DIRECTION-AWARE NEGATIVE REPULSION FOR REPLAY-FREE CLIP CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based continual learning can benefit from strong pre-trained representations and often learns sequential tasks without storing old samples. A common choice is task-local cross-entropy, where the softmax contains only classes from the current task. This objective pulls an image toward its ground-truth class text and pushes it away from other current-task class texts, while old classes receive no direct negative gradients. However, we observe that the distance between old image features and their corresponding text features can still increase after learning new tasks. This indicates that current-task optimization is still associated with changes in old cross-modal representations. We find that this effect is closely related to updates of the shared image encoder, and that different negative updates have very different effects on old tasks. In particular, negatives with stronger overlap with old-task-sensitive directions exhibit higher interference risk. To capture this directional sensitivity, we construct a sample-normalized Average Gradient Outer Product (AGOP) for each old task and use it to measure the interference risk of current-task negatives. Based on this idea, we propose **A**GOP-**G**uided **R**epulsion **A**ttenuation (AGRA), which preserves low-risk negative signals while selectively attenuating high-risk ones. Extensive experiments show that AGRA outperforms a wide range of CLIP-based continual learning methods and better preserves zero-shot generalization.

1 INTRODUCTION

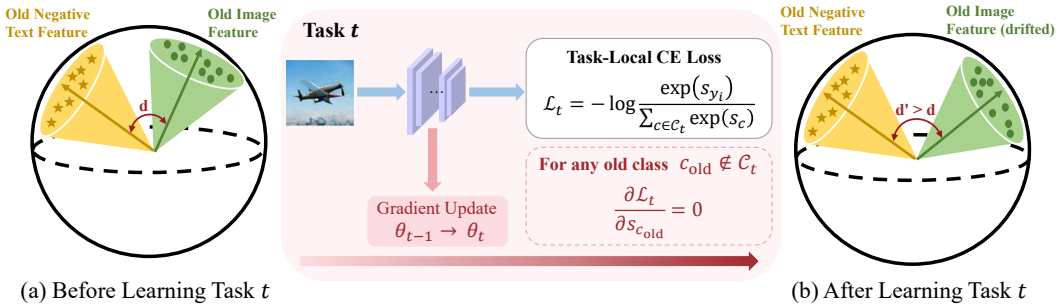
Continual learning aims to acquire new knowledge over time while preserving previously learned knowledge. Large-scale pre-trained models provide a strong starting point for this goal because they already contain rich and transferable representations. Among them, CLIP (Radford et al., 2021) shows strong transferability and zero-shot generalization across many downstream tasks. These properties have made CLIP-based continual learning an increasingly important direction (Zheng et al., 2023; Huang et al., 2024; Yu et al., 2024; Marczak et al., 2024; Xie et al., 2026).

Existing CLIP-based continual learning methods often learn sequential tasks without storing old samples. A common choice is *task-local cross-entropy* (Huang et al., 2025; Gong et al., 2026; Wang & Guo, 2026), where the softmax contains only classes from the current task. For each image, the objective increases its similarity to the ground-truth class text and decreases its similarities to the other current-task class texts. We refer to the latter as *negative repulsion*. Since old classes are excluded from the softmax, they are never directly involved in this repulsion and receive no direct negative gradients.

Recent work shows that such negative repulsion plays an important role in the change of CLIP's cross-modal geometry (Liang et al., 2022; Huang et al., 2025). In particular, the similarity between an image and its ground-truth (positive) class text quickly becomes stable, while its similarities to non-matching (negative) class texts continue to decrease during training (Huang et al., 2025). This motivates us to examine how the image–text relation of previously learned tasks changes when a new task is learned. Interestingly, although old classes are absent from the current loss, the gap between old image features and their text features still increases. Thus, removing old classes from the softmax does not fully isolate old cross-modal representations from changes during current-task optimization.

We find that this change is closely related to updates of the shared image encoder. The repulsion from negative classes in the current task produces feature gradients that update the encoder shared by all tasks. These gradients can therefore change old image representations even though old classes

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.



(a) Before Learning Task t (b) After Learning Task t

Figure 1: Task-local CE removes direct competition with old classes, yet shared encoder updates can still drift old image features and change their cross-modal geometry ($d' > d$).

do not appear in the current loss. However, not all current-task negatives have the same effect. Negatives with similar gradient strengths can be associated with very different levels of old-task feature drift. Our analysis shows that this difference is closely related to direction: a negative has higher interference risk when its gradient is more aligned with feature directions that are important to old tasks.

To characterize these directions, we build on the Average Gradient Outer Product (AGOP) (Radhakrishnan et al., 2024). We construct a sample-normalized AGOP for each old task to describe its directional sensitivity while reducing the influence of gradient magnitude. The interference risk of a current-task negative is then determined by both its gradient strength and its overlap with old-task-sensitive directions. Based on this insight, we propose **A**GOP-**G**uided **R**epulsion **A**ttenuation (AGRA). AGRA keeps low-risk negative signals unchanged and selectively attenuates high-risk negative repulsion. Extensive experiments show that AGRA outperforms a wide range of CLIP-based continual learning methods and better preserves CLIP's zero-shot generalization.

Our contributions are summarized as follows:

- We identify an indirect interference path under task-local training. Although old classes receive no direct negative gradients, current-task negative repulsion can still influence old representations through updates of the shared image encoder.
- We show that this interference is direction-dependent. We introduce a sample-normalized AGOP to characterize old-task-sensitive directions and use it to measure the interference risk of current-task negatives.
- We propose **AGRA**, which selectively attenuates high-risk negative repulsion while preserving low-risk negative signals. Experiments show consistent improvements over strong CLIP-based continual learning methods and better preservation of zero-shot generalization.

2 Related Work

Class-Incremental Learning. Class-incremental learning (CIL) aims to learn new classes sequentially while preserving previously acquired knowledge. Existing methods mainly follow three directions. Replay-based approaches revisit stored samples or generated features from previous tasks (Rebuffi et al., 2017; Buzzega et al., 2020; Zhu et al., 2021); regularization-based methods constrain model updates to reduce changes to important knowledge (Kirkpatrick et al., 2017; Li & Hoiem, 2017; Saha et al., 2021); and architecture-based methods allocate or adapt model components for incoming tasks (Mallya & Lazebnik, 2018; Hu et al., 2021b; Rypeść et al., 2024). While these approaches are effective for conventional visual models, the emergence of large-scale vision-language models introduces strong pre-trained representations and cross-modal knowledge, leading to a different continual learning setting.

CLIP-Based Continual Learning. Recent studies leverage the transferability and zero-shot capability of CLIP for continual learning (Thengane et al., 2022; Huang et al., 2024; Yu et al., 2024; Marczak et al., 2024; Gong et al., 2026). These methods improve continual adaptation through

parameter-efficient modules, knowledge integration, or cross-modal regularization. Another line of work studies the geometry of CLIP representations. The image and text features of CLIP exhibit a modality gap (Liang et al., 2022), and recent methods preserve this geometry during continual learning to reduce representation degradation (Huang et al., 2025; Peng et al., 2025). Our work takes a different view. Rather than directly preserving the modality gap or constraining parameter updates, we study how negative repulsion within the current task affects old representations through the shared image encoder. We further characterize this interference from a directional perspective and selectively attenuate high-risk negative updates. A more detailed discussion of related methods is provided in Appendix A.

