CIRCE: COUNTERFACTUALLY INFORMED RELIABLE COMPONENT EVIDENCE FOR CONTINUAL PHRASE-TO-PATCH ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual adaptation enables vision language models to acquire new concepts from sequential data streams. Existing methods mainly assess whether global predictions and image-text representations remain stable, but do not determine whether a model retains the local visual evidence required to understand a phrase in a complex image. We formulate continual phrase-to-patch alignment, which studies whether a model retains phrase-conditioned local evidence as it adapts to successive stages. The setting reveals a gap between global and local retention: a model preserves its global phrase decision while the supporting patch evidence drifts or becomes diffuse. To address this gap, we propose **CIRCE**, **C**ounterfactually **I**nformed **R**eliable **C**omponent **E**vidence. CIRCE learns phrase-conditioned mixture anchors from target phrases and minimally edited counterfactuals at each stage, then freezes the acquired local state as a stage expert. For a new target phrase, CIRCE infers the most compatible stage from Gaussian summaries in the shared text space and invokes its corresponding expert for local grounding. Experiments on COCO80-CI8 and PACO-CI8 show that CIRCE achieves superior Grounded Counterfactual Accuracy and retains local phrase evidence more effectively than representative continual learning baselines.

1 INTRODUCTION

Continual multimodal learning enables vision language models to adapt as new environments and data arrive while retaining previously acquired capabilities. Existing work (Zheng et al., 2023; Liu et al., 2025; Huang et al., 2025) largely focuses on continual classification (Guo et al., 2026; Levinstein et al., 2026), retrieval (Weller et al., 2026), and global image-text matching. It typically measures forgetting through the stability of predictions on earlier classes or tasks. In long term deployment, however, vision language systems also handle fine-grained referring requests. Users ask a system not only whether an image and a text are globally compatible, but also which region corresponds to "the red backpack next to the chair" Such requests arise in interactive retrieval, visual assistance, and open world perception, where historical data cannot be retained indefinitely and fine-grained region annotations remain costly. Crucially, a correct global image-text decision does not imply that the model retains the local visual evidence supporting a particular phrase.

Existing continual multimodal methods address forgetting through three broad directions. One line uses replay, distillation, or regularization to preserve previous model behavior. A second line uses prompts, adapters, low rank modules, or parameter isolation to limit interference between new and old knowledge. A third line expands the model with dynamic modules or expert banks that retain stage specific capabilities. These approaches improve continual adaptation from the perspectives of data, parameters, or model capacity. Their objectives, however, usually preserve global predictions, holistic representations, or class level decisions. They rarely examine whether a model grounds a particular phrase in the correct local visual evidence or whether its available knowledge suits a given request when external task information is absent. As a result, stable global performance does not ensure continual fine-grained vision language understanding. This raises a fundamental question.

Can a continually adapted vision language model retain global image-text understanding while interpreting individual phrases through the correct local visual evidence?

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.

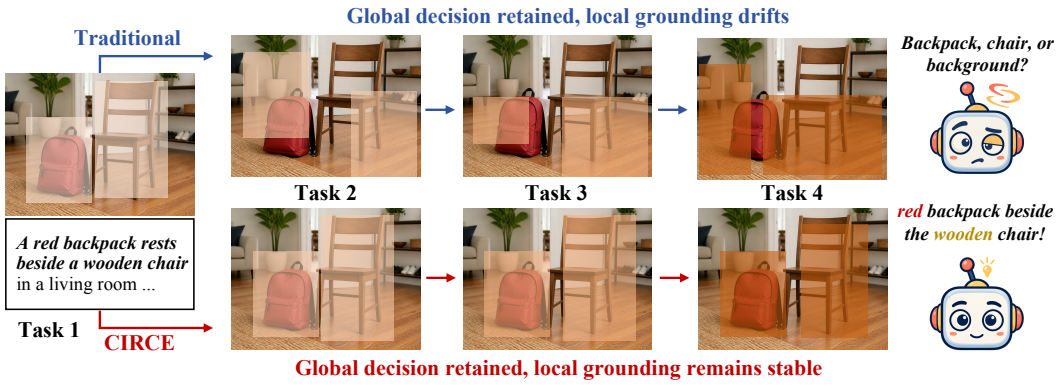


Figure 1: Conceptual illustration of continual local grounding. Traditional adaptation retains the global decision while local evidence drifts, whereas CIRCE preserves evidence on the phrase referents across later stages.

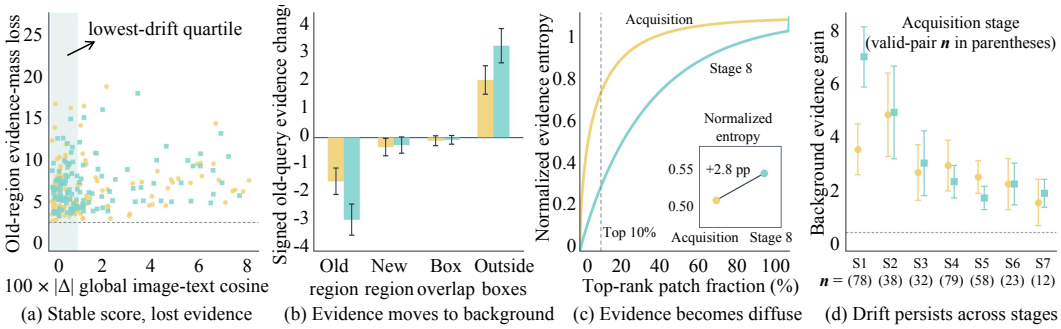
In this work, we formulate *continual phrase-to-patch alignment*, a continual learning setting in which a vision language model retains local understanding of individual phrases and associates them with minimally edited counterfactual phrases and the correct image regions. As shown in Fig.1, it presents two coupled technical challenges. **Global Local Retention Decoupling** arises because stable global image-text decisions do not ensure correct local phrase evidence. A model still prefers "the red backpack" to "the blue backpack," while the patches supporting this decision shift from the backpack to the background. Global outputs and holistic representations therefore provide no direct control over phrase region bindings. **Provenance Competence Mismatch** arises because the stage at which knowledge is acquired does not determine whether that knowledge best serves a particular image phrase request. The same concept occurs in attribute, action, and relation contexts, while task identities, region annotations, and correctness feedback remain unavailable during inference. The system therefore identifies the local knowledge that best supports the current image and query.

Building on this formulation, we present **CIRCE** (**C**ounterfactually **I**nformed **R**eliable **C**omponent **E**vidence), which preserves and invokes reliable local phrase evidence throughout continual learning. **Stage local mixture anchor acquisition** learns phrase-conditioned mixture anchors from target phrases and minimally edited counterfactuals at each stage, then freezes the completed local state as a stage expert. **Target phrase stage inference** stores a Gaussian summary of each stage in the shared text space and uses the target phrase to identify the most compatible frozen expert for a new request. CIRCE therefore preserves structured phrase-region knowledge across stages and selects it without relying on acquisition provenance or a task identity.

2 Related Work

2.1 Continual Vision Language Learning

Continual learning methods preserve previous capabilities through replay, distillation, parameter regularization, or gradient constraints (Li & Hoiem, 2016; Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Saha et al., 2021). Recent vision language methods adapt these principles through reference distillation, frozen backbone projection modules, prompts, adapters, and expert expansion (Zheng et al., 2023; Zhou et al., 2023; Wang et al., 2022b;a; Smith et al., 2023; Yu et al., 2024; Liu et al., 2025). Other work preserves modality alignment or compositional structure during sequential adaptation (Huang et al., 2025; Zhang et al., 2026b;a; Xiao et al., 2026; Zhou et al., 2026). These approaches mainly evaluate global recognition, retrieval, zero shot transfer, or compositional prediction, and do not directly assess whether a model retains the patch evidence supporting an individual phrase. Historical Prompt Calibration (Jin et al., 2024) studies exemplar free continual vision language retrieval and grounding without an inference task identity, but learns its grounding heads from box or mask supervision and selects prompts from stored task statistics. In contrast, CIRCE's focus is continual preservation and selection of fine-grained phrase region evidence rather than only global vision language behavior without region supervision or replay.



(a) Stable score, lost evidence (b) Evidence moves to background (c) Evidence becomes diffuse (d) Drift persists across stages

Figure 2: Global decisions remain stable while local phrase evidence degrades. Evidence shifts from the old region to background, becomes diffuse, and persists across stages. See Sec. 3.2.

2.2 Fine-grained Vision Language Alignment

CLIP and SigLIP learn transferable global image-text representations from large scale contrastive supervision (Radford et al., 2021; Zhai et al., 2023). Fine-grained vision language alignment builds on these representations to associate text with local visual tokens, supporting phrase level recognition, dense prediction, and open vocabulary perception. β-CLIP (Zohra et al., 2026) improves multi granularity phrase patch alignment through text conditioned spatial pooling, which CIRCE adopts as its local alignment backbone. These approaches establish that global vision language compatibility does not directly provide reliable local evidence. They generally learn local correspondence under a fixed data distribution and focus on improving the alignment of a single model. CIRCE instead studies how phrase region evidence evolves as new stages arrive.

3 Continual Phrase-to-Patch Alignment

3.1 Problem Formulation

We study *continual phrase-to-patch alignment*, where a sequence of learning stages $\{\mathcal{D}_t\}_{t=1}^{T}$ arrives over time. Each record in $\mathcal{D}_t$ contains a whole image, its caption, a target phrase q^+, and a set of minimally edited counterfactual phrases $\mathcal{Q}^- = \{q_1^-, \ldots, q_m^-\}$. For example, "a red backpack beside a wooden chair" is paired with "a blue backpack beside a wooden chair" and "a red backpack behind a wooden chair." At Stage t, the learner accesses only $\mathcal{D}_t$. It receives no box, mask, or region proposal, and neither stores nor revisits samples from preceding stages. A frozen vision encoder maps an image x to patch tokens $\{\mathbf{z}_p\}_{p=1}^{P}$, while a frozen text encoder maps a phrase q to $\mathbf{h}_q$. The grounding model produces a nonnegative phrase-conditioned evidence field $a_p(x, q)$ over patches, where $\sum_{p=1}^{P} a_p(x, q) = 1$. The objective is to retain the local evidence that supports an individual phrase as the model adapts to later stages, rather than only preserving global image-text compatibility.

At inference, the system receives only an image x and a target phrase q^+. It does not observe counterfactual phrases, a stage identity, a class label, a target location, or correctness feedback. The offline evaluator retains a hidden target region $\mathcal{B}$ and the counterfactual set $\mathcal{Q}^-$. It measures

$$\text{Point} = \mathbb{I}\left[\arg\max_p a_p(x, q^+) \in \mathcal{B}\right], \qquad \text{Disc} = \mathbb{I}\left[s(x, q^+) > \max_{q^- \in \mathcal{Q}^-} s(x, q^-)\right], \tag{1}$$

where $s(x, q)$ denotes the image phrase compatibility score, and reports Grounded Counterfactual Accuracy $(\text{GCA}) = \text{Point} \cdot \text{Disc}$. Counterfactual phrases provide training supervision and a shared evaluation probe, but do not enter the deployed grounding or state selection procedure. The task therefore requires the model to preserve reliable local phrase evidence without replay and identify a phrase compatible grounding state for the current request.

