifying existing support images rather than generating samples from scratch. By retaining visual cues from the original exemplars, editing-based augmentation maintains stronger instance-level grounding while expanding the diversity of the support set.

Finally, combining knowledge guidance with image editing achieves the best performance across FSCIL frameworks, suggesting that knowledge provides semantically meaningful variation directions, while image editing anchors such variations to observed instances. Together, they produce augmented samples with broader variation coverage and stronger alignment with the original data.

4.3 Distributional Analysis of Augmented Samples

The analysis above shows that knowledge guidance and image editing provide complementary benefits for FSCIL. To further understand the underlying mechanism, we analyze how different augmentation strategies affect the distribution of the augmented support set. Specifically, we evaluate the coverage and distributional alignment of generated samples under different augmentation settings.

Quantitative results are reported in Tab. 2, with metrics computed using SwAV and DINOv2 features, while the complete evaluation with additional metrics is provided in the Appendix. We mainly focus on *coverage* (Naeem et al., 2020) and *adjusted IRS* (Dombrowski et al., 2025), which characterize distribution expansion and alignment, respectively. *Coverage* is emphasized because FSCIL augmentation primarily aims to recover missing intra-class variations caused by limited observations, whereas *adjusted IRS* measures whether the expanded distribution remains aligned with real samples.

Knowledge-informed strategies achieve higher *coverage* scores while maintaining comparable distributional alignment, suggesting that the generated samples better complement the missing regions of the original support distribution. The t-SNE visualization in Fig. 3 further illustrates these differences. Conventional augmentation methods mainly introduce local perturbations around observed samples, resulting in relatively limited expansion of the intra-class variation space. In con-

Table 2: Distributional evaluation of different augmentation strategies across FSCIL benchmarks using *coverage* and *adjusted IRS* (IRS for short) based on SwAV and DINOv2 features. Higher values indicate broader coverage of intra-class variations. The highest value is highlighted in **bold**.

Feature	Method	CIFAR-100		miniImageNet		CUB	
		Coverage	IRS	Coverage	IRS	Coverage	IRS
SwAV	CutMix	0.3358	0.1227	0.4395	0.1042	0.6943	0.5417
	RandAugment	0.4008	0.1318	0.3620	0.1230	0.6999	0.6167
	T2I Synthesis	0.3467	0.1255	0.1805	0.1275	0.7425	0.5714
	T2I Synthesis (w/ knowledge)	0.3525	0.1445	0.1960	0.1323	0.7329	0.6058
	I2I Editing	0.4183	0.1795	0.3502	0.1700	0.7748	0.7285
	I2I Editing (w/ knowledge, KICA)	**0.4800**	**0.2812**	**0.4317**	**0.3178**	**0.7845**	**0.7590**
DINOv2	CutMix	0.4155	0.1253	0.3930	0.1227	0.6403	0.7293
	RandAugment	0.3633	0.1143	0.3785	0.1115	0.6731	0.7072
	T2I Synthesis	0.3595	0.1135	0.4002	0.1337	0.6204	0.7094
	T2I Synthesis (w/ knowledge)	0.3478	0.1155	0.4125	0.1293	0.6447	0.7646
	I2I Editing	0.4412	0.1998	0.4423	0.1547	**0.7503**	0.8242
	I2I Editing (w/ knowledge, KICA)	**0.4785**	**0.2975**	**0.5320**	**0.3495**	0.7262	**1.2345**

trast, KICA produces broader variations with diverse editing operations by incorporating class-level knowledge beyond the observed examples, which is consistent with Fig. 1. Notably, KICA achieves the most effective coverage expansion among the evaluated augmentation strategies. This improvement suggests that combining class-level knowledge with image-conditioned editing enables augmentation beyond local perturbations while maintaining consistency with observed samples.