3 PRELIMINARIES

3.1 CLASS-INCREMENTAL LEARNING WITH TASK-LOCAL OPTIMIZATION

We consider class-incremental learning with a pre-trained CLIP model (Radford et al., 2021). The model sequentially learns a stream of T tasks, where task t consists of a training set $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ and a class set $\mathcal{C}_t$. Classes are disjoint across tasks, i.e., $\mathcal{C}_t \cap \mathcal{C}_{t'} = \varnothing$ for $t \neq t'$. During training on task t, only samples from the current task are available, while previous-task images are not retained. After learning task t, the model is evaluated over all classes observed so far, $\mathcal{C}_{1:t} = \bigcup_{k=1}^{t} \mathcal{C}_k$, without access to task identities at inference time.

Given an image x_i, let $\mathbf{z}_i$ denote its ℓ_2-normalized CLIP image representation and $\mathbf{w}_c$ the normalized text representation of class c. Their classification logit is $s_{ic} = a\, \mathbf{z}_i^\top \mathbf{w}_c$, where a is the CLIP logit scale. Following the task-local training protocol adopted in continual CLIP (Huang et al., 2025; 2024), the softmax during task t is computed only over the current class set $\mathcal{C}_t$:

$$\mathcal{L}_i = -\log \frac{\exp(s_{iy_i})}{\sum_{c \in \mathcal{C}_t} \exp(s_{ic})}, \tag{1}$$

Previous classes are therefore excluded from the current softmax. For any $c \in \mathcal{C}_{1:t-1}$, $\partial \mathcal{L}_i^{\mathrm{loc}} / \partial s_{ic} = 0$, meaning that an old class is never directly treated as a negative while learning the current task.

3.2 MODALITY GAP

CLIP exhibits a characteristic modality gap between image and text representations, where the two modalities occupy separated regions of the joint embedding space (Liang et al., 2022). Following (Huang et al., 2025), we characterize the evolution of this cross-modal geometry using image–text cosine similarities. Given N images and K class text representations, we distinguish the similarity between an image and its ground-truth class text from its similarities to the remaining class texts:

$$\mathrm{Pos} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{sim}_{\cos}(\mathbf{z}_i, \mathbf{w}_{y_i}), \qquad \mathrm{Neg} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K-1} \sum_{c \neq y_i} \mathrm{sim}_{\cos}(\mathbf{z}_i, \mathbf{w}_c). \tag{2}$$

Here, Pos measures how well an image aligns with the text of its correct class, whereas Neg measures its average similarity to non-matching class texts. Prior work observes that Pos tends to stabilize early during downstream training, while Neg continues to decrease at later stages (Huang et al., 2025). We revisit this behavior under task-local continual learning and investigate how current-task negative repulsion is related to the representation drift of previously learned tasks.

3.3 REVISITING AVERAGE GRADIENT OUTER PRODUCT

The key challenge is to characterize which directions in the feature space are important to previously learned tasks. A suitable measure should capture the sensitivity of an old task to feature changes, while being independent of the magnitude of individual updates. For this purpose, we adopt the Average Gradient Outer Product (AGOP) (Radhakrishnan et al., 2024), which provides a gradient-based characterization of task-sensitive feature directions.

For a predictor f and a set of feature representations $\{\mathbf{z}_i\}_{i=1}^{N}$, the empirical AGOP is defined as

$$\mathbf{G} = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\mathbf{z}} f(\mathbf{z}_i) \nabla_{\mathbf{z}} f(\mathbf{z}_i)^\top. \tag{3}$$

Unlike scalar importance measures that only describe the magnitude of feature changes, AGOP preserves directional information through the outer product of gradients. For any unit direction $\mathbf{v}$, its quadratic form can be written as

$$\mathbf{v}^\top \mathbf{G}\mathbf{v} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{v}^\top \nabla_{\mathbf{z}} f(\mathbf{z}_i)\right)^2. \tag{4}$$

Therefore, a larger value indicates that the direction $\mathbf{v}$ has stronger alignment with the gradients that are relevant to the predictor. This property makes AGOP suitable for our setting. The goal is not to measure how large an old-task gradient is, but to identify which feature directions are more sensitive to changes from new-task updates. Thus, AGOP provides a task-dependent sensitivity geometry that allows us to measure the directional overlap between current-task negative updates and previously learned representations.

4 METHODOLOGY

4.1 DIAGNOSING DIRECTIONAL NEGATIVE INTERFERENCE

We first study how previously learned representations change during task-local training of a new task. All analyses are conducted on ImageNet-100 (Deng et al., 2009; Rebuffi et al., 2017), split into 10 sequential class-incremental tasks. As described in Sec. 3.1, only current-task classes participate in the softmax, and old classes therefore receive no direct logit-level negative gradients. We examine how current-task negative repulsion evolves together with old-task representation changes and analyze the relationship between them.

Current-task negative repulsion is associated with old-task drift. We first examine how image–text similarities evolve during current-task training. For task t, let $\mathrm{Pos}_t(e)$ and $\mathrm{Neg}_t(e)$ denote the average positive and negative image–text similarities at epoch e, respectively. We measure their changes from the beginning of training as

$$\Delta\mathrm{Pos}_t(e) = \mathrm{Pos}_t(e) - \mathrm{Pos}_t(0), \qquad \Delta\mathrm{Neg}_t(e) = \mathrm{Neg}_t(e) - \mathrm{Neg}_t(0). \tag{5}$$

As shown in Fig. 2(a), the positive similarity of the current task quickly stabilizes, whereas its negative similarity continues to decrease. Interestingly, although old classes are excluded from the current softmax, their image–text similarities also change during current-task training. Since the visual encoder is shared across tasks, current-task gradients update the same encoder used to represent old samples, providing a potential pathway for these changes. A detailed gradient analysis is provided in Appendix B.1.

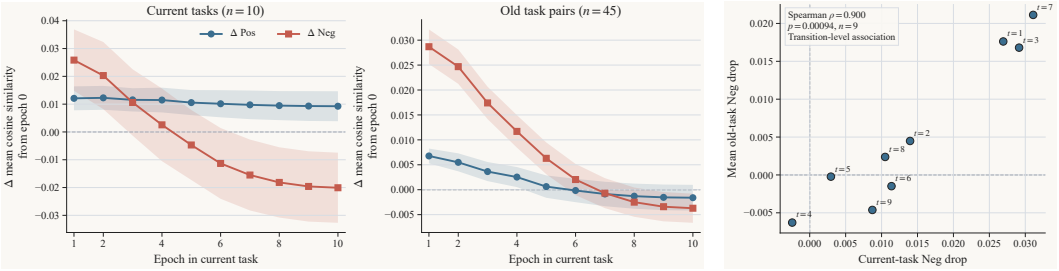
We further compare the decrease in negative similarity between the current and previously learned tasks. At epoch e, we define

$$\Delta_t^{\mathrm{cur}}(e) = \mathrm{Neg}_t(0) - \mathrm{Neg}_t(e), \qquad \Delta_t^{\mathrm{old}}(e) = \frac{1}{t-1}\sum_{k<t}\left[\mathrm{Neg}_k(0) - \mathrm{Neg}_k(e)\right]. \tag{6}$$

Here, Δ_t^{cur} measures the negative-similarity drop of the current task, while Δ_t^{old} measures the average drop over previously learned tasks. As shown in Fig. 2(b), the two quantities are strongly associated. This observation suggests that task-local masking removes direct old-class competition, but does not fully isolate old representations from changes during current-task learning. This raises a more specific question: *are all current-task negatives equally likely to interfere with old representations?*

Negative interference is direction-dependent. We find that the interference risk of a negative update depends not only on its magnitude, but also on its direction. To characterize the feature directions that are sensitive to an old task k, we construct a sample-normalized Average Gradient Outer Product (AGOP) from its task-local cross-entropy gradients.