3.2 Diagnostic Analysis of Global-Local Retention Decoupling

We examine whether a preserved global vision language decision also preserves the local evidence supporting an earlier phrase. For a query q acquired at Stage s, we compare the acquisition checkpoint

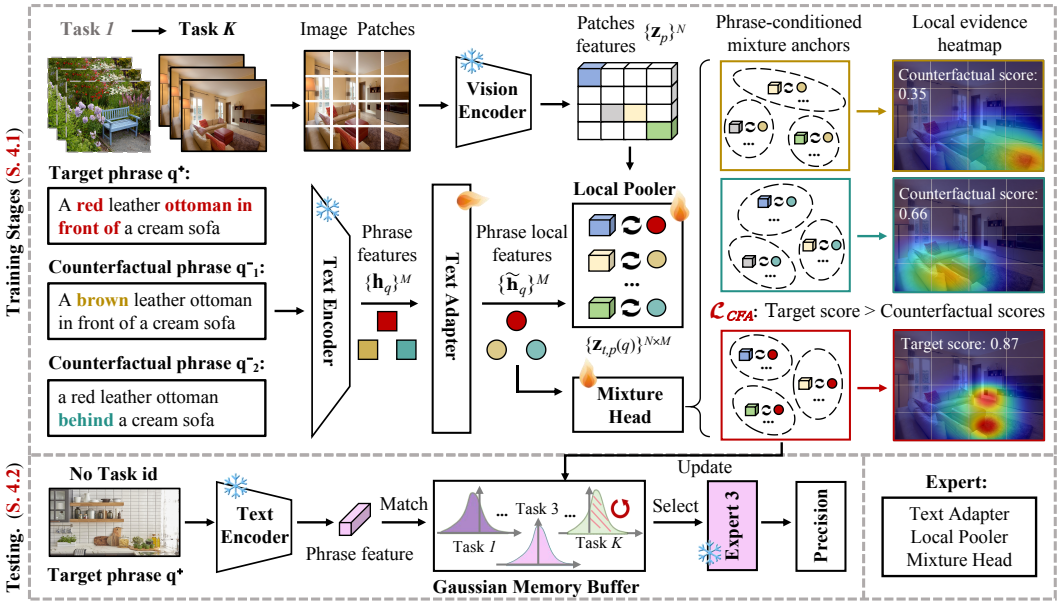


Figure 3: Overview of CIRCE. Target and counterfactual phrases guide mixture anchor learning from frozen features at each stage. The trained modules form a frozen expert. At inference, Gaussian text summaries route the target phrase to an expert that produces its local evidence heatmap.

with the final model at Stage T. At checkpoint $u \in \{s, T\}$, we define the raw patch evidence distribution as $\rho_{q,p}^{(u)} = \mathrm{softmax}_p\left(\cos\left(\mathbf{z}_p^{(u)}, \bar{\mathbf{h}}_q^{(u)}\right)/\tau_{\mathrm{d}}\right)$. The evaluator accesses the target region $\mathcal{B}$ only for this diagnostic and measures the evidence mass $\sum_{p\in\mathcal{B}} \rho_{q,p}^{(u)}$. Fig. 2(a) shows that this mass decreases even when global image-text similarity changes little. We then retain examples whose global target decision remains correct at both checkpoints. For these examples, Fig. 2(b) shows that the missing evidence moves mainly outside the target region rather than toward a valid new region. Fig. 2(c) further shows a higher spatial entropy and lower evidence concentration, while Fig. 2(d) shows positive background evidence gain across acquisition stages. Together, these results establish global-local retention decoupling. A model retains its global phrase decision while losing the compact local structure that supports it. This observation motivates stage local mixture anchor acquisition, which learns and preserves reliable local phrase evidence across stages.

4 Method

We propose **CIRCE**, a **C**ounterfactually **I**nformed **R**eliable **C**omponent **E**vidence framework for continual phrase to patch alignment. CIRCE separates local evidence retention from stage inference. As shown in Fig. 3, **Stage local mixture anchor acquisition** learns phrase-conditioned mixture anchors from the current stage using target and counterfactual phrases. It then freezes the completed local state as a stage expert, preventing later updates from directly overwriting its acquired phrase region evidence. **Target phrase stage inference** stores a Gaussian summary of target phrases for each stage in the shared frozen text space. Given a new image and target phrase, CIRCE first identifies the stage whose phrase distribution best matches the target, then invokes the corresponding frozen expert for local grounding. Standard CIRCE therefore retains local grounding states without replay and selects an expert without a task identity or counterfactual phrase at inference.

4.1 Stage Local Mixture Anchor Acquisition

The first module learns a local grounding state from the current stage and preserves it without retaining previous samples. At Stage t, CIRCE updates only the local pooling path, the dual text adapter, and the mixture anchor head using $\mathcal{D}_t$. The shared vision and text encoders remain frozen.

The module constructs phrase-conditioned local anchors, then uses them to ensure that the patches supporting a prediction also distinguish the target phrase from its counterfactuals.

To begin, CIRCE represents how each patch supports a phrase. The inherited local pooling path produces an affinity $\ell_{t,p}(q)$ for patch token $\mathbf{z}_p$ and phrase representation $\mathbf{h}_q$. It then maps this patch phrase pair to an anchor feature:

$$\mathbf{u}_{t,p}(q) = \phi_t(\mathbf{z}_p, \mathbf{h}_q) \in \mathbb{R}^{d_{\mathrm{a}}}. \tag{2}$$

This feature allows the same patch to express different local roles for phrases such as "red backpack", "blue backpack", and "backpack beside a chair." The next step organizes these phrase-conditioned features into a small number of local evidence modes.

CIRCE therefore predicts K diagonal Gaussian components from the phrase representation:

$$\{\pi_{t,k}(q), \boldsymbol{\mu}_{t,k}(q), \boldsymbol{\Sigma}_{t,k}(q), \chi_{t,k}(q)\}_{k=1}^{K} = \eta_t(\mathbf{h}_q). \tag{3}$$

Here, $\pi_{t,k}(q)$ is normalized over components, $\boldsymbol{\mu}_{t,k}(q)$ and $\boldsymbol{\Sigma}_{t,k}(q)$ define a local evidence mode, and $\chi_{t,k}(q)$ determines its contribution to grounding. These parameters define the likelihood of an anchor feature:

$$p_t(\mathbf{u} \mid q) = \sum_{k=1}^{K} \pi_{t,k}(q) \mathcal{N}\big(\mathbf{u}; \boldsymbol{\mu}_{t,k}(q), \boldsymbol{\Sigma}_{t,k}(q)\big). \tag{4}$$

The mixture likelihood alone does not identify which component explains a particular patch. CIRCE next computes its component responsibilities:

$$r_{t,p,k}(q) = \frac{\pi_{t,k}(q) \mathcal{N}\big(\mathbf{u}_{t,p}(q); \boldsymbol{\mu}_{t,k}(q), \boldsymbol{\Sigma}_{t,k}(q)\big)}{p_t\big(\mathbf{u}_{t,p}(q) \mid q\big)}. \tag{5}$$

Each component and its assigned patches form a local anchor. Multiple anchors capture complementary phrase evidence, while soft assignments retain overlap at object boundaries.

The responsibilities describe local modes, but grounding requires one normalized evidence value for every patch. CIRCE therefore combines the inherited affinity with the learned component contributions:

$$a_{t,p}(x, q) = \mathrm{softmax}_p \left(\ell_{t,p}(q) + \log \sum_{k=1}^{K} r_{t,p,k}(q) \, \mathrm{softplus}\left(\chi_{t,k}(q)\right) \right). \tag{6}$$

The evidence field now identifies the patches used by the prediction. CIRCE pools their visual features and obtains a phrase compatibility score:

$$\mathbf{v}_t(x, q) = \sum_{p=1}^{P} a_{t,p}(x, q) \mathbf{z}_p, \qquad s_t(x, q) = \cos\big(\mathbf{v}_t(x, q), \mathbf{h}_q\big). \tag{7}$$

Equations **??** through **??** produce a complete phrase grounding prediction. However, a correct image level score still allows generic background context to receive high evidence. The final step therefore connects target discrimination directly to the selected patches.

For $\mathcal{Q} = \{q^+\} \cup \mathcal{Q}^-$, CIRCE first measures the target preference of each patch:

$$\delta_{t,p} = \ell_{t,p}(q^+) - \max_{q^- \in \mathcal{Q}^-} \ell_{t,p}(q^-). \tag{8}$$

A positive margin means that a patch favors the target over every controlled alternative. CIRCE then uses these margins to train the complete grounding prediction. Let $\mathbf{s}_t(x, \mathcal{Q})$ contain the target score followed by the counterfactual scores. The counterfactual anchor loss is

$$\mathcal{L}_{\mathrm{CFA}} = \mathrm{CE}\left(\zeta \mathbf{s}_t(x, \mathcal{Q}), 0\right) - \alpha \sum_{p=1}^{P} a_{t,p}(x, q^+) \log \sigma\big(\delta_{t,p}/\tau\big), \tag{9}$$

where ζ is a learned capped score scale. The first term promotes target discrimination at the image level. The second term requires the patches receiving high evidence to favor the target over its controlled alternatives. CIRCE uses $\mathcal{L}_{\mathrm{CFA}}$ as the stage objective. After optimizing it on $\mathcal{D}_t$, CIRCE freezes the local pooler, text adapter, and mixture head as expert E_t. Later stages do not update completed experts, thereby preserving their acquired local phrase evidence.

4.2 TARGET PHRASE STAGE INFERENCE

Stage local mixture anchor acquisition preserves one local grounding expert at each stage. The remaining question is which frozen expert should answer a new request. CIRCE uses the target phrase to infer the most compatible stage before grounding the image. This choice operates in the shared frozen text space and remains independent of the local evidence scales produced by different experts. At the end of Stage t, CIRCE summarizes the target phrase representations $\{\mathbf{h}_{q_i^+}\}_{i=1}^{N_t}$ from the current stage with their empirical mean and covariance:

$$\boldsymbol{\mu}_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbf{h}_{q_i^+}, \qquad \hat{\boldsymbol{\Sigma}}_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\big(\mathbf{h}_{q_i^+}-\boldsymbol{\mu}_t\big)\big(\mathbf{h}_{q_i^+}-\boldsymbol{\mu}_t\big)^\top. \tag{10}$$

The empirical covariance captures the phrase variation within a stage, but its estimate can be poorly conditioned when the stage contains limited phrase diversity. CIRCE therefore regularizes it with a shrinkage estimate:

$$\boldsymbol{\Sigma}_t = (1-\lambda_{\mathrm{s}})\hat{\boldsymbol{\Sigma}}_t + \lambda_{\mathrm{s}}\frac{\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_t\right)}{d}\mathbf{I} + \epsilon\mathbf{I}. \tag{11}$$

CIRCE stores only $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ with expert E_t, rather than the target phrases used to estimate them.

Given a target phrase q^+, CIRCE evaluates its Gaussian log likelihood under each available stage distribution:

$$g_t(q^+) = -\frac{1}{2}\left[\big(\mathbf{h}_{q^+}-\boldsymbol{\mu}_t\big)^\top\boldsymbol{\Sigma}_t^{-1}\big(\mathbf{h}_{q^+}-\boldsymbol{\mu}_t\big) + \log|\boldsymbol{\Sigma}_t| + d\log(2\pi)\right]. \tag{12}$$

The quadratic term measures the distance from the stage phrase center, while the determinant term accounts for the spread of its phrase distribution. CIRCE then selects the highest likelihood:

$$\hat{t} = \arg\max_{e\in\{1,\ldots,t\}} g_e(q^+). \tag{13}$$

Finally, the selected $E_{\hat{t}}$ receives the image x and target phrase q^+, then returns the grounded patch:

$$\hat{p} = \arg\max_{p} a_{\hat{t},p}(x, q^+). \tag{14}$$

This procedure separates phrase-based stage inference from local image grounding. It preserves completed local experts while selecting a phrase compatible state for each target-only request.