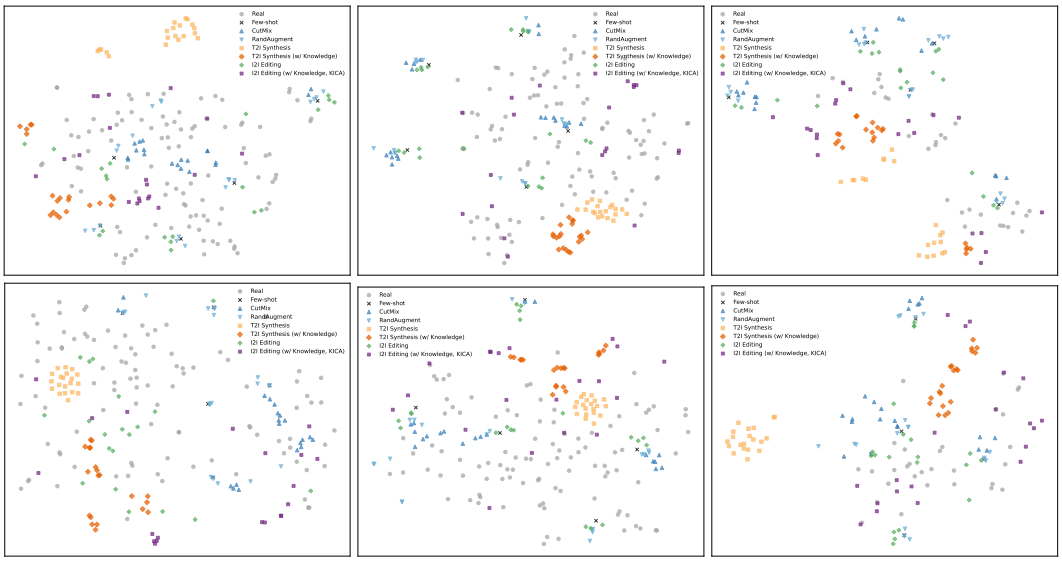


Figure 3: t-SNE visualization of feature distributions under different augmentation strategies. Each column corresponds to one dataset: CIFAR-100 (*plain* and *raccoon*), miniImageNet (*bolete* and *hotdog*), and CUB (*Black Tern* and *Black-and-White Warbler*). Different colors and markers indicate different augmentation methods. Magnify for better view.

4.4 Extra Results and Analyses

Class Attribute Base Analysis. We first analyze the scale and composition of the constructed attribute base. As illustrated in Fig. 4 and summarized in Tab. 3, the attribute base covers a large number of category-related attributes across different semantic dimensions, providing broad prior knowledge for controllable augmentation. Beyond the scale, different knowledge sources exhibit distinct characteristics. Lexical databases provide more reliable but relatively limited attributes due to their curated nature, whereas LLMs and visual attribute statistics offer broader coverage

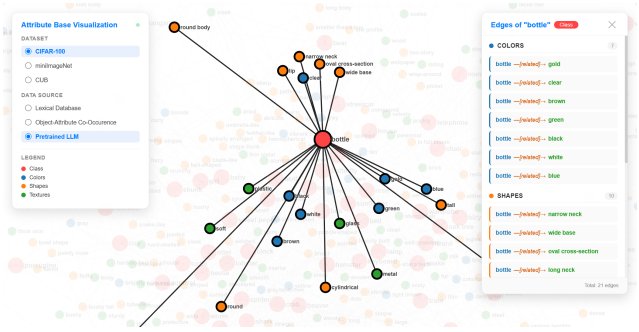


Figure 4: Visualization of the attribute base. The class *bottle* is selected and its associated attributes are highlighted.

Table 3: The number of attribute items across color, shape, and texture for CIFAR-100, miniImageNet, and CUB, respectively.

Attr.	CIFAR	mini	CUB
Colors	5,547	3,416	1,117
Shapes	6,529	4,186	1,563
Textures	8,665	1,159	305
Total	20,741	8,761	2,985

with more diverse candidates. However, automatic extraction may introduce noisy or redundant attributes. Since the downstream augmentation results already demonstrate the effectiveness of the constructed knowledge base, improving attribute validation and redundancy-aware organization remains an interesting direction for future work.