For an old-task sample $\mathbf{x}_i$ with label y_i, let $\mathbf{z}_i$ denote its image feature and $\mathcal{L}_i^{(k)}$ denote its task-local cross-entropy loss over the class set $\mathcal{C}_k$, We compute its feature gradient $\mathbf{g}_i^{(k)} = \nabla_{\mathbf{z}_i}\mathcal{L}_i^{(k)}$ and



(a) Current-task and old-task similarity dynamics. (b) Current–old Neg association.

Figure 2: **Current-task negative repulsion is associated with old-task representation changes.** (a) Negative similarity keeps decreasing during current-task training, while old-task similarities also drift. (b) Larger current-task Neg drops are associated with larger old-task Neg drops.

construct the task sensitivity matrix as

$$\widehat{\mathbf{g}}_i^{(k)} = \frac{\mathbf{g}_i^{(k)}}{\|\mathbf{g}_i^{(k)}\|_2}, \qquad \mathbf{M}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \widehat{\mathbf{g}}_i^{(k)} \widehat{\mathbf{g}}_i^{(k)\top}. \tag{7}$$

The sample-wise normalization removes the influence of gradient scale and allows $\mathbf{M}_k$ to focus on the sensitive directions of the old task. For a unit direction $\mathbf{v}$, a larger value of $\mathbf{v}^\top \mathbf{M}_k \mathbf{v}$ indicates stronger alignment with these directions. A detailed discussion of the sample-normalized AGOP formulation is provided in Appendix B.2.

For a current sample $\mathbf{x}_i$ and a negative class $j \neq y_i$, the corresponding feature-gradient contribution is $\mathbf{g}_{ij}^- = a p_{ij}^{\mathrm{loc}} \mathbf{w}_j$. We measure its interference risk with respect to old task k as

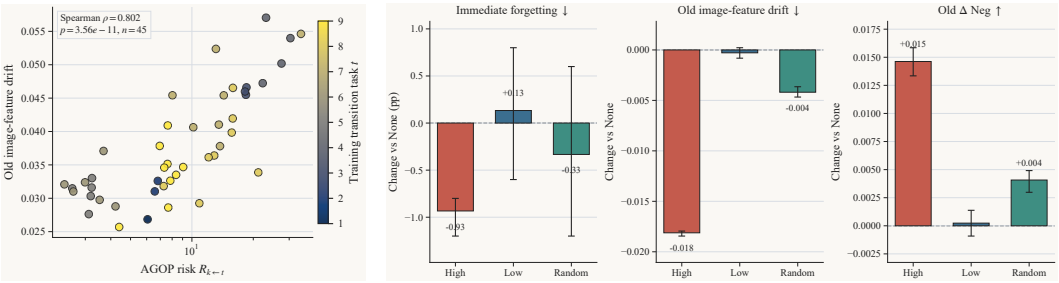
$$R_{ij}^k = (\mathbf{g}_{ij}^-)^\top \mathbf{M}_k \mathbf{g}_{ij}^- = \underbrace{a^2 (p_{ij}^{\mathrm{loc}})^2}_{\text{current gradient strength}} \cdot \underbrace{\mathbf{w}_j^\top \mathbf{M}_k \mathbf{w}_j}_{\text{old-task directional exposure}}. \tag{8}$$

Thus, the old task specifies *where* it is sensitive, while the current negative gradient specifies *how strongly* it overlaps with these sensitive directions. Aggregating this risk over current samples and negative classes gives the task-pair risk $R_{k \leftarrow t}$. As shown in Fig. 3(a), $R_{k \leftarrow t}$ is strongly associated with old image-feature drift, indicating that negatives with greater exposure to old-task-sensitive directions tend to accompany larger representation changes.

Directional effects under matched gradient strength. The strong association above suggests that negatives with larger interference risk are more likely to accompany old-task drift. However, since R_{ij}^k jointly depends on gradient strength and directional exposure, this observation alone cannot determine whether the difference is simply caused by larger negative gradients. We therefore conduct a gradient-strength-matched intervention to isolate the effect of direction. Specifically, negatives with similar gradient strength are grouped together, from which we select *High*-risk, *Low*-risk, and *Random* negatives and apply the same suppression.

As shown in Fig. 3(b), suppressing high-risk negatives consistently yields the largest reduction in immediate forgetting and old image-feature drift, while better preserving old-task negative similarity. In contrast, suppressing low-risk negatives produces only limited improvements despite having comparable gradient strength. Together with the risk–drift association in Fig. 3(a), these results indicate that the interference of current-task negatives cannot be explained by gradient magnitude alone; their exposure to old-task-sensitive directions is also critical.

Overall, our analyses reveal a structured interference mechanism under task-local training. Although old classes receive no direct negative gradients, current-task negative optimization can still affect old representations through the shared image encoder. More importantly, this interference is highly heterogeneous: negatives that are both strong and highly aligned with old-task-sensitive directions pose greater risk. This finding naturally motivates a selective strategy that preserves most negative signals for learning the current task, while attenuating only those with excessive interference risk.



(a) AGOP risk vs. old feature drift.

(b) Gradient-strength-matched intervention.

Figure 3: **Negative interference is direction-dependent.** (a) Higher AGOP risk is associated with larger old image-feature drift. (b) Under matched gradient strength, suppressing high-risk negatives provides the strongest protection.

4.2 AGOP-Guided Repulsion Attenuation

The above analysis suggests that current-task negatives exhibit substantially different interference risks. This motivates **A**GOP-**G**uided **R**epulsion **A**ttenuation (AGRA), which preserves low-risk negative signals while selectively attenuating negatives that are highly exposed to old-task-sensitive directions.

Historical directional sensitivity. After completing task k, we construct its directional sensitivity matrix $\mathbf{M}_k$ according to Eq. 7. When learning a new task t, we summarize the sensitivity of all previously learned tasks as

$$\mathbf{M}_{\text{old}} = \frac{1}{t-1} \sum_{k=1}^{t-1} \mathbf{M}_k. \tag{9}$$

Since each $\mathbf{M}_k$ is trace-normalized, the aggregation assigns comparable weight to different tasks rather than favoring tasks with larger gradient scales. The sensitivity matrix of task t is computed only after finishing its training and is used for subsequent tasks. Therefore, AGRA does not require storing or replaying previous-task images.

Negative interference risk. For a current sample i and a negative class $j \neq y_i$, its task-local feature-gradient contribution is $\mathbf{g}_{ij}^{-} = a p_{ij}^{\text{loc}} \mathbf{w}_j$. Using the aggregated historical sensitivity, we measure its interference risk as

$$R_{ij} = a^2 \, \text{sg}\left[\left(p_{ij}^{\text{loc}} \right)^2 \right] \mathbf{w}_j^\top \mathbf{M}_{\text{old}} \mathbf{w}_j, \qquad j \neq y_i, \tag{10}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation. The two factors in Eq. 10 have complementary roles: $a^2 (p_{ij}^{\text{loc}})^2$ measures how strongly the current negative pushes the representation, whereas $\mathbf{w}_j^\top \mathbf{M}_{\text{old}} \mathbf{w}_j$ measures how strongly its direction overlaps with the sensitivity geometry of previous tasks. The risk is used only to determine the attenuation weight and does not introduce an additional gradient path.