4.3 OPTIONAL COUNTERFACTUAL CASCADE

Standard CIRCE selects an expert from the target phrase alone, without checking whether it distinguishes competing descriptions in the current image, such as "red backpack" and "blue backpack." When inference allows additional computation to construct controlled counterfactuals, an optional cascade uses them to verify the default expert and refine expert selection. Let $\hat{t}$ denote the expert selected by target phrase stage inference. CIRCE first checks whether this expert ranks the target above every counterfactual phrase:

$$m_{\hat{t}}(x, q^+) = s_{\hat{t}}(x, q^+) - \max_{q^-\in\mathcal{Q}^-} s_{\hat{t}}(x, q^-). \tag{15}$$

If $m_{\hat{t}}(x, q^+) > 0$, CIRCE returns the selected expert's grounding result. Otherwise, $\mathrm{Disc} = 0$, and CIRCE evaluates the target counterfactual margin of every available expert:

$$m_e(x, q^+) = s_e(x, q^+) - \max_{q^-\in\mathcal{Q}^-} s_e(x, q^-). \tag{16}$$

The cascade selects the expert with the strongest target preference:

$$\hat{e}_{\mathrm{CF}} = \arg\max_{e} m_e(x, q^+), \qquad e \in \{1, \ldots, t\}. \tag{17}$$

It then returns the corresponding grounded patch:

$$\hat{p}_{\mathrm{CF}} = \arg\max_{p} a_{\hat{e}_{\mathrm{CF}},p}(x, q^+). \tag{18}$$

This branch trades additional test time and counterfactual construction for a stronger expert selection signal. We report it separately as CIRCE-CF because standard CIRCE remains target-only.

Table 1: Evaluation on COCO80-CI8. Each S_t reports the average GCA over all stages observed through Stage t, measured immediately after learning Stage t. Avg. averages stage-wise results and $\mathcal{F}_{\mathrm{GCA}}$ measures forgetting.

Method	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	Avg. ↑	$\mathcal{F}_{\mathrm{GCA}}$ ↓
(a) Conventional Continual Learning										
β-CLIP	$59.80_{\pm 0.1}$	$57.40_{\pm 0.1}$	$53.90_{\pm 0.1}$	$51.23_{\pm 0.1}$	$48.76_{\pm 0.1}$	$48.10_{\pm 0.1}$	$47.04_{\pm 0.1}$	$47.86_{\pm 0.1}$	$51.76_{\pm 0.1}$	$0.00_{\pm 0.0}$
Full Tuning	$66.20_{\pm 0.3}$	$63.35_{\pm 0.4}$	$56.73_{\pm 0.4}$	$58.38_{\pm 0.4}$	$55.50_{\pm 0.4}$	$54.53_{\pm 0.5}$	$51.17_{\pm 0.5}$	$51.14_{\pm 0.5}$	$57.13_{\pm 0.4}$	$11.33_{\pm 0.3}$
LwF	$66.50_{\pm 0.3}$	$55.60_{\pm 0.4}$	$56.40_{\pm 0.3}$	$54.52_{\pm 0.4}$	$50.70_{\pm 0.4}$	$49.43_{\pm 0.4}$	$48.81_{\pm 0.4}$	$48.35_{\pm 0.4}$	$53.79_{\pm 0.4}$	$7.70_{\pm 0.3}$
EWC	$66.50_{\pm 0.2}$	$66.15_{\pm 0.3}$	$58.53_{\pm 0.3}$	$58.40_{\pm 0.3}$	$55.42_{\pm 0.4}$	$55.48_{\pm 0.4}$	$53.10_{\pm 0.4}$	$51.85_{\pm 0.4}$	$58.18_{\pm 0.3}$	$11.54_{\pm 0.3}$
GPM	$66.50_{\pm 0.2}$	$65.20_{\pm 0.2}$	$63.40_{\pm 0.2}$	$62.95_{\pm 0.2}$	$59.72_{\pm 0.3}$	$58.32_{\pm 0.3}$	$57.07_{\pm 0.3}$	$57.94_{\pm 0.3}$	$61.39_{\pm 0.2}$	$5.17_{\pm 0.2}$
(b) Parameter Efficient Continual Learning										
L2P	$58.70_{\pm 0.2}$	$56.80_{\pm 0.2}$	$53.33_{\pm 0.2}$	$51.85_{\pm 0.3}$	$48.82_{\pm 0.3}$	$47.67_{\pm 0.3}$	$47.56_{\pm 0.3}$	$48.69_{\pm 0.3}$	$51.68_{\pm 0.3}$	$0.16_{\pm 0.0}$
DualPrompt	$57.60_{\pm 0.2}$	$55.65_{\pm 0.2}$	$52.37_{\pm 0.3}$	$49.90_{\pm 0.3}$	$46.88_{\pm 0.3}$	$46.33_{\pm 0.3}$	$45.10_{\pm 0.3}$	$45.84_{\pm 0.3}$	$49.96_{\pm 0.3}$	$0.63_{\pm 0.1}$
CODA	$58.80_{\pm 0.2}$	$56.75_{\pm 0.2}$	$54.10_{\pm 0.2}$	$52.02_{\pm 0.2}$	$48.60_{\pm 0.3}$	$47.30_{\pm 0.3}$	$46.81_{\pm 0.3}$	$48.02_{\pm 0.3}$	$51.55_{\pm 0.2}$	$0.51_{\pm 0.1}$
MoE-Adapters	$56.20_{\pm 0.3}$	$53.65_{\pm 0.3}$	$48.87_{\pm 0.3}$	$48.77_{\pm 0.3}$	$44.86_{\pm 0.3}$	$44.07_{\pm 0.3}$	$43.46_{\pm 0.4}$	$43.29_{\pm 0.3}$	$47.90_{\pm 0.3}$	$1.66_{\pm 0.1}$
(c) Continual Vision Language Learning										
ZSCL	$66.20_{\pm 0.3}$	$56.75_{\pm 0.4}$	$50.30_{\pm 0.4}$	$49.73_{\pm 0.4}$	$44.32_{\pm 0.5}$	$44.13_{\pm 0.5}$	$41.14_{\pm 0.5}$	$40.80_{\pm 0.5}$	$49.17_{\pm 0.5}$	$8.46_{\pm 0.3}$
PROOF	$63.90_{\pm 0.2}$	$59.90_{\pm 0.3}$	$56.17_{\pm 0.3}$	$53.50_{\pm 0.3}$	$51.24_{\pm 0.3}$	$49.57_{\pm 0.4}$	$49.36_{\pm 0.3}$	$49.51_{\pm 0.3}$	$54.14_{\pm 0.3}$	$4.60_{\pm 0.2}$
C-CLIP	$56.80_{\pm 0.4}$	$54.10_{\pm 0.4}$	$45.33_{\pm 0.5}$	$32.08_{\pm 0.6}$	$22.64_{\pm 0.7}$	$19.20_{\pm 0.7}$	$15.73_{\pm 0.8}$	$15.64_{\pm 0.8}$	$32.69_{\pm 0.6}$	$18.16_{\pm 0.4}$
MG-CLIP	$60.00_{\pm 0.2}$	$57.00_{\pm 0.2}$	$52.77_{\pm 0.2}$	$50.65_{\pm 0.3}$	$47.84_{\pm 0.3}$	$47.37_{\pm 0.3}$	$45.99_{\pm 0.3}$	$46.90_{\pm 0.3}$	$51.06_{\pm 0.2}$	$0.41_{\pm 0.1}$
Pi-CCA	$63.80_{\pm 0.2}$	$61.95_{\pm 0.3}$	$59.88_{\pm 0.3}$	$58.71_{\pm 0.3}$	$56.45_{\pm 0.3}$	$55.86_{\pm 0.3}$	$55.20_{\pm 0.3}$	$55.54_{\pm 0.3}$	$58.42_{\pm 0.3}$	$1.92_{\pm 0.1}$
CCA-CL	$64.10_{\pm 0.2}$	$62.42_{\pm 0.2}$	$60.35_{\pm 0.3}$	$59.48_{\pm 0.3}$	$57.32_{\pm 0.3}$	$56.95_{\pm 0.3}$	$56.33_{\pm 0.3}$	$56.76_{\pm 0.2}$	$59.21_{\pm 0.3}$	$1.35_{\pm 0.1}$
Logits CL	$64.45_{\pm 0.2}$	$63.18_{\pm 0.2}$	$61.27_{\pm 0.2}$	$60.92_{\pm 0.3}$	$58.84_{\pm 0.3}$	$58.36_{\pm 0.2}$	$57.98_{\pm 0.2}$	$58.41_{\pm 0.2}$	$60.43_{\pm 0.2}$	$0.74_{\pm 0.1}$
CIRCE	$\mathbf{72.53}_{\pm 1.3}$	$71.07_{\pm 1.2}$	$68.19_{\pm 1.3}$	$67.68_{\pm 1.3}$	$64.77_{\pm 1.0}$	$64.48_{\pm 0.8}$	$63.92_{\pm 0.7}$	$64.02_{\pm 0.6}$	$67.08_{\pm 1.0}$	$0.01_{\pm 0.0}$
CIRCE-CF	$\mathbf{72.53}_{\pm 1.3}$	$\mathbf{73.28}_{\pm 1.0}$	$\mathbf{71.39}_{\pm 1.0}$	$\mathbf{71.32}_{\pm 1.2}$	$\mathbf{68.85}_{\pm 0.8}$	$\mathbf{68.99}_{\pm 0.7}$	$\mathbf{68.46}_{\pm 0.7}$	$\mathbf{68.65}_{\pm 0.5}$	$\mathbf{70.44}_{\pm 0.9}$	$\mathbf{-0.28}_{\pm 0.1}$

Table 2: Evaluation on PACO-CI8. Each S_t reports the average GCA over all stages observed through Stage t, measured immediately after learning Stage t. Avg. averages stage-wise results and $\mathcal{F}_{\mathrm{GCA}}$ measures forgetting.