Efficiency. KICA introduces an offline preprocessing cost due to generative image editing. This cost is incurred only once before incremental training and is independent of the downstream FSCIL method. As reported in Tab. 4, editing the 800 images required for CIFAR-100 and miniImageNet takes approximately 20 hours, while editing the 2,000 images for CUB takes approximately 66.7 hours. The average editing time is around 90 seconds per image for CIFAR-100 and miniImageNet and 120 seconds per image for CUB. Since KICA only augments the support set without modifying the FSCIL architecture, it introduces no additional model parameters or inference-time computation.

Table 4: Offline image editing cost of KICA on an A800 GPU. The table reports the number of edited images, total editing time, time per image, memory use, and disk use for each dataset.

Dataset	Images	Time (h)	s/image	Mem. (GB)	Disk (MB)	MB/image
CIFAR-100	800	20.0	90	76.23	932.0	1.17
miniImageNet	800	20.0	90	76.60	968.0	1.21
CUB	2000	66.7	120	76.58	2335.2	1.17

Limitations. KICA relies on the availability and quality of external knowledge to identify plausible intra-class variations. Although the current attribute base integrates multiple knowledge sources, the extracted attributes are still represented at a relatively coarse semantic level and do not explicitly model fine-grained part-level relationships or structural compositions. More advanced visual knowledge representations could further improve the discovery of class-specific variations. In addition, the quality of generated samples is bounded by the capability of the underlying image editing model, and improving controllability for complex semantic changes remains an important direction.

5 Conclusion

Few-Shot Class-Incremental Learning (FSCIL) remains challenging because limited observations provide incomplete coverage of novel-class variations. We propose Knowledge-Informed Controllable Augmentation (KICA), a data-level framework that enriches few-shot support sets for FSCIL by combining class-level knowledge with observed visual instances. Through a class attribute base and controllable image editing, KICA generates diverse yet class-consistent variations without modifying downstream FSCIL architectures. Experiments demonstrate that KICA alleviates the limited observation problem by expanding intra-class variations in few-shot incremental learning. This work highlights the potential of integrating external knowledge with visual evidence to enable more robust adaptation under limited observations.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks: editing the paper to improve readability. We have not used generative AI tools for any tasks with required disclosure, and the remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. The authors carefully checked all LLM-edited text to ensure it accurately reflects our original meaning and claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility and to ensure the authenticity of our proposed method, we provide the core implementation of the key algorithm in the supplementary materials. This code covers attribute base visualization, image editing, t-SNE visualization, and coverage calculation, which we believe is sufficient for readers to verify the novelty and correctness of our approach. Due to the risk of misuse, we are unable to release the complete codebase and all experimental scripts at this time. We will release the full implementation upon acceptance. Detailed descriptions of the experimental configurations are provided in Section 4.1 and Appendix A.1 to facilitate reproduction.

REFERENCES

Noor Ahmed, Anna Kukleva, and Bernt Schiele. OrCo: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28762–28771, 2024.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves ImageNet classification. *Transactions on Machine Learning Research*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Yiyang Chen, Tianyu Ding, Lei Wang, Jing Huo, Yang Gao, and Wenbin Li. Enhancing few-shot class-incremental learning via training-free bi-level modality calibration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9881–9890, 2025.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18613–18624, 2020.

Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3029–3039, 2025.

Yongwei Jiang, Yixiong Zou, Yuhua Li, and Ruixuan Li. Revisiting pool-based prompt learning for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1303–1313, 2025.

Junsu Kim, Yunhoe Ku, Dongyoon Han, and Seungryul Baek. Beyond synthetic replays: Turning diffusion features into few-shot class-incremental learning knowledge. *arXiv preprint arXiv:2503.23402*, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Juntae Lee, Munawar Hayat, and Sungrack Yun. Tripartite weight-space ensemble for few-shot class-incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15329–15338, 2025.