Selective risk-adaptive attenuation. Uniformly weakening all negative terms would also reduce useful discriminative signals for learning the current task. We therefore use the risk distribution of each sample to identify only relatively high-risk negatives. Specifically, we define a sample-wise threshold

$$\tau_i = Q_q \left(\{ R_{ij} \mid j \in \mathcal{C}_t, \, j \neq y_i \} \right), \tag{11}$$

where $Q_q(\cdot)$ denotes the q-th quantile. The quantile q determines which portion of negatives is considered high-risk. Rather than introducing an additional attenuation-strength hyperparameter, we directly determine the attenuation from the risk relative to this threshold:

$$\alpha_{ij} = \min \left(1, \, \frac{\tau_i + \epsilon}{R_{ij} + \epsilon} \right), \qquad j \neq y_i. \tag{12}$$

For negatives with $R_{ij} \leq \tau_i$, we have $\alpha_{ij} = 1$, and their original learning signal is fully preserved. For $R_{ij} > \tau_i$, the attenuation becomes stronger as the relative interference risk increases. Thus,

q determines *which* negatives are attenuated, while the risk itself determines *how much* they are attenuated. All quantities used to compute α_{ij} are treated with stop-gradient.

We incorporate the resulting weights into the negative terms of the task-local softmax:

$$\mathcal{L}_i^{\mathrm{AGRA}} = -\log \frac{\exp(s_{iy_i})}{\exp(s_{iy_i}) + \sum_{j \in \mathcal{C}_t, j \neq y_i} \alpha_{ij} \exp(s_{ij})}. \tag{13}$$

AGRA does not directly reweight the positive term and leaves all low-risk negatives identical to the original task-local objective. Only negatives whose interference risk exceeds the sample-wise threshold are attenuated. In this way, AGRA preserves most negative signals required for current-task discrimination while reducing updates that are more strongly exposed to previously learned sensitive directions.

For the first task, no historical sensitivity matrix is available, and we simply optimize the standard task-local cross-entropy. After finishing each task, its sensitivity matrix is computed from the current-task data before the data are discarded. AGRA therefore introduces no replay buffer and requires only the historical sensitivity statistics during subsequent training.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate AGRA on CIFAR-100 (Krizhevsky et al., 2009), ImageNet-R (Hendrycks et al., 2021), ImageNet-100 (Rebuffi et al., 2017), ImageNet-1K (Deng et al., 2009), and VTAB (Zhai et al., 2019), using the continual learning protocol adopted in (Zhou et al., 2025a). CIFAR-100, ImageNet-R, ImageNet-100, and ImageNet-1K are evenly divided into 10 class-incremental tasks, while VTAB contains 5 tasks. The class sets are disjoint across tasks, and only current-task data are available during training. At inference time, the model predicts over all classes observed so far without task identity. Results are averaged over three different class orders.

Baselines. We compare AGRA against a broad set of representative continual learning methods, covering prompt-based approaches, parameter-efficient adaptation methods, and recent CLIP-based continual learning algorithms. Specifically, we include L2P++ (Wang et al., 2022c), DualPrompt (Wang et al., 2022b), CODA (Smith et al., 2023), SLCA (Zhang et al., 2023), Continual-CLIP (Thengane et al., 2022), RAPF (Huang et al., 2024), PROOF (Zhou et al., 2025b), CLAP (Jha et al., 2024), MOE4CL (Yu et al., 2024), MagMax (Marczak et al., 2024), MGCLIP (Huang et al., 2025), and LfI (Gong et al., 2026). A detailed description and categorization of all compared methods are provided in the Appendix.

Evaluation metrics. We report *Average Accuracy* (Avg) and *Last Accuracy* (Last) as the main evaluation metrics. Avg measures the average class-incremental accuracy throughout the task sequence, while Last measures the accuracy over all learned classes after the final task.

Implementation details. All experiments are implemented in PyTorch and run on a single NVIDIA A100 GPU. Unless otherwise specified, we retain the continual training paradigm of MGCLIP, and introduce AGRA on top of the same adaptation pipeline. We use OpenAI CLIP ViT-B/16 as the default backbone and adapt the image encoder with LoRA (Hu et al., 2021a), where rank-8 adapters are inserted into the key and value projections of the attention layers, while the text encoder remains frozen. Optimization is performed with Adam using an initial learning rate of 1×10^{-3} together with a cosine decay schedule. For AGRA, the directional sensitivity matrix of each task is computed once the task is completed and reused for subsequent tasks, without storing past images. We use a fixed risk quantile of $q = 0.8$ throughout the main experiments.

5.2 Comparison Results

Table 1 compares AGRA with representative continual learning methods under different replay settings. AGRA achieves the best performance across all reported metrics on CIFAR-100, ImageNet-R, ImageNet-100, ImageNet-1K, and VTAB, while requiring no replay. The consistent improvements across datasets with different scales and distributions demonstrate the effectiveness and robustness of our approach. Compared with MGCLIP, whose continual training paradigm is retained in our

Table 1: Comparison with different methods. DR denotes replay with real data, FR denotes replay with generated features of old classes, and NR denotes methods without replay. Most results are taken from prior work (Jha et al., 2024; Huang et al., 2025; Gong et al., 2026) or reproduced using the publicly available code. Our results are averaged over three different class orders.

Method	Exemplar	CIFAR-100		ImageNet-R		ImageNet100		ImageNet-1K		VTAB	
		Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last
PROOF	DR	84.88	76.29	82.83	77.05	84.71	72.48	76.23	65.26	89.09	83.97
CLAP	DR	86.13	78.21	85.77	79.98	87.76	79.16	81.72	73.19	91.37	89.67
SLCA	FR	80.53	67.58	75.92	70.37	78.63	59.92	79.10	68.27	84.25	82.54
RAPF	FR	86.19	79.04	85.58	80.28	87.51	*80.23*	81.73	72.58	90.88	82.31
L2P++	NR	81.90	73.08	81.67	75.98	80.51	67.22	79.30	69.60	63.23	38.37
DualPrompt	NR	81.45	72.51	82.01	75.77	80.65	67.38	79.39	69.79	61.89	37.58
CODA	NR	76.98	62.25	78.00	67.52	64.13	34.76	76.99	66.96	62.51	38.25
Continual-CLIP	NR	75.15	66.68	79.12	72.00	84.98	75.40	72.96	64.44	53.64	31.50
Aper-Adapter	NR	75.76	65.50	78.65	71.35	85.84	76.40	76.60	68.74	80.75	71.21
MOE4CL	NR	85.36	78.37	85.28	80.77	86.39	76.66	81.29	72.73	68.49	61.70
CLAP*	NR	74.19	63.45	81.22	75.80	81.07	72.00	75.85	67.36	82.11	80.11
MagMax	NR	85.63	79.00	87.13	80.85	86.33	75.92	80.74	71.31	64.63	53.90
MGCLIP	NR	87.00	80.57	87.58	82.67	*87.31*	78.38	81.88	73.68	*94.67*	*91.53*
LfI	NR	*87.24*	*81.35*	*87.62*	*82.83*	87.26	78.10	*82.38*	*74.23*	-	-
AGRA (Ours)	NR	**88.21**	**81.87**	**88.76**	**83.72**	**88.42**	**81.05**	**82.73**	**74.95**	**95.07**	**92.64**

implementation, AGRA consistently improves both Avg and Last accuracy on all benchmarks. This shows that selectively attenuating high-risk negative repulsion provides complementary gains beyond the underlying continual adaptation strategy.

Notably, AGRA also outperforms methods that rely on real-data or feature replay. These results suggest that explicitly mitigating directional negative interference can effectively preserve previously learned knowledge without storing old samples, while maintaining strong adaptation to new tasks.