Method	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	Avg. ↑	$\mathcal{F}_{\mathrm{GCA}}$ ↓
(a) Conventional Continual Learning										
β-CLIP	$16.10_{\pm 0.1}$	$15.80_{\pm 0.1}$	$18.47_{\pm 0.2}$	$18.20_{\pm 0.2}$	$18.70_{\pm 0.1}$	$18.10_{\pm 0.1}$	$18.19_{\pm 0.1}$	$17.58_{\pm 0.1}$	$17.64_{\pm 0.1}$	$0.00_{\pm 0.0}$
Full Tuning	$24.50_{\pm 0.4}$	$18.20_{\pm 0.5}$	$18.97_{\pm 0.6}$	$15.43_{\pm 0.6}$	$17.48_{\pm 0.6}$	$16.35_{\pm 0.6}$	$15.53_{\pm 0.7}$	$16.68_{\pm 0.6}$	$17.89_{\pm 0.6}$	$8.74_{\pm 0.4}$
LwF	$24.50_{\pm 0.4}$	$18.50_{\pm 0.5}$	$19.57_{\pm 0.5}$	$15.95_{\pm 0.6}$	$18.42_{\pm 0.5}$	$16.03_{\pm 0.6}$	$16.01_{\pm 0.6}$	$15.73_{\pm 0.6}$	$18.09_{\pm 0.5}$	$10.74_{\pm 0.4}$
EWC	$24.50_{\pm 0.3}$	$18.10_{\pm 0.4}$	$18.80_{\pm 0.5}$	$15.93_{\pm 0.5}$	$17.38_{\pm 0.5}$	$16.25_{\pm 0.5}$	$15.49_{\pm 0.6}$	$15.64_{\pm 0.6}$	$17.76_{\pm 0.5}$	$9.30_{\pm 0.4}$
GPM	$22.20_{\pm 2.1}$	$21.37_{\pm 1.4}$	$24.49_{\pm 1.2}$	$23.18_{\pm 0.6}$	$22.78_{\pm 0.6}$	$21.88_{\pm 0.5}$	$22.78_{\pm 0.5}$	$21.43_{\pm 0.2}$	$22.51_{\pm 0.7}$	$2.43_{\pm 0.4}$
(b) Parameter-Efficient Continual Learning										
L2P	$16.90_{\pm 0.2}$	$16.10_{\pm 0.2}$	$20.07_{\pm 0.3}$	$18.50_{\pm 0.2}$	$20.96_{\pm 0.3}$	$19.23_{\pm 0.3}$	$18.99_{\pm 0.2}$	$17.60_{\pm 0.2}$	$18.54_{\pm 0.2}$	$2.29_{\pm 0.1}$
DualPrompt	$15.80_{\pm 0.3}$	$5.80_{\pm 0.3}$	$8.10_{\pm 0.4}$	$5.50_{\pm 0.4}$	$12.32_{\pm 0.4}$	$11.57_{\pm 0.4}$	$7.04_{\pm 0.4}$	$9.89_{\pm 0.4}$	$9.50_{\pm 0.4}$	$2.09_{\pm 0.2}$
CODA-Prompt	$17.60_{\pm 0.2}$	$13.35_{\pm 0.3}$	$11.13_{\pm 0.3}$	$13.50_{\pm 0.3}$	$16.74_{\pm 0.3}$	$18.27_{\pm 0.2}$	$15.49_{\pm 0.3}$	$17.46_{\pm 0.2}$	$15.44_{\pm 0.3}$	$0.09_{\pm 0.0}$
MoE-Adapters	$18.80_{\pm 0.3}$	$17.95_{\pm 0.3}$	$20.86_{\pm 0.4}$	$19.43_{\pm 0.4}$	$20.72_{\pm 0.4}$	$19.68_{\pm 0.3}$	$19.34_{\pm 0.3}$	$18.83_{\pm 0.3}$	$19.45_{\pm 0.4}$	$1.24_{\pm 0.1}$
(c) Continual Vision–Language Learning										
ZSCL	$22.40_{\pm 0.4}$	$18.25_{\pm 0.4}$	$18.72_{\pm 0.5}$	$16.80_{\pm 0.5}$	$17.35_{\pm 0.5}$	$16.72_{\pm 0.5}$	$16.05_{\pm 0.5}$	$15.88_{\pm 0.5}$	$17.77_{\pm 0.5}$	$6.38_{\pm 0.3}$
PROOF	$21.90_{\pm 0.3}$	$19.35_{\pm 0.3}$	$21.48_{\pm 0.4}$	$20.05_{\pm 0.3}$	$20.74_{\pm 0.3}$	$19.82_{\pm 0.3}$	$19.56_{\pm 0.3}$	$19.24_{\pm 0.3}$	$20.27_{\pm 0.3}$	$1.86_{\pm 0.1}$
C-CLIP	$11.90_{\pm 0.3}$	$10.65_{\pm 0.4}$	$9.80_{\pm 0.4}$	$5.78_{\pm 0.5}$	$4.00_{\pm 0.5}$	$2.08_{\pm 0.6}$	$1.19_{\pm 0.6}$	$0.89_{\pm 0.6}$	$5.79_{\pm 0.5}$	$5.41_{\pm 0.3}$
MG-CLIP	$15.70_{\pm 0.1}$	$16.20_{\pm 0.2}$	$18.93_{\pm 0.2}$	$18.25_{\pm 0.2}$	$18.90_{\pm 0.2}$	$18.32_{\pm 0.2}$	$18.06_{\pm 0.2}$	$17.45_{\pm 0.2}$	$17.73_{\pm 0.2}$	$0.26_{\pm 0.0}$
Pi-CCA	$21.30_{\pm 0.2}$	$19.55_{\pm 0.3}$	$22.08_{\pm 0.3}$	$20.86_{\pm 0.3}$	$21.42_{\pm 0.3}$	$20.65_{\pm 0.3}$	$20.58_{\pm 0.3}$	$20.12_{\pm 0.3}$	$21.07_{\pm 0.3}$	$0.92_{\pm 0.1}$
CCA-CL	$21.60_{\pm 0.2}$	$20.08_{\pm 0.3}$	$22.52_{\pm 0.3}$	$21.35_{\pm 0.3}$	$21.81_{\pm 0.3}$	$21.22_{\pm 0.3}$	$21.05_{\pm 0.2}$	$20.58_{\pm 0.2}$	$21.28_{\pm 0.3}$	$0.71_{\pm 0.1}$
Logits CL	$21.95_{\pm 0.2}$	$20.53_{\pm 0.2}$	$22.86_{\pm 0.3}$	$21.72_{\pm 0.3}$	$22.35_{\pm 0.3}$	$21.66_{\pm 0.2}$	$21.54_{\pm 0.2}$	$21.14_{\pm 0.2}$	$21.72_{\pm 0.2}$	$0.48_{\pm 0.1}$
CIRCE	$\mathbf{25.57}_{\pm 0.50}$	$23.28_{\pm 0.62}$	$26.01_{\pm 0.66}$	$24.36_{\pm 0.64}$	$25.27_{\pm 0.59}$	$24.85_{\pm 0.57}$	$24.90_{\pm 0.55}$	$24.55_{\pm 0.56}$	$24.85_{\pm 0.59}$	$0.00_{\pm 0.00}$
CIRCE-CF	$\mathbf{25.57}_{\pm 0.50}$	$\mathbf{24.30}_{\pm 0.61}$	$\mathbf{27.63}_{\pm 0.67}$	$\mathbf{26.49}_{\pm 0.65}$	$\mathbf{27.64}_{\pm 0.60}$	$\mathbf{27.24}_{\pm 0.58}$	$\mathbf{27.59}_{\pm 0.56}$	$\mathbf{27.14}_{\pm 0.57}$	$\mathbf{26.70}_{\pm 0.60}$	$\mathbf{-0.08}_{\pm 0.06}$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and Evaluation Protocol. We evaluate on COCO80-CI8, derived from MS COCO, and PACO-CI8, derived from PACO part and attribute annotations. Both benchmarks comprise eight stages with 5,000 training and 1,000 evaluation records per stage. Training uses whole images and target and counterfactual phrases without region supervision or image replay. Standard inference receives only an image and its target phrase. The offline evaluator uses boxes and counterfactuals to compute Point, discrimination accuracy, and GCA (Eq. 1). CIRCE-CF additionally uses counterfactuals at inference and is reported separately. Dataset construction and protocol details appear in the appendix.

Implementation Details. CIRCE initializes from CLIP ViT-B/16 and keeps its shared vision and text encoders frozen throughout continual learning. At each stage, it trains the stage local pooler, text adapter, and mixture anchor head using only the arriving data. It then freezes these modules as a

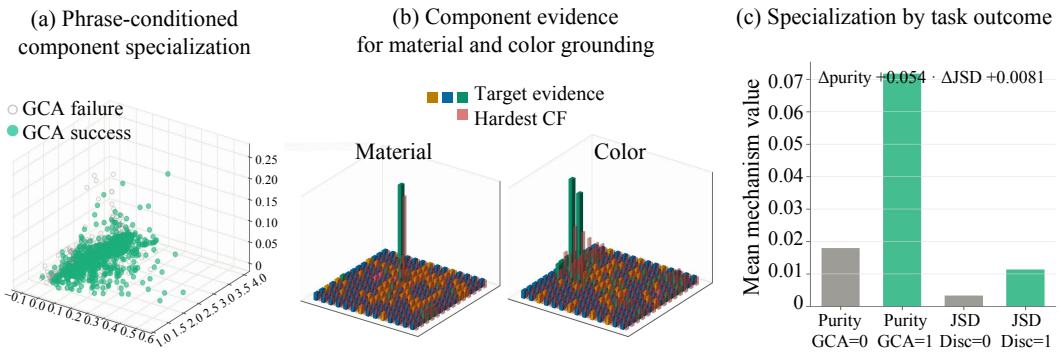


Figure 4: **Phrase-conditioned mixture-anchor acquisition.** On PACO-CI8, the panels (a) (b) (c) show global component geometry, material and color evidence landscapes, and the relationship between component structure and grounding outcomes.

stage expert and stores the mean and shrinkage covariance of the stage target phrase embeddings for target phrase stage inference. We optimize the trainable modules with AdamW and cosine learning rate decay, using the same stage-wise training budget across CIRCE variants. We provide the number of mixture components, loss weights, learning rates, and training epochs in the appendix.

Baselines. We compare CIRCE with several groups of continual learning methods. *Frozen β-CLIP* (Zohra et al., 2026) applies the pretrained local alignment model without continual adaptation, while *Sequential Full Fine Tuning* updates the model on each incoming stage and serves as a direct forgetting baseline. We include replay-free regularization and subspace methods, including Learning without Forgetting (LwF) (Li & Hoiem, 2016), Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), and Gradient Projection Memory (GPM) (Saha et al., 2021). We further evaluate parameter-isolation methods, including Learning to Prompt (L2P) (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), CODA-Prompt (Smith et al., 2023), and MoE-Adapters (Yu et al., 2024). Continual vision–language baselines include ZSCL (Zheng et al., 2023), PROOF (Zhou et al., 2023), C-CLIP (Liu et al., 2025), MG-CLIP (Huang et al., 2025), Pi-CCA (Zhang et al., 2026b), CCA-CL (Zhang et al., 2026a), and the logits-based framework of Wang & Zhu (2026). All methods use the same pretrained initialization, continual data stream, phrase-to-patch scoring protocol, and target-only inference interface.

5.2 Main results

Tables 1 and 2 show that CIRCE consistently improves continual phrase-to-patch alignment across both object-centric and part-centric streams. On COCO80-CI8, standard CIRCE achieves the best target-only performance, reaching 67.08 average GCA and 64.02 final-stage GCA, outperforming the strongest baseline by a clear margin while nearly eliminating measured forgetting. The same pattern holds on the harder PACO-CI8 benchmark, where absolute scores are lower due to fine-grained part, material, and pattern distinctions, but CIRCE still obtains the highest target-only average and final GCA. CIRCE-CF further improves both datasets by using counterfactual phrases at inference, indicating that counterfactual verification provides useful expert-selection signal beyond target-only stage inference.

6 Additional Analysis

Local Structure of Mixture Anchors. Fig. 4(a) shows that the model effectively uses approximately three of four components and achieves positive target-region purity relative to a spatial-shift null, indicating component utilization and localization beyond incidental overlap. With the image and expert fixed, Fig. 4(b) shows that changing the queried material or color reorganizes component evidence, revealing sensitivity to fine-grained phrase semantics. Fig. 4(c) further links this structure to performance. Target purity is 0.054 higher for successful GCA predictions, while target–counterfactual component-map JSD is 0.0081 higher for successful discrimination. These associations support the relevance of mixture anchors to local grounding and discrimination. Boxes and outcomes are used only for offline analysis.