George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning (ICML)*, pp. 7176–7185, 2020.

Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. Closer: Towards better representation learning for few-shot class-incremental learning. In *European Conference on Computer Vision (ECCV)*, pp. 18–35, 2024.

Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23881–23890, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12183–12192, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3630–3638, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15060–15076, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 6023–6032, 2019.

A Appendix

A.1 More Experimental Details

Implementations of Compared Methods. We evaluate our augmentation method on three representative FSCIL approaches using their official implementations: CLOSER (Oh et al., 2024) [2],

[2] https://github.com/JungHunOh/CLOSER_ECCV2024

LGSP (Jiang et al., 2025) [3], and BiMC (Chen et al., 2025) [4]. To incorporate the generated images, we modify only the data loading pipeline for the incremental sessions, while leaving the original model architectures and training procedures unchanged. Specifically, for each novel class, the original five support images are retained and supplemented with 20 additional images loaded from the corresponding augmentation directory, resulting in an expanded support set of 25 images per class. Except for this modification to the incremental support set, all training configurations, including optimization settings, learning rates, training schedules, and other method-specific hyperparameters, strictly follow the official implementations without further tuning. This setup ensures that the performance differences are attributable to the augmented training samples rather than changes in the downstream FSCIL methods.

Details of the prompts for KICA. We construct three prompt templates for image editing augmentation: *background replacement*, *attribute editing*, and *viewpoint adjustment* (Fig. 5). Each template specifies the intended modification while explicitly constraining the remaining visual properties of the input image. The final editing prompts are instantiated from these templates using class specific information and image dependent descriptions:

- **Background Replacement**. This template modifies the scene background while preserving the foreground object, including its category identity, pose, appearance, and key visual characteristics. The prompt specifies the target background theme, scene details, and color tone, while encouraging consistency in illumination, shadows, and depth of field. The corresponding negative prompt suppresses unwanted changes to the foreground object, such as deformation, pose variation, color alteration, or structural modification.
- **Attribute Editing**. This template modifies a specified visual attribute of the foreground object while keeping the remaining object and scene properties unchanged. In our experiments, we mainly use color as the editable attribute. The prompt specifies the target attribute value and requires the original pose, shape, texture, background, lighting, and composition to be preserved. The negative prompt further discourages structural changes, background modification, and other unintended visual variations.
- **Viewpoint Adjustment**. This template changes the viewpoint of the foreground object while preserving its semantic identity and other visual properties. The prompt specifies a target viewpoint, such as a side view, top down view, or three quarter view, and constrains the object color, texture, shape, and other characteristic details to remain consistent with the input image. The background is also kept unchanged. The negative prompt suppresses modifications unrelated to the viewpoint transformation.

For each input image, the placeholders in these templates are instantiated with the corresponding subject name, object state, key visual features, and the target editing specification. These instantiated prompts are then provided to Qwen-Image-Edit to generate the augmented samples. Generated samples are shown in Fig. 6. Starting from the initial 5-shot images, we generate four additional images for each image using Qwen-Image-Edit, resulting in an expanded set of 20 images.

Details of the prompts for KICA variants. Below, we show how the prompts are constructed for three KICA variants *T2I Synthesis*, *T2I Synthesis (w/ knowledge)*, and *I2I Editing*.

- **T2I Synthesis.** We use the Qwen-Image image generation model to synthesize 20-shot images with the prompt `a photo of {class}`.
- **T2I Synthesis (w/ knowledge).** We adopt the knowledge-enhanced prompts from KICA. The reference image component is removed, while the remaining slots are filled with the corresponding knowledge attributes following the original template. Specifically, two groups are generated based on *background replacement*, while *attribute editing* and *viewpoint adjustment* each contribute one group. For each group, one prompt is used to generate a batch of five images, resulting in 20-shot images in total.
- **I2I Editing.** We use the prompt `Create a different photo of the same {class}. {photo} based on this image.` Starting from the initial 5-shot