5.3 Preserving Zero-Shot Generalization

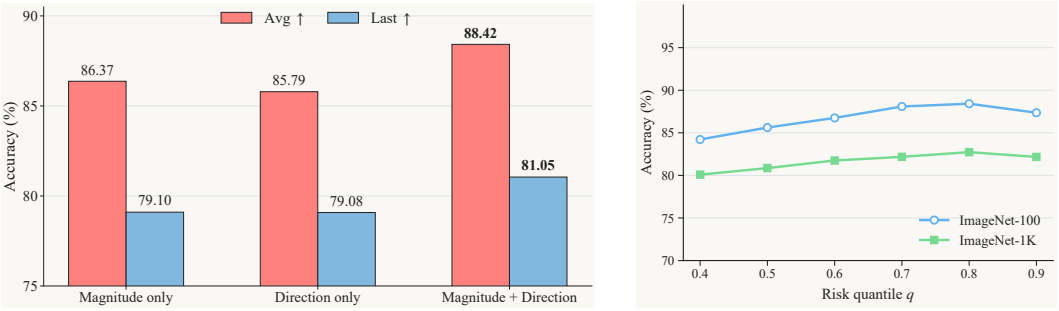
An important advantage of CLIP is its strong zero-shot generalization acquired from large-scale vision–language pre-training. Therefore, continual adaptation should not only retain knowledge from previously learned downstream tasks, but also avoid degrading the generalizable representations inherited from the pre-trained model. To evaluate this property, we measure zero-shot performance on independent datasets after completing continual learning on CIFAR-100.

Table 2: Zero-shot performance after continual learning on CIFAR-100.

	Food101	Pets	ImageNet-1K	Avg
CLIP	85.14	87.60	64.44	79.06
PROOF	9.75	22.81	11.18	14.58
RAPF	17.18	28.56	15.20	20.31
CLAP	80.82	74.68	56.04	70.51
MOE4CL	82.85	84.06	66.02	77.64
MagMax	81.67	85.96	66.44	78.02
MGCLIP	85.70	88.17	67.29	80.39
LfI	85.98	88.80	67.86	80.88
AGRA (Ours)	**86.42**	**89.43**	**68.25**	**81.37**

As shown in Table 2, AGRA achieves the best zero-shot performance on Food101 (Bossard et al., 2014), Oxford Pets (Parkhi et al., 2012), and ImageNet-1K (Deng et al., 2009), with an average accuracy of 81.37%. Compared with the original CLIP model, AGRA improves performance on all three datasets, indicating that continual adaptation does not compromise its original zero-shot capability. In particular, the improvement on ImageNet-1K suggests that AGRA can preserve the semantic structure of the pre-trained representation while incorporating knowledge from sequential downstream tasks.

Compared with existing continual learning methods, AGRA also consistently retains stronger zero-shot generalization. Replay-based approaches such as PROOF, RAPF, and CLAP exhibit substantial degradation after continual training, while recent replay-free CLIP-based methods preserve the pre-trained capability more effectively. AGRA further improves upon MGCLIP and LfI, suggesting that selectively attenuating high-risk negative repulsion not only reduces forgetting on learned tasks, but also better preserves the generalizable representation inherited from CLIP.



(a) Ablation on interference risk.

(b) Sensitivity to the risk quantile q.

Figure 4: Ablation results. *Left:* the full interference risk, which combines gradient strength and directional exposure, performs best on ImageNet-100. *Right:* AGRA remains stable across a broad range of q.

5.4 Ablation Study

Ablation on Interference Risk. We study the design of the interference risk on ImageNet-100 by comparing three variants: *Magnitude only*, *Direction only*, and the full *Magnitude + Direction* risk. As shown in Fig. 4 *left*, combining gradient strength and directional exposure achieves the best performance on both Avg and Last accuracy. This indicates that gradient magnitude alone is insufficient to identify harmful negatives, while directional information provides complementary signals for measuring interference risk.

Sensitivity to the Risk Quantile q. We evaluate the sensitivity of AGRA to the risk quantile q on ImageNet-100 and ImageNet-1K. As shown in Fig. 4 *right*, AGRA achieves stable performance across different values of q and obtains the best results around $q = 0.8$ on both datasets. This indicates that AGRA is robust to the choice of the risk threshold.

Additional Analysis. Further analyses are provided in Appendix B. We investigate the characteristics of the proposed task sensitivity matrix from a spectral perspective, showing how sensitive directions are distributed in the feature space. We also provide a detailed analysis of the storage and computational overhead of AGRA, including the additional memory required for maintaining historical sensitivity information and the extra computation introduced during training.

6 Conclusion

In this work, we study how current-task negative optimization affects previously learned representations in CLIP-based continual learning. We show that task-local training removes direct old-class competition, but current-task negative gradients can still change old representations through the shared image encoder. We further find that this interference is direction-dependent and is better captured by both gradient strength and overlap with old-task-sensitive directions. Based on this insight, we propose AGRA, which selectively attenuates high-risk negative repulsion while preserving low-risk signals. Experiments show that AGRA improves continual learning performance and better preserves CLIP's zero-shot generalization.

AI Assistance Disclosure

In this work, we used generative AI tools to assist with writing and language polishing, as well as literature retrieval and discovery of relevant prior work. All AI-assisted content and retrieved references were reviewed and verified by the authors. We take full responsibility for the final content of this work.

REFERENCES

Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3366–3375, 2017.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. *Advances in Neural Information Processing Systems 33*, pp. 15920–15930, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Qiankun Gao, Chen Zhao, Bernard Ghanem, and Jian Zhang. R-dfcil: Relation-guided representation learning for data-free class incremental learning. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Yizheng Gong, Siyue Yu, Waleed Al-Nuaimy, and Jimin Xiao. Learning from itself: Mining internal knowledge from vision language models for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10830–10839, 2026.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021a.

Wenpeng Hu, Qi Qin, Mengyu Wang, Jinwen Ma, and Bing Liu. Continual learning by using information of each class holistically. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 7797–7805, 2021b.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pp. 214–231. Springer, 2024.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3777–3786. IEEE, 2025.

Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in neural information processing systems*, 37:129146–129186, 2024.

Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16050–16059. IEEE, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, pp. 3521–3526, 2017.

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

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Zilin Luo, Yaoyao Liu, Bernt Schiele, and Qianru Sun. Class-incremental exemplar compression for class-incremental learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11371–11380. IEEE, 2023.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pp. 379–395. Springer, 2024.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Tiantian Peng, Yuyang Liu, Shuo Yang, Qiuhe Hong, and YongHong Tian. Gnsp: Gradient null space projection for preserving cross-modal alignment in vlms continual learning. *arXiv preprint arXiv:2507.19839*, 2025.

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

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, pp. 123–146, 1995.

Grzegorz Rypeść, Sebastian Cygert, Valeriya Khan, Tomasz Trzciński, Bartosz Zieliński, and Bartłomiej Twardowski. Divide and not forget: Ensemble of selectively trained experts in continual learning. *arXiv preprint arXiv:2401.10191*, 2024.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*, 2022.

Liyuan Wang, Xingxing Zhang, Qian Li, Jun Zhu, and Yi Zhong. Coscl: Cooperation of small continual learners is stronger than a big one. In *Proceedings of the 17th European Conference on Computer Vision*, pp. 254–271, 2022a.

Tianqi Wang and Jingcai Guo. Gr4cil: Gap-compensated routing for clip-based class incremental learning. *arXiv preprint arXiv:2604.17822*, 2026.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149. IEEE, 2022c.

Zhen-Hao Xie, Yu-Cheng Shi, and Da-Wei Zhou. AREA: Attribute extraction and aggregation for CLIP-based class-incremental learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5th8Uu1HjJ.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23219–23230. IEEE, 2024.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19091–19101. IEEE, 2023.