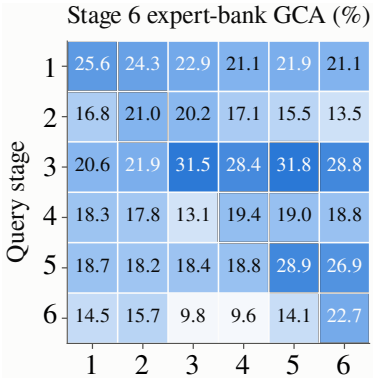


Figure 5: Expert GCA Across Stages.

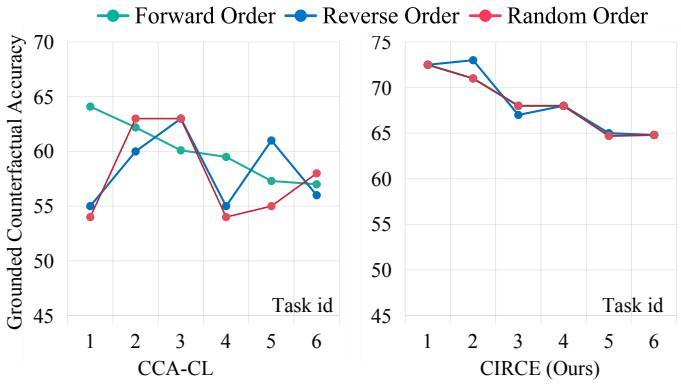


Figure 6: Robustness to forward, reverse, and random stage orders.

Table 3: Ablation study of CIRCE under the target-only inference protocol.

Variant	COCO80-CI8			PACO-CI8		
	Avg. ↑	Final ↑	$\mathcal{F}_{\text{GCA}}$ ↓	Avg. ↑	Final ↑	$\mathcal{F}_{\text{GCA}}$ ↓
Single anchor ($K = 1$)	$64.91_{\pm 0.86}$	$61.72_{\pm 0.64}$	$0.18_{\pm 0.05}$	$23.41_{\pm 1.11}$	$22.96_{\pm 0.94}$	$0.05_{\pm 0.03}$
w/o component contribution weights χ	$65.58_{\pm 0.91}$	$62.47_{\pm 0.61}$	$0.09_{\pm 0.04}$	$23.76_{\pm 1.18}$	$23.31_{\pm 1.01}$	$0.03_{\pm 0.02}$
w/o patch counterfactual term	$64.12_{\pm 0.95}$	$60.88_{\pm 0.72}$	$0.42_{\pm 0.10}$	$23.05_{\pm 1.24}$	$22.41_{\pm 1.07}$	$0.16_{\pm 0.06}$
Mean-only stage routing	$65.96_{\pm 0.88}$	$62.91_{\pm 0.66}$	$0.12_{\pm 0.04}$	$24.03_{\pm 1.20}$	$23.61_{\pm 1.04}$	$0.02_{\pm 0.02}$
Most-recent expert E_t	$62.73_{\pm 0.83}$	$59.34_{\pm 0.70}$	$2.31_{\pm 0.28}$	$22.18_{\pm 1.05}$	$21.52_{\pm 0.91}$	$1.14_{\pm 0.21}$
CIRCE	$\mathbf{67.08}_{\pm 0.99}$	$\mathbf{64.02}_{\pm 0.58}$	$\mathbf{0.01}_{\pm 0.00}$	$\mathbf{24.85}_{\pm 0.59}$	$\mathbf{24.55}_{\pm 0.56}$	$\mathbf{0.00}_{\pm 0.00}$

Expert Specialization Across Stages. Fig. 5 evaluates each frozen expert on queries from all six stages after Stage 6. The diagonal generally achieves higher GCA, indicating that experts retain capabilities specialized to their acquisition stages. However, acquisition provenance does not always identify the strongest expert. For Stage 3 queries, Expert 5 achieves 31.8% GCA, slightly exceeding Expert 3 at 31.5%. This cross-stage variation motivates expert selection, while the routing comparison separately evaluates the effectiveness of target phrase stage inference.

Robustness to Stage Order. Fig. 6 compares CCA-CL and CIRCE under forward, reverse, and random stage orders. CCA-CL exhibits substantial variation across orders, indicating sensitivity to the sequence of incoming data. In contrast, CIRCE maintains closely aligned GCA trajectories, with nearly identical results at the final two stages. CIRCE also achieves higher GCA across all displayed stages and orders. These results demonstrate more consistent local alignment performance under changes in the continual learning sequence.

Ablation Study. Table 3 verifies both CIRCE modules under target-only inference. Replacing the mixture bank with a single anchor or removing learned component contributions consistently reduces GCA on both datasets, showing the value of multiple adaptive local evidence modes. Removing the patch counterfactual term causes the largest acquisition drop and increases forgetting, which confirms that global target discrimination alone does not preserve locally discriminative evidence. Mean-only routing also underperforms Gaussian stage inference, while always selecting the most recent expert yields the largest degradation and forgetting. These results show that local mixture anchors and target phrase stage inference contribute complementary gains.

7 Conclusion

We introduce continual phrase-to-patch alignment and identify a gap between global decision retention and local evidence preservation. CIRCE learns stage local mixture anchors and uses target phrase stage inference to select frozen experts. Experiments on COCO80-CI8 and PACO-CI8 demonstrate improved GCA and robustness to stage order. Limitations include expert storage that grows with the number of stages and phrase-based routing that omits visual context during expert selection. Future work explores compact expert consolidation and efficient image-conditioned routing.

AI USE STATEMENT

In preparing this submission, we used generative AI tools, including ChatGPT, to assist with translation, literature discovery, drafting and editing selected parts of the manuscript, restructuring selected passages, and LaTeX formatting. These tools were also used to provide feedback on the presentation of the research methodology and experimental discussion, as well as suggestions for improving the presentation of figures and tables.

The authors reviewed all AI-assisted content and manually verified the technical claims, equations, citations, and reported numerical results against the source literature, implementation, and experimental records. Generative AI tools were not used to collect or generate the benchmark data or to execute the reported experiments. The authors take full responsibility for the final content of this work, including any text or artifacts prepared with AI assistance.

ETHICS STATEMENT

This work studies continual phrase-to-patch alignment using COCO80-CI8 and PACO-CI8, constructed from existing public datasets and annotations. It does not involve new human-subject research or the collection of new personal information. We credit the original data sources and use them in accordance with their applicable licenses and terms of use.

CIRCE may inherit biases from the source datasets and pretrained encoders. Its evidence maps may localize incorrect regions or fail to distinguish fine-grained descriptions, and should not be interpreted as guaranteed explanations of model reasoning. The reported results do not establish safe or fair performance across deployment domains. Applications involving people or consequential decisions require additional evaluation of accuracy, fairness, privacy, and robustness.

REPRODUCIBILITY STATEMENT

Section 3 defines continual phrase-to-patch alignment and its evaluation setting. Section 4 describes CIRCE's mixture-anchor learning objective, target-phrase stage inference, and optional counterfactual cascade. Section 5 presents the experimental setup and main comparisons. The appendix provides dataset construction details, stage assignments, image-level split isolation, evaluation metrics, implementation hyperparameters, diagnostic protocols, and training and inference pseudocode. Random seeds are specified, and repeated-run results are reported using means and sample standard deviations where applicable.

We distinguish standard CIRCE, which receives only an image and a target phrase at inference, from CIRCE-CF, which additionally uses counterfactual phrases. Hidden spatial annotations are reserved for offline evaluation and analysis. An anonymized code package accompanies this submission, including preprocessing, training, and evaluation code together with the experimental configurations, to support reproduction of the benchmark construction and both inference protocols.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, pp. 139–154, 2018.

Yifu Guo, Yuquan Lu, Wentao Zhang, Zishan Xu, Dexia Chen, Siyu Zhang, Yizhe Zhang, and Ruixuan Wang. Decoupling continual semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4476–4484, 2026.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. MG-CLIP: Continual learning for vision-language models via modality gap. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3777–3786, 2025.

Tao Jin, Weicai Yan, Ye Wang, Sihang Cai, Qifan Shuai, and Zhou Zhao. Calibrating prompt from history for continual vision-language retrieval and grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

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*, 114 (13):3521–3526, 2017.

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Daniel Soudry, Tomer Koren, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. *Advances in Neural Information Processing Systems*, 38:60456–60495, 2026.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629, 2016.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-CLIP: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023.

Sijie Wang and Yingying Zhu. Task-specific knowledge improves generalization: A logits-based framework for continual learning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7615–7624, 2026.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648, 2022a.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 11202–11216, 2026.

Canran Xiao, Tianxiang Xu, Siyuan Ma, Yiyang Jiang, Haoyu Gao, and Yuhan Wu. Reversible primitive–composition alignment for continual vision–language learning. In *International Conference on Learning Representations*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3987–3995, 2017.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Huan Zhang, Shuyu Dong, Yujin Zheng, Dingwen Wang, Shenghua Fan, and Fan Lyu. Subspace alignment for clip-based continual learning via canonical correlation analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32257–32266, 2026a.

Jiayu Zhang, Chuangxin Zhao, Canran Xiao, Ruibo Duan, Wenyi Mo, Haoyu Gao, and Wenshuo Wang. Pi-CCA: Prompt-invariant CCA certificates for replay-free continual multimodal learning. In *International Conference on Learning Representations*, 2026b.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19125–19136, 2023.

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. *arXiv preprint arXiv:2305.19270*, 2023.

Heng Zhou, Jing Tang, Jusheng Zhang, Yanshu Li, Canran Xiao, Liwei Hou, Zong Ke, and Jiawei Yao. CoMem: Compositional concept-graph memory for vision–language adaptation. In *International Conference on Learning Representations*, 2026.

Fatimah Zohra, Chen Zhao, Hani Itani, and Bernard Ghanem. Beta-CLIP: Text-conditioned contrastive learning for multi-granular vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 680–689, 2026.

A Appendix

A.1 Dataset Construction

We construct two benchmarks for continual phrase-to-patch alignment: COCO80-CI8 and PACO-CI8. Both benchmarks contain eight sequential stages. At Stage t, the learner receives 5,000 training records from the current stage only. After learning that stage, it is evaluated on the union of the held-out records from stages $1{:}t$. Each stage contributes 1,000 evaluation records, giving 40,000 training records and 8,000 evaluation records in total. A training record contains a whole image, a caption or phrase context, a caller-designated target phrase q^+, and a set of minimally edited counterfactual phrases $\mathcal{Q}^-$. It contains no box, mask, task identity, or correctness feedback. Spatial annotations are retained only by the offline evaluator.

COCO80-CI8. COCO80-CI8 uses the 80 MS COCO object categories and partitions them into eight disjoint class-incremental stages, with ten newly introduced categories per stage. The class order is fixed before training and is not selected from experimental outcomes:

S_1 : {bench, laptop, cat, elephant, umbrella, motorcycle, airplane, carrot, broccoli, oven},
S_2 : {bicycle, bottle, horse, bear, skateboard, boat, couch, refrigerator, teddy bear, tennis racket},
S_3 : {car, cup, scissors, sports ball, train, banana, bed, skis, suitcase, dining table},
S_4 : {cell phone, spoon, handbag, cow, sheep, pizza, toothbrush, apple, sandwich, fire hydrant},
S_5 : {chair, book, mouse, bird, frisbee, snowboard, fork, parking meter, hair drier, baseball glove},
S_6 : {dog, bowl, microwave, zebra, tie, cake, toilet, potted plant, stop sign, toaster},
S_7 : {knife, television, vase, giraffe, kite, backpack, sink, keyboard, orange, traffic light},
S_8 : {person, clock, remote, surfboard, bus, truck, donut, baseball bat, wine glass, hot dog}.