[3] https://github.com/Jywsuperman/LGSP

[4] https://github.com/yychen016/BiMC

Background replacement | Attribute Editing | Viewpoint Adjustment

Positive Prompt

A {Subject_Name} in {Subject_Action_or_State}, detailed {Key_Feature} and original pose preserved exactly as in the input image. Replace the background with a {Background_Theme}: {Background_Details}. Ensure the environmental lighting direction aligns with the shadows on the {Subject_Name}, creating a realistic interaction. Maintain natural depth of field: {Subject_Name} sharp in focus, background slightly blurred to highlight the subject. Use {Color_Tone} photorealistic colors. Output must have same dimensions and aspect ratio as input.

Negative Prompt

{Subject_Name} modification, changed {Key_Feature}, blurry {Subject_Name}, deformed anatomy, extra limbs, changed pose, cropped subject, altered color of the {Subject_Name}, low resolution, text, watermark, frame, border, cartoon, painting, sketch, floating subject (unless intended), mismatched lighting, harsh shadows, over-saturated colors, plastic look, cutout effect, background elements overlapping the {Subject_Name}, different species/model, resized output, bad composition, hallucinated details on {Subject_Name}.

Background replacement | Attribute Editing | Viewpoint Adjustment

Positive Prompt

A {Subject_Name} in {Subject_Action_or_State}, with only the color changed to {Target_Color}, while all other visual attributes—including {Key_Feature}, pose, shape, texture, lighting, shadows, and composition—preserved exactly as in the input image. Maintain original background, depth of field, and photorealistic style. Ensure the subject remains sharp in focus, with consistent environmental lighting direction and natural shadow alignment. Output must have the same dimensions and aspect ratio as the input.

Negative Prompt

{Subject_Name} modification beyond color change, altered {Key_Feature}, changed pose, deformed anatomy, extra or missing parts, cropped subject, resized output, low resolution, blurry {Subject_Name}, text, watermark, frame, border, cartoon, painting, sketch, floating subject, plastic look, cutout effect, mismatched lighting, harsh shadows, over-saturated or unnatural colors (except the specified {Target_Color}), background alteration, background elements overlapping the subject, hallucinated details, bad composition, different species/model, or any other structural or contextual changes.

Background replacement | Attribute Editing | Viewpoint Adjustment

Positive Prompt

A {Subject_Name} in {Subject_Action_or_State}, viewed from a {Target_Viewpoint} perspective (e.g., side view, top-down view, 45-degree angle), with all other attributes—including color, {Key_Feature}, shape, texture, size, lighting direction, shadows, and surface details—preserved exactly as in the input image. Maintain the original background unchanged, natural depth of field ({Subject_Name} sharp, background slightly blurred), and photorealistic rendering. Ensure consistent lighting that matches the new viewpoint while respecting the original shadow logic. Output must have the same dimensions and aspect ratio as the input.

Negative Prompt

{Subject_Name} with altered color, modified {Key_Feature}, changed material or texture, deformed anatomy, extra or missing parts, resized or scaled subject, cropped composition, low resolution, blurry {Subject_Name}, text, watermark, frame, border, cartoon, painting, sketch, plastic look, cutout effect, floating subject (unless intended by viewpoint), background alteration, background elements overlapping the subject, mismatched lighting inconsistent with the new viewpoint, harsh or unrealistic shadows, over-saturation, hallucinated details, wrong species/model, or any structural, stylistic, or contextual changes beyond the specified viewpoint transformation.