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19068–19079. IEEE, 2023.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025a.

Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4489–4504, 2025b.

Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5871–5880, 2021.

Appendix

CONTENTS

A Related Work **14**

A.1 Class-Incremental Learning 14

A.2 CLIP-Based Continual Learning 14

A.3 Modality Gap in Continual CLIP 14

B More Analysis and Discussion **15**

B.1 Analysis of Negative Gradient Interference 15

B.2 Sample-Normalized AGOP for Task Sensitivity 16

B.3 Spectral Structure of Task Sensitivity 16

B.4 Efficiency Analysis 17

B.5 Replay-Free Continual Learning Setting 18

C Experimental Details **18**

C.1 Competitors 18

D Limitations and Future Work **19**

A Related Work

A.1 Class-Incremental Learning

Class-incremental learning requires a model to learn a sequence of disjoint class sets while predicting all previously observed classes at inference time. A major challenge is catastrophic forgetting, where learning new classes degrades knowledge acquired from earlier tasks. Existing approaches can be broadly grouped into replay-based, regularization-based, and architecture-based methods.

Replay-based methods. Replay-based approaches preserve previous knowledge by revisiting historical information during new-task training. Early experience-replay methods store a subset of old samples (Robins, 1995), while iCaRL (Rebuffi et al., 2017) maintains representative exemplars for old classes. Later methods improve replay through better memory selection, distillation, or representation constraints (Buzzega et al., 2020; Zhu et al., 2021; Luo et al., 2023). Although replay is effective for reducing forgetting, its storage requirement becomes less desirable when strong pre-trained representations already provide substantial prior knowledge.

Regularization-based methods. Regularization approaches preserve old knowledge by restricting model updates. EWC (Kirkpatrick et al., 2017) penalizes changes to parameters that are important to previous tasks, while LwF (Li & Hoiem, 2017) uses knowledge distillation to maintain previous predictions. Later studies introduce gradient constraints or representation-level regularization to further reduce interference across tasks (Saha et al., 2021; Gao et al., 2022; Kang et al., 2022). These methods mainly control the overall update of the model rather than identifying which components of the current-task optimization are more likely to interfere with old knowledge.

Architecture-based methods. Another family of approaches modifies the network architecture as new tasks arrive. Some methods allocate task-specific components (Aljundi et al., 2017), while others identify reusable and task-specific parameters through pruning or network expansion (Mallya & Lazebnik, 2018; Hu et al., 2021b). More recent methods improve parameter sharing and task-specific specialization to achieve a better stability–plasticity trade-off (Wang et al., 2022a; Rypeść et al., 2024). These methods provide explicit capacity for new tasks, but often require additional architectural design.

A.2 CLIP-Based Continual Learning

The strong transferability of CLIP has motivated increasing interest in continual learning with vision-language models. Continual-CLIP (Thengane et al., 2022) shows that the pre-trained CLIP representation itself provides strong zero-shot performance in class-incremental settings. Subsequent methods focus on adapting CLIP while preserving its pre-trained knowledge.

One line of work introduces lightweight or task-specific modules. RAPF (Huang et al., 2024) adapts CLIP representations and integrates parameters across tasks. MOE4CL (Yu et al., 2024) employs mixture-of-experts adapters and dynamically expands the model to improve adaptation to incoming tasks. PROOF (Zhou et al., 2025b) learns projection modules for the visual and textual branches to improve cross-modal representations. These methods exploit the strong pre-trained representation of CLIP while limiting changes to the original model.

Another line of work focuses on integrating knowledge across tasks. MagMax (Marczak et al., 2024) introduces model merging into continual learning and combines task updates according to their parameter magnitudes. LfI (Gong et al., 2026) generates pseudo-text descriptions as additional supervision and transfers knowledge through adaptive mutual distillation. These methods mainly address how knowledge from different tasks can be preserved or combined during continual adaptation.

A.3 Modality Gap in Continual CLIP

CLIP image and text representations are not perfectly aligned in the shared embedding space. Liang et al. (Liang et al., 2022) show that the two modalities occupy separated regions, forming a characteristic *modality gap*. This geometric structure is closely related to the pre-trained representation of CLIP and can change during downstream adaptation.

Recent continual learning methods explicitly consider this cross-modal geometry. MGCLIP (Huang et al., 2025) studies the evolution of positive and negative image–text similarities during continual training. It controls fine-tuning to preserve the pre-trained modality gap and introduces a visual classifier to compensate for reduced adaptability. GNSP (Peng et al., 2025) protects historical knowledge by projecting new-task gradients into a constrained space and further introduces cross-modal preservation objectives.

Our work is closely related to these studies but focuses on a different question. Under task-local cross-entropy, old classes are removed from the current softmax and therefore receive no direct negative gradients. Nevertheless, old image–text relations can still change when a new task is learned. We study this indirect interference path and show that negative repulsion within the current task can affect old representations through the shared image encoder. More importantly, we characterize the interference at the level of individual negative directions. This leads to our AGOP-based risk measure and the selective attenuation strategy used in AGRA.

B More Analysis and Discussion

B.1 Analysis of Negative Gradient Interference

Although task-local cross-entropy removes old classes from the current-task softmax, current-task negative optimization can still influence previously learned representations through the shared visual encoder. In this section, we provide a gradient-level analysis to explain this interference pathway.

For a current-task sample x_i and its non-target class j, the negative component of the task-local classification loss can be written as

$$\mathcal{L}_{ij}^{-} = -\log \frac{\exp(s_{iy})}{\exp(s_{iy}) + \sum_{m \neq y} \exp(s_{im})}, \tag{14}$$

where s_{ij} denotes the similarity between the image feature and the text feature of class j. The gradient with respect to a negative logit s_{ij} is

$$\frac{\partial \mathcal{L}}{\partial s_{ij}} = p_{ij}^{\mathrm{loc}}, \qquad j \neq y, \tag{15}$$

where p_{ij}^{loc} is the task-local softmax probability. During gradient descent, this positive derivative decreases the corresponding negative similarity, resulting in negative repulsion between the image feature and non-target class text features.

Importantly, the negative gradient is not restricted to the text classifier. Since CLIP shares the visual encoder across all tasks, the gradient is back-propagated through the image representation:

$$\nabla_{\theta_v} \mathcal{L}_{ij}^{-} = \frac{\partial s_{ij}}{\partial f_v(x_i)} \frac{\partial f_v(x_i)}{\partial \theta_v}, \tag{16}$$

where $f_v(\cdot)$ denotes the visual encoder and θ_v represents its parameters. Therefore, optimizing current-task negative similarities updates the shared visual encoder parameters. Although these updates are computed using only current-task samples, the updated encoder is also used to extract features for previously learned tasks. Consequently, the feature representations of old-task samples may change after learning new tasks.

This analysis explains why removing old classes from the softmax eliminates direct logit-level competition but does not completely remove representation interference. The effect of a negative update depends on how the induced parameter change interacts with previously learned feature directions. Therefore, beyond measuring the magnitude of negative updates, it is necessary to characterize whether their directions overlap with sensitive directions of old tasks, which motivates the proposed AGOP-based interference risk.

B.2 SAMPLE-NORMALIZED AGOP FOR TASK SENSITIVITY

The goal of our sensitivity representation is to characterize which feature directions are important to previously learned tasks. A direct AGOP construction from raw gradients can capture gradient-sensitive directions, but it may be dominated by samples with large gradient magnitudes. In our setting, however, the magnitude of old-task gradients is not the main factor of interest. Instead, we aim to capture the directional structure of old tasks, which determines whether a current-task update overlaps with previously sensitive feature directions.