For each category, we retain 600 image-unique records, with 500 assigned to training and 100 assigned to evaluation. Each stage therefore contains 5,000 training records and 1,000 evaluation records. The records are balanced across object, attribute, action, and relation queries, with 1,250 training records and 250 evaluation records from each family per stage. Images are globally unique across the complete stream: no image occurs in both training and evaluation.

PACO-CI8. PACO-CI8 is derived from the PACO-LVIS training annotations (**?**) and targets fine-grained grounding beyond whole-object recognition. We retain part boxes covering at least 0.4% of the image and part categories with at least 50 usable annotations. A parent object must contain at least four retained sibling parts, allowing each part query to be paired with three controlled alternatives. We select the 32 parent objects with the greatest remaining annotation capacity and assign four objects to each stage using a fixed snake allocation over annotation counts. The selected annotations cover 187 part categories. The resulting parent-object schedule is fixed before model evaluation:

S_1 : {dog, table, telephone, television set},
S_2 : {bottle, can, chair, handbag},
S_3 : {cooking pan, plastic bag, plate, tray},
S_4 : {basket, car, cellular telephone, microwave oven},
S_5 : {bowl, bucket, lamp, vase},
S_6 : {cup, drink glass, mug, sweater},
S_7 : {bench, box, computer mouse, trash can},
S_8 : {bicycle, jar, laptop computer, shoe}.

PACO-CI8 contains four fine-grained query families: part, color, material, and pattern. A part query identifies one annotated component of a parent object. The other three families retain the same parent object and part while changing only the corresponding PACO attribute. Each stage contains 1,250 training records and 250 evaluation records from each family, matching the COCO80-CI8 stage budget. Image isolation is enforced before record sampling. We first collect the stages of all selected parent objects annotated in each image and discard an image if it contains selected objects assigned to different stages. This removes 6,001 ambiguous images and leaves 19,772 single-stage candidate

Table 4: Statistics and stream structure of the continual phrase-to-patch alignment benchmarks. Slash-separated counts denote training/evaluation.

Dataset	Stages	Stage units	Label inventory	Query families	Records/stage	Unique images
COCO80-CI8	8	10 object classes	80 objects	Object/attribute/action/relation	5,000/1,000	40,000/8,000
PACO-CI8	8	4 parent objects	32 parents, 187 parts	Part/color/material/pattern	5,000/1,000	13,243/3,090

images. Within each stage, a seed-17 stable hash assigns 20% of the images to evaluation and the remainder to training. Records are sampled only after this image-level stage and split assignment. The final 40,000 training records and 8,000 evaluation records use 13,243 and 3,090 unique images, respectively. The executable audit verifies zero train–evaluation image overlap, zero cross-stage image reuse in either split, unique sample and query identifiers, exact stage and family quotas, and zero missing image files. An earlier pilot split separated records by object rather than globally by image and failed this requirement. All results in this paper use the finalized PACO-CI8 construction described here; no result from the pilot split is retained.

Phrase and counterfactual generation. Counterfactual generation is dataset specific but follows a shared principle: each alternative preserves the phrase frame and replaces one controlled semantic slot. In COCO80-CI8, the four query families modify the object, attribute, action, or relation slot. For example, an attribute alternative may change "a red backpack" to "a blue backpack" while preserving the object noun. In PACO-CI8, a part query has the form "the [part] of the [object]" and replaces the target part with three sibling parts of the same parent object. Color, material, and pattern queries retain both the parent object and the part while replacing only the relevant PACO attribute. For example, "the brown side of the basket" may be paired with "the dark-grey side of the basket." A replacement may contain multiple words; minimality refers to one controlled semantic slot rather than one lexical token. For compatibility with the shared data loader, the PACO part, color, material, and pattern families are serialized internally through the object, attribute, action, and relation slots, respectively. These internal names do not imply that PACO provides action or spatial relation annotations.

Spatial evaluation annotations. Target regions are stored only in the evaluation manifests and are never exposed during training or inference. COCO80-CI8 uses the query-specific object or entity regions produced by its annotation pipeline. When a relation phrase refers to multiple participating entities, the evaluator accepts the corresponding combined valid region. In PACO-CI8, all four query families use the PACO box of the queried part. Part, color, material, and pattern queries can therefore share the same spatial target while differing in their semantic discrimination requirement. This design separates finding the correct local region from distinguishing controlled descriptions of that region. During evaluation, the predicted location is the center of the patch receiving maximum phrase-conditioned evidence. Localization is correct when that center lies inside a hidden target region. If a query has multiple valid target regions, a prediction inside any of them is accepted.

Training and inference interface. At Stage t, the model receives only current-stage images, target phrases, and their controlled alternatives. It receives no region supervision and does not replay images, text records, or per-example features from earlier stages. At standard inference, the model receives only an image and the target phrase. It does not receive counterfactual phrases, the stage identity, the class label, the target region, or correctness feedback. CIRCE-CF is reported separately because it additionally receives the controlled counterfactual phrases at inference for expert verification. Hidden regions and counterfactual sets are shared across methods only through the offline evaluator.

A.2 Diagnostic Protocol and Mechanism Metrics

We use two offline analyses: raw evidence retention across checkpoints (Figure 2) and phrase-conditioned component structure (Figure 4). Region annotations, acquisition stages, and prediction outcomes are used only for diagnosis and are not supplied to the deployed grounding procedure.

Raw evidence and global score drift. Figure 2 evaluates Full Tuning and LwF on held-out COCO80-CI8 pairs using training seed 17. For a query acquired at stage s, we compare its acquisition checkpoint with the final checkpoint $T = 8$. The diagnostic bypasses the local pooler, text adapter,

and mixture-anchor module. At checkpoint $u \in \{s, T\}$, raw patch evidence is

$$\rho_{q,p}^{(u)} = \mathrm{softmax}_p \left(\cos(\mathbf{z}_p^{(u)}, \mathbf{h}_q^{(u)}) / \tau_d \right), \qquad \tau_d = 0.07.$$

This distribution measures raw patch–text alignment and is distinct from CIRCE's learned evidence field $a_p(x, q)$. Global score drift is

$$D_{\mathrm{global}} = \left| \cos(\mathbf{v}_{\mathrm{global}}^{(T)}, \mathbf{h}_q^{(T)}) - \cos(\mathbf{v}_{\mathrm{global}}^{(s)}, \mathbf{h}_q^{(s)}) \right|.$$

Spatial evidence redistribution. Each pair contains old and new regions in the same image. We partition the image domain Ω into

$$R_{\mathrm{old}} = B_{\mathrm{old}} \setminus B_{\mathrm{new}}, \qquad R_{\mathrm{new}} = B_{\mathrm{new}} \setminus B_{\mathrm{old}},$$
$$R_{\mathrm{overlap}} = B_{\mathrm{old}} \cap B_{\mathrm{new}}, \quad R_{\mathrm{outside}} = \Omega \setminus (B_{\mathrm{old}} \cup B_{\mathrm{new}}).$$

A pair is valid when both exclusive regions contain at least one patch center. For any region R, define

$$M^{(u)}(R) = \sum_{p : c_p \in R} \rho_{q,p}^{(u)}, \qquad \Delta M(R) = M^{(T)}(R) - M^{(s)}(R).$$

Old-region evidence loss is $L_{\mathrm{old}} = -\Delta M(R_{\mathrm{old}})$. Figure 2(a) compares this loss with global score drift over all valid pairs, highlighting the lowest-drift quartile within each acquisition stage.

Figures 2(b,c) retain pairs whose old query outranks its paired swap query at both checkpoints. This condition preserves one controlled global ordering, rather than requiring correctness against every counterfactual. Figure 2(b) reports signed regional mass changes. Evidence outside the paired boxes is described as outside-region evidence, since it may include unannotated visual content.

Evidence concentration and aggregation. Spatial diffuseness and top-ranked evidence mass are

$$H^{(u)} = -\frac{1}{\log P} \sum_p \rho_{q,p}^{(u)} \log \rho_{q,p}^{(u)}, \qquad C^{(u)}(r) = \sum_{j=1}^{\max(1, \mathrm{round}(rP))} \rho_{q,(j)}^{(u)},$$

where $\rho_{q,(j)}^{(u)}$ is the j-th largest patch value. An increase in entropy or a decrease in $C^{(u)}(0.1)$ indicates less concentrated evidence. Figures 2(b,c) use acquisition-stage macro averages within the retained-order subset. Figure 2(d) uses all valid pairs and reports outside-region mass changes by acquisition stage. Its uncertainty intervals resample complete image clusters; the displayed counts are valid pair counts.

Component utilization. Figure 4 analyzes PACO-CI8 queries using their acquisition-stage experts, holding the expert fixed across phrase comparisons. This isolates component structure from target-only expert selection. For patch responsibilities $r_{p,k}(q)$, define

$$w_k(q) = \frac{1}{P} \sum_p r_{p,k}(q), \qquad K_{\mathrm{eff}}(q) = \exp\left(-\sum_k w_k(q) \log w_k(q) \right).$$

$K_{\mathrm{eff}} \in [1, K]$ measures how broadly responsibility is distributed across components. It does not by itself rule out redundant component maps.

Spatially adjusted component purity. Normalize each component over patches and measure its overlap with the hidden target region:

$$b_{p,k}(q) = \frac{r_{p,k}(q)}{\sum_j r_{j,k}(q)}, \qquad P_{\max}(q, B) = \max_k \sum_{p : c_p \in B} b_{p,k}(q).$$

We subtract a circular-shift control to account for incidental spatial overlap:

$$P_{\mathrm{adj}}(q, B) = P_{\max}(q, B) - \frac{1}{|\mathcal{U}_q|} \sum_{\Delta \in \mathcal{U}_q} \max_k \sum_{p : c_p \in B} [S_\Delta b_k(q)]_p.$$

Here S_Δ shifts the component maps while keeping the box fixed. $\mathcal{U}_q$ contains up to 64 nonzero shifts, sampled deterministically from the query identifier. Positive adjusted purity indicates stronger target alignment than under this spatial control.

Target–counterfactual divergence. The hardest counterfactual is selected using the inherited base evidence score before inspecting mixture-anchor output:

$$q^-_{\text{hard}} = \arg\max_{q^- \in \mathcal{Q}^-} s_{\text{base}}(x, q^-).$$

We compare normalized component maps using Jensen–Shannon divergence,

$$\text{JSD}(b, c) = \tfrac{1}{2}\,\text{KL}(b\|m) + \tfrac{1}{2}\,\text{KL}(c\|m), \qquad m = (b + c)/2.$$

To remove dependence on component ordering, we use the minimum-cost matching:

$$D_{\text{comp}} = \min_{\pi \in \mathfrak{S}_K} \frac{1}{K} \sum_{k=1}^{K} \text{JSD}\left(b_k(q^+), b_{\pi(k)}(q^-_{\text{hard}})\right).$$

For $K = 4$, all 24 permutations are evaluated. Higher divergence indicates a larger phrase-dependent change in spatial component structure.

Association with prediction outcomes. The component diagnostic was developed using training seed 17; Figure 4 reports results from seeds 7 and 23. We compare mean adjusted purity between GCA-successful and unsuccessful queries, and mean component divergence between Disc-successful and unsuccessful queries. The respective differences are 0.0538 and 0.00806. These are descriptive associations between component structure and prediction outcomes, rather than evidence of a causal effect.