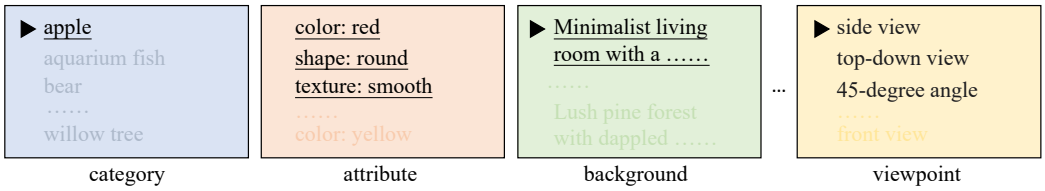


Figure 5: Prompt templates and the candidate values for different placeholders. Class-aware knowledge and predefined viewpoint candidates are inserted into operation-specific positive templates. Negative templates constrain changes to unrelated visual content. For *attribute editing*, color change is shown as an example. Colored placeholders indicate fields filled with candidate values in boxes of the corresponding colors, whereas uncolored placeholders are completed by an LLM.

Table 5: Distributional evaluation of different augmentation strategies on CIFAR-100 using *PRDC* and *adjusted IRS* (IRS for short) based on SwAV and DINOv2 features.

Feature	Method	Precision	Recall	Density	Coverage	IRS
SwAV	CutMix	0.6833	0.6808	0.5060	0.3358	0.1227
	RandAugment	0.8100	0.6233	0.9500	0.4008	0.1318
	T2I Synthesis	0.6533	0.8642	0.5287	0.3467	0.1255
	T2I Synthesis (w/ knowledge)	0.6267	0.7950	0.5167	0.3525	0.1445
	I2I Editing	0.8100	0.6975	0.7160	0.4183	0.1795
	I2I Editing (w/ knowledge, KICA)	0.7000	0.8883	0.6147	0.4800	0.2812
DINOv2	CutMix	0.8850	0.4910	1.0696	0.4155	0.1253
	RandAugment	0.8990	0.5865	0.9622	0.3633	0.1143
	T2I Synthesis	0.8100	0.8282	0.8888	0.3595	0.1135
	T2I Synthesis (w/ knowledge)	0.8300	0.7967	0.7650	0.3478	0.1155
	I2I Editing	0.8380	0.6960	0.9258	0.4412	0.1998
	I2I Editing (w/ knowledge, KICA)	0.7860	0.8815	0.7152	0.4785	0.2975

images, four additional images are generated from each image, resulting in 20-shot images in total.

Details of the distributional analysis. We use the official repositories of IRS[5] and PRDC[6]. We follow the official implementations and use their default hyperparameters. For IRS on CUB, we encountered computational issues where the evaluation occasionally stalled for certain classes. To resolve this issue, we split the CUB test data into two equal subsets and use them as "pathtrain" and "pathtest", respectively, following the recommendation in the official README.

A.2 More Experimental Results

Edited Image Analysis. Fig. 8 shows representative image editing results on the three benchmark datasets. Given an original support image, KICA applies *background replacement*, *attribute editing*, or *viewpoint adjustment* while retaining the object's class identity. The edits vary in appearance and context for the coarse-grained classes in CIFAR-100 and miniImageNet, and for the fine-grained bird classes in CUB. Overall, the results demonstrate strong controllable editing capability, with edits almost exclusively affecting the background, attribute, or viewpoint.

Additional Distributional Analysis. To further validate the distributional analysis, we provide additional evaluations using different feature extractors and metrics. In the main paper, we report *coverage* and *adjusted IRS* based on SwAV and DINOv2 features for clarity. Here, we report the complete PRDC (Naeem et al., 2020) and IRS (Dombrowski et al., 2025) metrics (Tabs. 5 to 7).

Overall, KICA generally achieves stronger coverage compared with other augmentation strategies, indicating that knowledge-informed editing is more effective at expanding the observed support distribution toward missing intra-class variations. Meanwhile, the results also reveal a trade-off between coverage-oriented and fidelity-oriented metrics. For example, methods with higher Precision or Density do not necessarily achieve higher Recall or Coverage, suggesting that local distribution fidelity and global variation expansion capture different aspects of augmented samples.

As pretrained generative models rapidly advance, the visual quality of generated images has improved, making fidelity-oriented metrics less directly correlated with the objective of FSCIL augmentation. While PRDC provides a comprehensive evaluation of generation quality from multiple perspectives, FSCIL primarily benefits from recovering missing class variations and improving the decision boundaries learned from limited observations. Therefore, coverage-related metrics, such as Recall and Coverage, are particularly relevant in this setting, as they measure whether augmented samples expand the support distribution to encompass more class-relevant variations.