For an old task k, let $\mathbf{g}_i^{(k)} = \nabla_{\mathbf{z}_i} \mathcal{L}_i^{(k)}$ denote the feature gradient of sample $\mathbf{x}_i$. The standard gradient outer product is defined as

$$\mathbf{G}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{g}_i^{(k)} \mathbf{g}_i^{(k)\top}. \tag{17}$$

Although $\mathbf{G}_k$ contains directional information, its value is also strongly affected by gradient scale. Specifically, samples with larger gradient norms contribute larger outer products and may dominate the resulting sensitivity matrix. This makes the matrix partially reflect the optimization state of individual samples rather than the intrinsic directional structure of the task.

To reduce this effect, we normalize each sample gradient before constructing the outer product:

$$\widehat{\mathbf{g}}_i^{(k)} = \frac{\mathbf{g}_i^{(k)}}{\|\mathbf{g}_i^{(k)}\|_2}. \tag{18}$$

The resulting sample-normalized AGOP is defined as

$$\mathbf{M}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \widehat{\mathbf{g}}_i^{(k)} \widehat{\mathbf{g}}_i^{(k)\top}. \tag{19}$$

The normalization has two effects. First, each sample contributes only its gradient direction, preventing a small number of high-magnitude samples from dominating the sensitivity estimation. Second, it makes $\mathbf{M}_k$ describe the distribution of sensitive directions rather than the remaining gradient strength of the old task.

The directional interpretation follows from the quadratic form. For any unit vector $\mathbf{v}$,

$$\mathbf{v}^\top \mathbf{M}_k \mathbf{v} \propto \frac{1}{N_k} \sum_{n=1}^{N_k} \left(\mathbf{v}^\top \widehat{\mathbf{g}}_n^{(k)} \right)^2. \tag{20}$$

Therefore, a larger value indicates that $\mathbf{v}$ has stronger alignment with the directions emphasized by the old task. This property makes sample-normalized AGOP suitable for our purpose: it provides a compact representation of old-task-sensitive directions, allowing the interference risk of a current-task negative update to be measured by its directional overlap rather than only its magnitude.

B.3 SPECTRAL STRUCTURE OF TASK SENSITIVITY

To further understand the structure captured by the sample-normalized AGOP, we analyze the eigenvalue spectrum of the task sensitivity matrix $\mathbf{M}_k$ on ImageNet-100. Figure 5 shows the cumulative normalized eigenvalue trace for all tasks, together with two measures of spectral dimensionality.

As shown in Fig. 5(a), the spectrum is highly concentrated. A relatively small number of leading eigencomponents already explains most of the total trace, indicating that the sensitivity of an old task is far from isotropic in the 512-dimensional feature space. Figure 5(b) gives a more direct view of this structure. Across tasks, the entropy effective rank remains roughly between 12 and 30, while about 20–105 components are sufficient to explain 95% of the total trace. These values are much smaller than the full feature dimension $d = 512$.

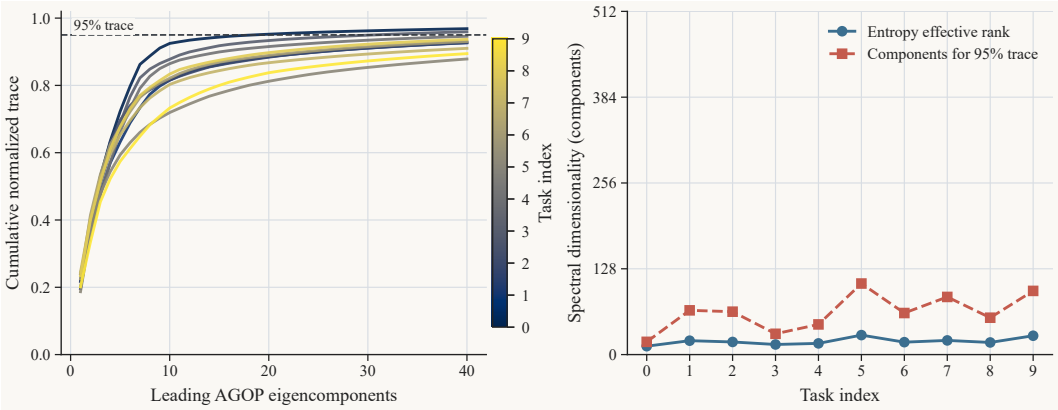


Figure 5: **Spectral structure of the task sensitivity matrices on ImageNet-100.** *left:* Cumulative normalized eigenvalue trace. *right:* Entropy effective rank and the number of components required to explain 95% of the trace.

This result shows that old-task sensitivity is concentrated along a limited set of feature directions rather than being distributed uniformly across the representation space. It also supports our directional view of negative interference: current-task negatives can have very different effects depending on how strongly they overlap with these sensitive directions.

B.4 Efficiency Analysis

We provide a more detailed analysis of the storage and computational overhead introduced by AGRA. All measurements are conducted on ImageNet-100 with the same 10-task continual learning protocol used in the main experiments. Training time is measured on a single NVIDIA A100 GPU, and computational cost is reported in FP32-equivalent PFLOPs over the complete task sequence.

Storage Overhead. AGRA represents the directional sensitivity of a task using a matrix $\mathbf{M}_k \in \mathbb{R}^{d\times d}$. If the sensitivity matrix of every previous task were stored independently, the memory cost after K tasks would be

$$4Kd^2 \text{ bytes}, \tag{21}$$

where each matrix element is stored in FP32. For the CLIP ViT-B/16 backbone used in our experiments, $d = 512$. Therefore, one sensitivity matrix requires

$$512^2 \times 4 = 1 \text{ MiB}, \tag{22}$$

and naively storing ten task-specific matrices would require approximately 10 MiB.

In practice, however, AGRA does not need to retain the individual matrices $\{\mathbf{M}_k\}$. As described in Sec. 4.2, the interference risk during task t only depends on their aggregated form,

$$\mathbf{M}_{\text{old}}^{(t)} = \frac{1}{t-1}\sum_{k=1}^{t-1}\mathbf{M}_k. \tag{23}$$

This quantity can be updated incrementally after completing task t:

$$\mathbf{M}_{\text{old}}^{(t+1)} = \frac{(t-1)\mathbf{M}_{\text{old}}^{(t)} + \mathbf{M}_t}{t}. \tag{24}$$

Once the update is completed, the task-specific matrix $\mathbf{M}_t$ does not need to be stored separately. Therefore, the persistent historical storage required by AGRA is only

$$4d^2 \text{ bytes}, \tag{25}$$

which is independent of the number of tasks. For $d = 512$, this corresponds to approximately 1 MiB. A temporary matrix may be required when constructing $\mathbf{M}_t$ at the end of a task, but this memory is released after the running aggregate is updated and does not grow with the task sequence.

Table 3: Computational overhead of AGRA on ImageNet-100 using a single NVIDIA A100 GPU.

Method	Training time (GPU-h) ↓	FP32 compute (PFLOPs) ↓
Baseline	2.702	116.72
AGRA	3.004	128.09
Relative change	$+11.2\%$	$+9.7\%$

This property makes the storage cost of AGRA fundamentally different from replay-based approaches. The memory required for historical information remains $\mathcal{O}(d^2)$ rather than increasing with the number of stored samples or tasks. Moreover, AGRA does not retain any previous-task images.

Computational Overhead. AGRA introduces two additional computations compared with the underlying continual learning baseline. First, the sensitivity matrix $\mathbf{M}_t$ is constructed once after completing each task. Second, during subsequent training, AGRA evaluates the directional risk of current-task negatives with respect to the aggregated historical sensitivity matrix. No additional forward pass through old-task samples is required during current-task training.