A.3 Evaluation Metrics

We evaluate continual phrase-to-patch alignment with metrics that require both local grounding and fine-grained counterfactual discrimination. Each evaluation record contains an image x, a target phrase q^+, a set of minimally edited counterfactual phrases $Q^- = \{q^-_1, \ldots, q^-_m\}$, and a hidden target region B. The model predicts a phrase-conditioned evidence distribution $a_p(x, q^+)$ over image patches and an image–phrase compatibility score $s(x, q)$ for each phrase. The hidden box B and the counterfactual set are used only by the offline evaluator, except for the separately reported CIRCE-CF variant, which also uses counterfactual phrases at inference.

Pointing accuracy. Pointing accuracy measures whether the most strongly supported patch for the target phrase falls inside the hidden target region. Let c_p denote the normalized center coordinate of patch p. We define

$$\text{Point}(x, q^+, B) = \mathbb{1}\left[c_{\arg\max_p a_p(x,q^+)} \in B\right].$$

Thus, a prediction is locally correct only if the peak target evidence lies in the annotated target region. For records with multiple valid target regions, the prediction is counted as correct if the selected patch center lies inside any valid region.

Counterfactual discrimination. Discrimination accuracy measures whether the model assigns the target phrase a higher compatibility score than all controlled counterfactual phrases:

$$\text{Disc}(x, q^+, Q^-) = \mathbb{1}\left[s(x, q^+) > \max_{q^- \in Q^-} s(x, q^-)\right].$$

This metric prevents a model from receiving credit for localizing a plausible region while failing to distinguish the target phrase from minimally edited alternatives.

Grounded Counterfactual Accuracy. Our primary metric is Grounded Counterfactual Accuracy (GCA), which requires both localization and discrimination to be correct:

$$\text{GCA} = \text{Point} \cdot \text{Disc}.$$

GCA is therefore a stricter metric than either pointing accuracy or phrase discrimination alone. A sample receives credit only when the model grounds the target phrase in the correct region and ranks the target phrase above its counterfactuals. We report GCA as a percentage.

Stage-wise continual evaluation. Let $R_{t,k}$ denote the average GCA on evaluation records from stage group k after the model has completed training through Stage t, where $k \leq t$. After Stage t, the reported stage-wise score is the macro-average over all stages observed so far:

$$\mathrm{GCA}_t = \frac{1}{t} \sum_{k=1}^{t} R_{t,k}.$$

This seen-stage macro-average gives equal weight to each acquired stage and prevents later evaluation from being dominated by stages with more records or easier phrase types. The "Avg." column in the main tables averages these stage-wise scores across the continual trajectory:

$$\mathrm{Avg.} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{GCA}_t,$$

where $T = 8$ in our benchmarks.

Forgetting. We report GCA forgetting, denoted $\mathcal{F}_{\mathrm{GCA}}$, over stages introduced before the final stage. For each stage group $k < T$, we compare its best GCA after acquisition with its final GCA after Stage T:

$$\mathcal{F}_{\mathrm{GCA}} = \frac{1}{T-1} \sum_{k=1}^{T-1} \left(\max_{t \in \{k,\ldots,T\}} R_{t,k} - R_{T,k} \right).$$

Lower values indicate less forgetting. Negative values can occur when final performance exceeds all earlier post-acquisition values, corresponding to positive backward transfer.

Auxiliary diagnostics. In addition to GCA, we sometimes report Point and Disc separately to distinguish localization failures from counterfactual ranking failures. We also use evidence-in-box diagnostics for analysis. Given the target evidence distribution $a_p(x, q^+)$ and hidden region B, evidence-in-box is

$$\mathrm{EIB} = \sum_{p : c_p \in B} a_p(x, q^+).$$

A normalized version subtracts the fraction of patches covered by the box, so that diffuse evidence over a large region does not receive the same interpretation as concentrated evidence on the target. These diagnostics are used only for analysis and are not the primary ranking metric.

A.4 Implementation Details

Backbone and preprocessing. All main experiments initialize from the same beta-CLIP ViT-B/16 checkpoint. Images are mapped to the 224×224 CLIP input resolution, producing a 14×14 grid of patch tokens after removing the class token. Both local visual tokens and text embeddings have dimension 512. The shared vision encoder, text encoder, and global projection layers remain frozen throughout continual learning. Consequently, every stage operates in the same shared image and text spaces, while only its local grounding state is optimized.

Stage-local trainable state. At Stage t, CIRCE trains a phrase-conditioned local pooler, a dual text adapter, and a mixture-anchor module. The local pooler produces the inherited patch–phrase affinity $\ell_{t,p}(q)$. The dual text adapter contains rank-128 content and target–counterfactual residual branches and is used only inside the local expert. Stage inference continues to use the unadapted frozen text representation, which keeps the Gaussian stage summaries comparable across experts. The mixture-anchor module contains a phrase-modulated evidence-feature head and a phrase-conditioned mixture head. The evidence-feature head maps each patch–phrase pair to a 32-dimensional anchor feature. The mixture head uses hidden width 128 and predicts four diagonal Gaussian components, including their weights, means, log variances, and contribution weights. Log variances are clamped to $[-6, 4]$. Phrase modulation starts from unit gain and zero bias. Component means receive a small random initialization with standard deviation 0.25 to break responsibility symmetry. After training, the pooler, adapter, and mixture-anchor parameters are frozen as expert E_t.

Table 5: Implementation hyperparameters for the formal CIRCE experiments. Parameters are shared by COCO80-CI8 and PACO-CI8 unless stated otherwise.

Hyperparameter	Value
Architecture	
Backbone	beta-CLIP ViT-B/16
Input resolution	224×224
Patch grid	14×14
Shared embedding dimension	512
Dual text-adapter rank	128
Anchor feature dimension	32
Anchor hidden dimension	128
Mixture components K	4
Mixture covariance	Diagonal
Log-variance range	$[-6, 4]$
Component-mean initialization std.	0.25
Optimization	
Epochs per stage	1
Batch size	64
Optimizer	AdamW
AdamW (β_1, β_2)	$(0.9, 0.999)$
AdamW ϵ	10^{-8}
Weight decay	0.01
Pooler learning rate	10^{-4}
Text-adapter learning rate	3×10^{-4}
Mixture-anchor learning rate	3×10^{-4}
Warmup fraction	5%
Learning-rate schedule	Cosine decay
Automatic mixed precision	bfloat16
Gradient clipping	1.0
Acquisition objective	
Local-pooling temperature	0.07
Evidence exponent γ	1.0
Local-margin weight α	0.5
Local-margin temperature τ	1.0
Stability weight λ	0
Stability inner weights $(\beta, \gamma_{\text{ent}})$	$(1.0, 0.0)$
Feature-perturbation scale	0.05
Learned score-scale cap	100
Anchor numerical epsilon	10^{-8}
Calibration records per expert	512
Stage inference	
Stage-prior feature space	Frozen target-phrase embeddings
Prior fit records per stage	Up to 5,000
Covariance type	Full, shrinkage regularized
Covariance shrinkage λ_s	0.1
Covariance ridge ϵ_s	10^{-6}
Standard CIRCE expert calls/query	1
Reporting	
COCO80-CI8 seeds	17, 23, 42
PACO-CI8 primary seeds	7, 17, 23
Uncertainty	Sample standard deviation

Optimization. We train each stage for one epoch with batch size 64. AdamW uses $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay 0.01. The local pooler uses learning rate 10^{-4}. The dual text adapter and the complete mixture-anchor module both use learning rate 3×10^{-4}. The optimizer is reset at the beginning of every stage. We apply linear warmup during the first 5% of stage updates, followed by cosine learning-rate decay. Gradients are clipped to norm 1.0, and training uses bfloat16

Table 6: Effect of mixture count K under target-only inference. Bold denotes the default configuration.

	COCO80-CI8			PACO-CI8		
K	Avg. ↑	Final ↑	$\mathcal{F}_{\mathrm{GCA}}$ ↓	Avg. ↑	Final ↑	$\mathcal{F}_{\mathrm{GCA}}$ ↓
1	$64.91_{\pm 0.86}$	$61.72_{\pm 0.64}$	$0.18_{\pm 0.05}$	$23.41_{\pm 1.11}$	$22.96_{\pm 0.94}$	$0.05_{\pm 0.03}$
2	$66.21_{\pm 0.91}$	$63.10_{\pm 0.62}$	$0.07_{\pm 0.03}$	$24.22_{\pm 0.68}$	$23.91_{\pm 0.53}$	$0.03_{\pm 0.02}$
4	$\mathbf{67.08}_{\pm 0.99}$	$\mathbf{64.02}_{\pm 0.58}$	$\mathbf{0.01}_{\pm 0.00}$	$\mathbf{24.85}_{\pm 0.59}$	$\mathbf{24.55}_{\pm 0.56}$	$\mathbf{0.00}_{\pm 0.00}$
8	$66.95_{\pm 1.04}$	$63.86_{\pm 0.63}$	$0.02_{\pm 0.01}$	$24.76_{\pm 0.61}$	$24.42_{\pm 0.62}$	$0.01_{\pm 0.01}$

Table 7: Sensitivity to counterfactual count and covariance shrinkage. Bold denotes the default configuration in each block.

	COCO80-CI8			PACO-CI8		
Setting	Avg. ↑	Final ↑	$\mathcal{F}_{\mathrm{GCA}}$ ↓	Avg. ↑	Final ↑	$\mathcal{F}_{\mathrm{GCA}}$ ↓
$m = 1$	$65.47_{\pm 0.92}$	$62.36_{\pm 0.65}$	$0.18_{\pm 0.05}$	$23.81_{\pm 0.67}$	$23.37_{\pm 0.63}$	$0.06_{\pm 0.03}$
$\boldsymbol{m = 2}$	$\mathbf{67.08}_{\pm 0.99}$	$\mathbf{64.02}_{\pm 0.58}$	$\mathbf{0.01}_{\pm 0.00}$	$\mathbf{24.85}_{\pm 0.59}$	$\mathbf{24.55}_{\pm 0.56}$	$\mathbf{0.00}_{\pm 0.00}$
$\lambda_s = 0$	$66.31_{\pm 0.94}$	$63.14_{\pm 0.67}$	$0.11_{\pm 0.04}$	$24.18_{\pm 0.31}$	$23.74_{\pm 0.56}$	$0.04_{\pm 0.02}$
$\lambda_s = 0.05$	$66.86_{\pm 0.98}$	$63.71_{\pm 0.60}$	$0.04_{\pm 0.02}$	$24.61_{\pm 0.64}$	$24.22_{\pm 0.59}$	$0.01_{\pm 0.01}$
$\boldsymbol{\lambda_s = 0.1}$	$\mathbf{67.08}_{\pm 0.99}$	$\mathbf{64.02}_{\pm 0.58}$	$\mathbf{0.01}_{\pm 0.00}$	$\mathbf{24.85}_{\pm 0.59}$	$\mathbf{24.55}_{\pm 0.56}$	$\mathbf{0.00}_{\pm 0.00}$
$\lambda_s = 0.2$	$66.74_{\pm 1.02}$	$63.52_{\pm 0.64}$	$0.06_{\pm 0.02}$	$24.49_{\pm 0.68}$	$24.10_{\pm 0.31}$	$0.02_{\pm 0.01}$

automatic mixed precision. COCO80-CI8 and PACO-CI8 use the same optimization schedule and per-stage training budget.