As a complementary view of t-SNE and the aforementioned metrics, we show the generated images of KICA and three variants, i.e., *T2I Synthesis*, *T2I Synthesis (w/knowledge)*, and *I2I Editing*. Note

[5] https://github.com/mischad/beyondfid

[6] https://github.com/clovaai/generative-evaluation-prdc

Table 6: Distributional evaluation of different augmentation strategies on miniImageNet using *PRDC* and *adjusted IRS* (IRS for short) based on SwAV and DINOv2 features.

Feature	Method	Precision	Recall	Density	Coverage	IRS
SwAV	CutMix	0.8180	0.3427	1.2626	0.4395	0.1042
	RandAugment	0.7390	0.5607	0.8300	0.3620	0.1230
	T2I Synthesis	0.5710	0.7632	0.4396	0.1805	0.1275
	T2I Synthesis (w/ knowledge)	0.6000	0.7105	0.4414	0.1960	0.1323
	I2I Editing	0.7200	0.5740	0.6010	0.3502	0.1700
	I2I Editing (w/ knowledge, KICA)	0.6580	0.7750	0.6186	0.4317	0.3178
DINOv2	CutMix	0.9010	0.4650	1.0854	0.3930	0.1227
	RandAugment	0.8900	0.6345	1.0346	0.3785	0.1115
	T2I Synthesis	0.8700	0.7067	1.2926	0.4002	0.1337
	T2I Synthesis (w/ knowledge)	0.8980	0.6552	1.1380	0.4125	0.1293
	I2I Editing	0.8990	0.6123	1.0600	0.4423	0.1547
	I2I Editing (w/ knowledge, KICA)	0.8160	0.8122	0.8148	0.5320	0.3495

Table 7: Distributional evaluation of different augmentation strategies on CUB using *PRDC* and *adjusted IRS* (IRS for short) based on SwAV and DINOv2 features.

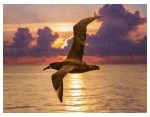
Feature	Method	Precision	Recall	Density	Coverage	IRS
SwAV	CutMix	0.8496	0.4294	0.8475	0.6943	0.5417
	RandAugment	0.8113	0.5861	0.8917	0.6999	0.6167
	T2I Synthesis	0.7145	0.7722	0.6989	0.7425	0.5714
	T2I Synthesis (w/ knowledge)	0.7770	0.7818	0.6156	0.7329	0.6058
	I2I Editing	0.8055	0.6556	0.7314	0.7748	0.7285
	I2I Editing (w/ knowledge, KICA)	0.6373	0.8659	0.5106	0.7845	0.7590
DINOv2	CutMix	0.6605	0.5790	0.6012	0.6403	0.7293
	RandAugment	0.7443	0.6641	0.7735	0.6731	0.7072
	T2I Synthesis	0.4135	0.9237	0.3248	0.6204	0.7094
	T2I Synthesis (w/ knowledge)	0.4842	0.8529	0.3708	0.6447	0.7646
	I2I Editing	0.7291	0.6791	0.7597	0.7503	0.8242
	I2I Editing (w/ knowledge, KICA)	0.5041	0.9258	0.3618	0.7262	1.2345

that although these images have different resolusion from the figure, they will go through the same preprocessing procedure for benchmark comparisons, t-SNE visualization, distributional evaluation etc. Here, we use the non-preprocessed images simply for better view (e.g. images of CIFAR-100 are only 32×32). Visualization of diversity introduced by different augmentation strategies. From left to right, the diversity generally increases. KICA exhibits richer variations in background, viewpoint, and foreground attributes, demonstrating the effectiveness of knowledge-informed augmentation.