Table 3 reports the end-to-end cost over the full ImageNet-100 continual learning sequence. The baseline requires 2.702 GPU-hours and 116.72 PFLOPs, whereas AGRA requires 3.004 GPU-hours and 128.09 PFLOPs. This corresponds to an increase of 11.2% in training time and 9.7% in FP32 computation. The small difference between the two relative increases is expected, since wall-clock time also includes the task-end sensitivity construction and other implementation overhead that is not fully reflected by FLOP counts.

Overall, AGRA introduces a modest increase in computation while keeping its historical storage constant throughout continual learning. The additional cost comes from explicitly modeling old-task directional sensitivity, while the memory footprint remains approximately 1 MiB for the default CLIP ViT-B/16 backbone.

B.5 Replay-Free Continual Learning Setting

AGRA follows a replay-free continual learning setting, where no historical images or old-task samples are stored or replayed during the learning of new tasks. Instead of relying on previous data, AGRA preserves information from earlier tasks through a compact representation of their sensitivity directions.

Specifically, after completing each task, AGRA computes the task sensitivity matrix using the gradients of the completed task and only maintains the aggregated historical sensitivity matrix. During subsequent task learning, this matrix is used to estimate the interference risk of current-task negative gradients. Therefore, AGRA does not require access to old samples when optimizing new tasks.

This design is different from replay-based methods, which reduce forgetting by revisiting stored examples or generated samples from previous tasks. By storing only task-level directional statistics rather than historical data, AGRA avoids the memory growth associated with replay buffers while still preserving information about which feature directions are important to previous tasks. Such a design is particularly suitable for CLIP-based continual learning, where strong pre-trained representations provide prior knowledge and the goal is to adapt the model while minimizing unnecessary changes to the learned representation space.

C Experimental Details

C.1 Competitors

We provide a brief overview of the baseline methods considered in our experiments.

- **L2P++**(Wang et al., 2022c): L2P++ maintains a pool of learnable prompts and retrieves relevant prompts through input-dependent key–query matching while keeping the pre-trained back-

bone frozen. It adopts prefix tuning in place of the original prompt-tuning formulation of L2P(Li & Hoiem, 2017) to improve adaptation in rehearsal-free continual learning.

- **DualPrompt** (Wang et al., 2022b):DualPrompt introduces complementary general and expert prompts to capture task-invariant and task-specific knowledge, respectively. These prompts are attached to different layers of a frozen pre-trained transformer, enabling continual adaptation without storing previous training examples.
- **CODA** (Smith et al., 2023): CODA constructs input-conditioned prompts as weighted combinations of learnable prompt components through an attention-based mechanism that supports end-to-end optimization. It incrementally introduces new components while freezing previously learned ones and imposing orthogonality constraints to reduce interference across tasks.
- **SLCA** (Zhang et al., 2023): SLCA mitigates progressive overfitting by updating the pre-trained feature extractor with a substantially smaller learning rate than the classifier. It further models class-wise feature distributions and performs post-hoc classifier alignment to improve discrimination across incrementally learned classes.
- **Continual-CLIP** (Thengane et al., 2022):Continual-CLIP directly applies a frozen CLIP model to continual recognition through zero-shot image–text matching. As new classes arrive, it expands the set of textual class representations used for prediction without updating model parameters or replaying previous data.
- **RAPF** (Huang et al., 2024): RAPF uses textual features to estimate how newly introduced classes affect previously learned classes and adaptively adjusts visual representations during adapter training. It subsequently applies decomposed parameter fusion to consolidate previously acquired and newly learned knowledge, further alleviating forgetting.
- **PROOF** (Zhou et al., 2025b): PROOF incrementally learns task-specific projections on top of frozen image and text encoders while preserving projections learned for previous tasks. It further incorporates a cross-modal fusion module that jointly refines visual and textual representations to improve incremental recognition.
- **CLAP** (Jha et al., 2024): CLAP introduces probabilistic fine-tuning by modeling distributions over visually guided textual features for individual tasks. It exploits pre-trained knowledge for parameter initialization and distribution regularization, alleviating forgetting while supporting uncertainty estimation.
- **MOE4CL** (Yu et al., 2024): MOE4CL progressively expands a pre-trained CLIP model with mixture-of-experts adapters to acquire task-specific knowledge through parameter-efficient adaptation. A distribution-discriminative auto-selector routes inputs to the adapted model or the original CLIP according to their distribution, helping preserve zero-shot recognition capability.
- **MagMax** (Marczak et al., 2024): MagMax combines sequential fine-tuning with model merging to consolidate knowledge acquired across tasks. It constructs parameter-update vectors relative to the pre-trained model and merges them by selecting, at each parameter coordinate, the update with the largest absolute magnitude.
- **MGCLIP** (Huang et al., 2025): MGCLIP preserves the image–text modality gap during continual adaptation to retain pre-trained knowledge and mitigate forgetting. It additionally introduces a complementary classifier in the visual feature space to compensate for the limitations imposed by the modality gap and improve adaptation to new classes.
- **LfI** (Gong et al., 2026): LfI mines internal knowledge from CLIP by optimizing learnable tokens to generate pseudo-captions, providing auxiliary supervision without external models. It further employs adaptive mutual distillation between the text-based prediction branch and a temporary visual classifier, dynamically weighting knowledge transfer according to their relative performance.

D Limitations and Future Work

Although AGRA effectively reduces directional negative interference in CLIP-based continual learning, several limitations remain.

First, AGRA is not a completely memory-free method. Instead of storing old samples, it maintains compact historical sensitivity information in the form of an aggregated AGOP matrix. This introduces additional computation for estimating task sensitivity and requires storing a $d \times d$ matrix. However, this design provides an important advantage: the memory cost is constant with respect to the number of tasks and does not grow with the size of the replay buffer. More importantly, the stored statistics preserve directional information that cannot be captured by simple scalar measurements such as gradient norms. Therefore, AGRA trades a small amount of statistical memory for a more informative description of old-task sensitivity.

Second, AGRA requires computing feature gradients to construct the task sensitivity matrix after each task. This introduces an additional analysis step compared with methods that only rely on parameter updates. However, this computation is performed once per task and the resulting sensitivity information can be reused throughout subsequent training. Future work may explore more efficient approximations of the sensitivity matrix to further reduce this overhead while preserving its directional information.

Third, AGRA focuses on one specific interference mechanism: negative repulsion from current-task non-target classes. Our analysis shows that this mechanism provides an important path through which current-task optimization affects old representations under task-local training. However, forgetting in continual learning can also arise from other sources, such as changes in positive alignment, parameter drift, or distribution shifts between tasks. AGRA does not attempt to model all possible forgetting factors. Instead, it adopts a targeted strategy that identifies and reduces harmful negative updates while preserving useful optimization signals for learning new tasks. Extending AGRA to jointly consider multiple interference sources is an interesting future direction.

Fourth, AGRA currently summarizes previous tasks using an aggregated sensitivity matrix. This design keeps the memory overhead independent of the number of tasks and simplifies risk estimation. However, it may lose task-specific information when different historical tasks have substantially different sensitivity structures. Future work could investigate more adaptive aggregation strategies or task-aware sensitivity modeling to better preserve diverse historical knowledge.

Finally, AGRA is designed specifically for CLIP-based continual learning, where the shared vision encoder and cross-modal representation structure play a central role. Whether the proposed directional interference perspective can be directly extended to other types of pre-trained models or different multimodal architectures remains an open question. We believe that studying update directions relative to previously learned representation geometry provides a promising direction for broader continual learning scenarios.