Acquisition objective. The target phrase occupies index zero in every training query set, and the remaining phrases are controlled counterfactuals. The acquisition objective is

$$\mathcal{L}_{\mathrm{acq}} = \mathcal{L}_{\mathrm{CFA}} + \lambda\, \mathcal{L}_{\mathrm{SMA}},$$

where

$$\mathcal{L}_{\mathrm{CFA}} = \mathcal{L}_{\mathrm{image}} + \alpha\, \mathcal{L}_{\mathrm{local}}.$$

The image-level term makes the target phrase outrank its alternatives. The local term weights the target-to-counterfactual patch margin by the evidence assigned to each patch. We use $\alpha = 0.5$ and margin temperature $\tau = 1.0$. Local pooling and evidence computation use temperature 0.07 and evidence exponent $\gamma = 1.0$. The score scale is learned within each local pooler and capped at 100; no fixed score scale is used. The stable-mixture term contains evidence consistency, component consistency, and component spatial entropy. Its internal weights are $\beta = 1.0$ and $\gamma_{\mathrm{ent}} = 0$, with perturbation scale 0.05. All main experiments set its outer weight to $\lambda = 0$. Thus, the submitted acquisition objective is exactly $\mathcal{L}_{\mathrm{CFA}}$. Nonzero stability weights are retained only as negative controls because they do not improve acquisition or final GCA.

Expert calibration. After acquiring expert E_t, we compute its target-only calibration statistics from at most 512 current-stage records. These statistics contain means and standard deviations of aggregate evidence qualities; they contain no stored phrase, image, patch feature, or per-example state. A feature perturbation with scale 0.05 is used only to measure the stability diagnostic. Because the main objective has $\lambda = 0$, this perturbation contributes no gradient to the final training objective.

Target-phrase stage inference. CIRCE performs task-free expert selection in the shared frozen text space. At the end of Stage t, we encode up to 5,000 current-stage target phrases with the frozen text tower and fit a full-covariance Gaussian. The covariance estimate is

$$\Sigma_t = (1 - \lambda_s)\widehat{\Sigma}_t + \lambda_s \frac{\mathrm{tr}(\widehat{\Sigma}_t)}{d} I + \epsilon_s I,$$

with shrinkage $\lambda_s = 0.1$, ridge $\epsilon_s = 10^{-6}$, and $d = 512$. At inference, the log likelihood of the target phrase is evaluated under every available stage Gaussian, and the corresponding expert with maximum likelihood is selected. The deployed selector retains one mean, precision matrix, log determinant,

Table 8: Optional counterfactual cascade behavior. Trigger is the percentage of queries failing initial verification; rescue is reported among triggered queries.

Dataset	Trigger (%)	Rescue (%)	Experts/query	Avg. GCA ↑	Final GCA ↑
COCO80-CI8	$31.6_{\pm 1.8}$	$74.2_{\pm 2.1}$	$3.21_{\pm 0.14}$	$70.44_{\pm 0.88}$	$68.65_{\pm 0.52}$
PACO-CI8	$38.9_{\pm 2.4}$	$68.5_{\pm 2.8}$	$3.72_{\pm 0.19}$	$26.70_{\pm 0.60}$	$27.14_{\pm 0.57}$

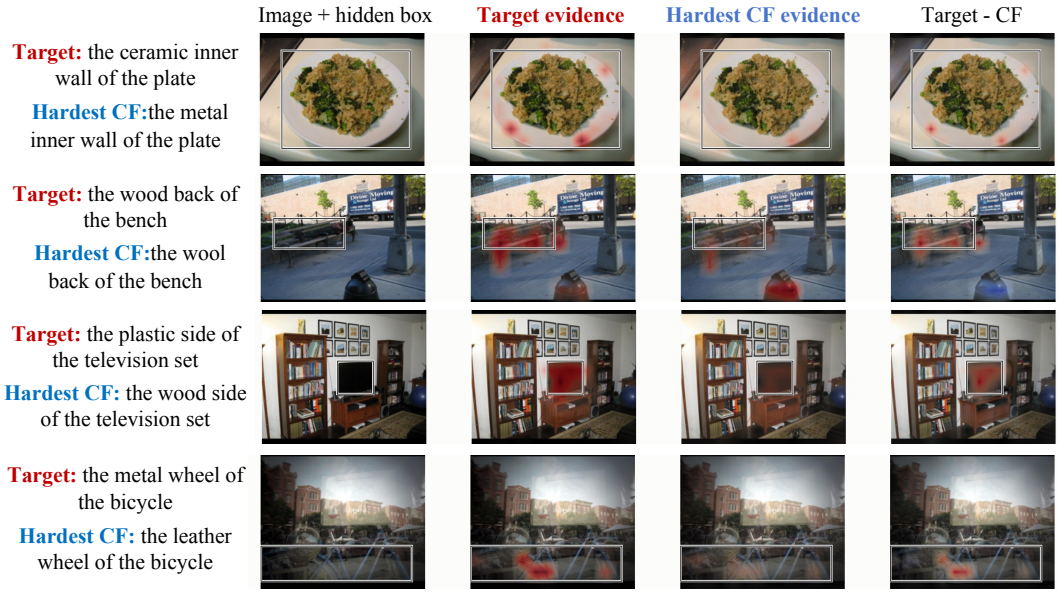


Figure 7: Target and hardest-counterfactual evidence maps with their signed difference (target minus counterfactual). Red and blue in the difference maps indicate target- and counterfactual-favored patches, respectively. Boxes are used only for offline evaluation.

sample count, and shrinkage value per stage. It stores no training phrase or per-example feature. For eight stages, the full float32 selector state occupies approximately 8.02 MiB. The per-query density dump used for offline analysis is not part of the deployed model state.

Retained expert state. Each frozen stage expert stores its local pooler, dual text adapter, mixture-anchor feature head, mixture parameters, and aggregate calibration statistics. The complete eight-stage expert bank occupies approximately 110.12 MiB. The shared 584.73 MiB backbone is stored once and is not duplicated across experts. No training image, text record, old feature, optimizer state, or teacher output is retained.

Inference variants. Standard CIRCE receives one image and one target phrase. It uses the target phrase to select one stage expert and returns the patch with maximum target evidence. It does not receive counterfactual phrases, task identity, class labels, target boxes, or correctness feedback, and it evaluates exactly one expert per query. CIRCE-CF uses a strictly richer inference interface. It receives the controlled counterfactual phrases and first evaluates the expert chosen by the same target-phrase stage prior. If that expert does not rank the target above every counterfactual, the query already has zero discrimination and therefore zero GCA. CIRCE-CF then evaluates the remaining frozen experts and switches to the expert with the largest target-to-counterfactual score margin. CIRCE-CF is reported separately from target-only CIRCE.

A.5 More Ablations

Mixture capacity. Table 6 shows that increasing the number of mixture components improves target-only grounding up to $K = 4$. A single anchor is insufficient to represent the diverse local evidence needed for object, attribute, relation, and part phrases, leading to lower average and final

Algorithm 1 CIRCE

Require: Stage datasets $\{\mathcal{D}_t\}_{t=1}^{T}$; frozen vision and text encoders f_v, f_q
Ensure: Expert bank $\mathcal{E}$ and stage summaries $\mathcal{S}$
$\mathcal{E} \leftarrow \varnothing$; $\mathcal{S} \leftarrow \varnothing$
Training: acquire and preserve local knowledge
for $t = 1, \ldots, T$ **do**
  Initialize the current expert E_t (local pooler, text adapter, mixture head)
  for each minibatch $(x, q^+, \mathcal{Q}^-)$ from $\mathcal{D}_t$ **do**
    Compute phrase-conditioned mixture anchors ▷ Eqs. (2)-(5)
    Obtain patch evidence and phrase scores ▷ Eqs. (6)-(7)
    Update E_t with $\mathcal{L}_{\mathrm{CFA}}$: target ranking and patch discrimination ▷ Eqs. (8)-(9)
  end for
  Freeze E_t and add it to $\mathcal{E}$
  Summarize $f_q(q^+)$ over $\mathcal{D}_t$ by mean μ_t and shrinkage covariance Σ_t ▷ Eqs. (10)-(11)
  Store (μ_t, Σ_t) in $\mathcal{S}$; discard current-stage examples
end for
Inference: select an expert and ground the phrase
function GROUND$(x, q^+, \mathcal{E}, \mathcal{S})$
  Encode the target phrase: $h \leftarrow f_q(q^+)$
  Compute stage compatibility: $g_e(q^+) \leftarrow \log \mathcal{N}(h; \mu_e, \Sigma_e)$ ▷ Eq. (12)
  Select an expert: $\hat{t} \leftarrow \arg\max_e g_e(q^+)$ ▷ Eq. (13)
  Compute patch evidence $a_{\hat{t},p}(x, q^+)$ using $E_{\hat{t}}$ ▷ Eqs. (2)-(6)
  return $\hat{p} = \arg\max_p a_{\hat{t},p}(x, q^+)$ ▷ Eq. (14)
end function

GCA on both benchmarks. Moving from $K = 1$ to $K = 2$ gives a clear gain, while $K = 4$ performs best. Increasing to $K = 8$ does not provide further improvement, suggesting that a small mixture already captures the dominant phrase-conditioned evidence modes and that additional components mainly add redundancy.

Counterfactual and shrinkage sensitivity. Table 7 studies the amount of counterfactual supervision and the covariance shrinkage used by target phrase stage inference. Using only one counterfactual weakens both average and final GCA, indicating that a single controlled alternative does not fully constrain local evidence. Two counterfactuals capture most of the benefit. For stage inference, moderate shrinkage performs best: no shrinkage makes covariance estimates less stable, while excessive shrinkage removes useful stage-specific variation. This supports using a regularized Gaussian summary rather than either a raw covariance or a mean-only stage representation.

Counterfactual cascade behavior. Table 8 summarizes the optional CIRCE-CF cascade. The cascade triggers only when the target-only selected expert fails counterfactual verification, then evaluates additional experts to find one with a stronger target preference. On both datasets, a large fraction of triggered cases are rescued, which explains the GCA improvement of CIRCE-CF over standard CIRCE. The gain is larger on PACO-CI8, where fine-grained part and attribute distinctions make target-only routing harder. This improvement comes with additional inference cost, as the cascade evaluates roughly three to four experts per query, so we report CIRCE-CF separately from the target-only protocol.

A.6 VISUALIZATION

Figure 7 compares the spatial evidence elicited by target phrases and their hardest counterfactuals. Because each pair shares an object and part description while changing a material attribute, substantial overlap between their evidence maps is expected. The relevant distinction is therefore where the edited phrase changes the allocation of local evidence. The signed difference maps expose these changes: positive values identify patches receiving greater evidence mass under the target phrase, whereas negative values identify patches favored by the counterfactual. This comparison helps examine whether phrase sensitivity is concentrated within the annotated region or extends to surrounding context. Since each evidence map is normalized over patches, these differences reflect a redistribution

of evidence rather than an absolute increase in visual support. Moreover, phrase-dependent changes alone do not establish that the model correctly recognizes the material or ranks the target above its counterfactual. The visualization thus complements Point and Disc by revealing the spatial structure underlying their aggregate outcomes.

A.7 ALGORITHM

Algorithm 1 summarizes the training and inference procedures of CIRCE. At each stage, CIRCE learns phrase-conditioned mixture anchors using target phrases and their counterfactuals, with an objective that couples phrase ranking with local patch discrimination. The resulting expert is frozen to preserve its acquired local knowledge, while a Gaussian summary of the stage's target phrases is stored in the shared frozen text space. At inference, CIRCE selects the expert with the highest Gaussian likelihood for the target phrase and uses its patch evidence to produce the grounding prediction. This procedure requires neither historical training examples nor counterfactual phrases or task identities at inference.