Input image



Edited image



Positive prompt:
A Black-footed Albatross in gliding flight, detailed dark brown plumage and long wingspan preserved exactly as in the input image. Replace the background with a dramatic sunset seascape: calm ocean waters reflecting golden sunlight, with scattered purple and orange cumulus clouds in the distance. Ensure the environmental lighting direction aligns with the shadows on the Black-footed Albatross, creating a realistic interaction with warm backlighting. Maintain natural depth of field: Black-footed Albatross sharp in focus, background slightly blurred to highlight the subject. Use warm, golden hour photorealistic colors. Output must have same dimensions and aspect ratio as input.

Negative prompt:
Black-footed Albatross modification, changed dark plumage, white feathers (unless natural), blurry bird, deformed anatomy, extra limbs, changed pose, cropped wings, altered color of the bird, low resolution, text, watermark, frame, border, cartoon, painting, sketch, floating bird without aerodynamic logic, mismatched lighting, harsh shadows, over-saturated colors, plastic look, cutout effect, background elements overlapping the bird, different species, resized output, bad composition, hallucinated details on bird.

Input image



Edited image



Positive prompt:
A jellyfish in gentle drifting state, with only the color changed to electric violet, while all other visual attributes—including translucent bell with visible radial canals, delicate trailing tentacles, and subtle surface sheen, pose, shape, texture, lighting, shadows, and composition—preserved exactly as in the input image. Maintain original background, depth of field, and photorealistic style. Ensure the subject remains sharp in focus, with consistent environmental lighting direction and natural shadow alignment. Output must have the same dimensions and aspect ratio as the input.

Negative prompt:
jellyfish modification beyond color change, altered translucent bell, radial canals, or tentacle morphology, changed pose (e.g., contracted, rigid, or unnaturally coiled), deformed anatomy, extra or missing tentacles, cropped subject, resized output, low resolution, blurry jellyfish, text, watermark, frame, border, cartoon, painting, sketch, floating subject (unless consistent with natural buoyancy), plastic look, cutout effect, mismatched lighting, harsh shadows, over-saturated or unnatural colors (except the specified electric violet), background alteration, background elements overlapping the subject, hallucinated details (e.g., eyes, mouth, internal organs, glowing patterns not in original), bad composition, different species/model (e.g., Portuguese man o' war, salp, sea nettle with different structure), or any other structural or contextual changes.

Input image



Edited image



Positive prompt:
An American Redstart perched on a branch, viewed from a front view perspective, with all other attributes—including color, yellow shoulder and chest patches, shape, texture, size, lighting direction, shadows, and surface details—preserved exactly as in the input image. Maintain the original background unchanged, natural depth of field (American Redstart sharp, background slightly blurred), and photorealistic rendering. Ensure consistent lighting that matches the new viewpoint while respecting the original shadow logic. Output must have the same dimensions and aspect ratio as the input.

Negative prompt:
American Redstart with altered color, modified yellow shoulder and chest patches, changed material or texture, deformed anatomy, extra or missing parts, resized or scaled subject, cropped composition, low resolution, blurry American Redstart, text, watermark, frame, border, cartoon, painting, sketch, plastic look, cutout effect, floating subject (unless intended by viewpoint), background alteration, background elements overlapping the subject, mismatched lighting inconsistent with the new viewpoint, harsh or unrealistic shadows, over-saturation, hallucinated details, wrong species/model, or any structural, stylistic, or contextual changes beyond the specified viewpoint transformation.

Figure 6: Examples of the final prompt and the corresponding edited image based on the *background replacement*, one *attribute editing*, and one *viewpoint adjustment* templates of KICA.



Figure 7: Representative KICA augmentations on CIFAR-100, miniImageNet, and CUB. From left to right, each row shows the original support image, two *background replacements*, one *attribute edit*, and one *viewpoint adjustment*. "bg" is short for background.



Figure 8: Examples of generated images of KICA and three variants, i.e., *T2I Synthesis*, *T2I Synthesis (w/knowledge)*, and *I2I Editing*. The classes are not cherry-picked.