E-Logic Prompt: Unified Energy-Logic Framework for Continual Visual Question Answering

Jiayao Tan[1,2,5], **Tianle Liu**[2], **Fuyuan Hu**[2*], **Wei Feng**[3,5], **Liang Wan**[4,5*]

[1]School of Artificial Intelligence, Tianjin University
[2]Suzhou University of Science and Technology
[3]School of Computer Science, Tianjin University
[4]School of Software, Tianjin University
[5]Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage
{jiayaotan,lwan}@tju.edu.cn, {tianleliu@post,fuyuanhu@mail}.usts.edu.cn, wfeng@ieee.org

Abstract

Prompt tuning has shown promise for continual visual question answering (CVQA), facilitating modular and transferable knowledge across tasks. However, existing approaches often overlook the guiding role of prompts in the model's implicit reasoning process. This oversight can lead to inconsistent reasoning paths and performance degradation across tasks. To address this issue, we propose the E-Logic Prompt framework, which employs energy-based models (EBMs) to model the semantic compatibility between prompts and queries. In this framework, prompts function not only as adapters but also as reasoning guides that help maintain coherence throughout the inference process. The framework enforces logical consistency at three levels. At the input level, it selects semantically aligned prompts by minimizing the energy between queries and prompts. Within the model, it aligns intermediate representations with prompts across layers to preserve step-by-step reasoning. Across tasks, it applies energy-based constraints to regulate prompt behavior, effectively suppressing semantic drift and enabling prompt reuse. These three levels of consistency together enhance the guiding capacity of prompts, allowing them to steer the model toward coherent reasoning. Experiments show that E-Logic Prompt outperforms existing methods in both accuracy and knowledge retention, while effectively maintaining balanced cross-modal reasoning throughout continual learning.

Code — https://github.com/tjy1423317192/E-logic-prompt

Introduction

Pretrained models (PTMs) have made significant advancements in the traditional multimodal task of Visual Question Answering (VQA), where the goal is to generate answers to questions based on relevant images. Recently, researchers have extended PTMs to the task of Continual Visual Question Answering (CVQA), where new content emerges over time, necessitating methods to address the challenge of catastrophic forgetting. Prompt tuning, a PTM-based approach, has proven effective in Continual Learning (CL) (Zhang, Zhang, and Xu 2023; Nikandrou et al. 2024; Lin et al. 2022; Zhou et al. 2024; Douillard et al. 2022), where it first selects prompts based on the similarity of input representations and then injects these selected prompts into the encoder to guide the learning process. This approach has shown promising results in CVQA (Wang et al. 2022a,b; Menabue et al. 2024; Lyu et al. 2021; Sun et al. 2022; Lyu et al. 2023; Ramakrishnan, Agrawal, and Lee 2018) For example, (Qian et al. 2023) introduced a fused prompt pool to supplement both visual and textual prompts, (Lei et al. 2023) enhanced generalization by replaying scene graph prompts, and (Cai and Rostami 2024) applied K-means clustering to define modality-aware prompt centers.

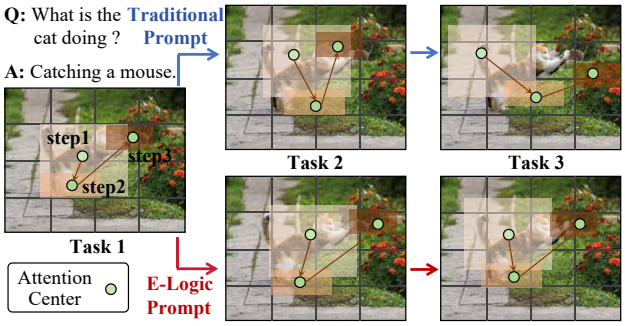


Figure 1: E-Logic Prompt enables consistent reasoning across tasks by aligning semantically related prompts, allowing the model to reuse logical structures and integrate new knowledge without forgetting prior reasoning paths.

Although traditional prompt learning has improved adaptability and efficiency in continual visual question answering, *the implicit reasoning logic between visual and language modalities remains largely underexplored*. In fact, prompts play a crucial role not only in injecting task-specific information, but also in guiding the model along coherent reasoning paths. When prompts fail to capture and reinforce correct reasoning trajectories, especially those shared across tasks, the model becomes susceptible to reasoning drift or modality disalignment, ultimately degrading cross-task consistency and stability. As shown in Fig.1. In contrast, different question types often share similar intermediate rea-

*Corresponding authors: Liang Wan (first), Fuyuan Hu

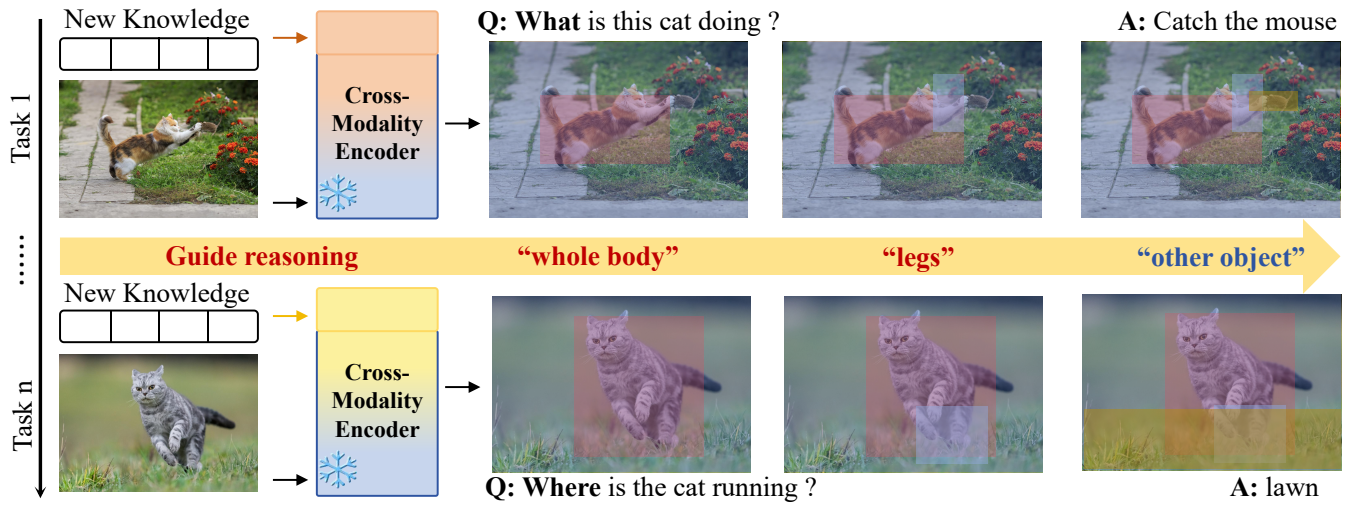


Figure 2: E-Logic Prompt leverages shared reasoning structures across tasks to maintain logical consistency and mitigate forgetting in continual visual question answering.

soning structures. If the prompt mechanism can identify and align such logical commonalities across tasks, it enables the model to continually acquire new tasks while retaining the reasoning capabilities required for previous ones, thus effectively mitigating catastrophic forgetting, As shown in Fig.2.

To address this issue, we explore the use of Energy-Based Models (LeCun et al. 2006) as a means of modeling latent logical consistency. Recent work (Xingsi Dong 2025) suggests that EBMs can softly encode preferences for semantically consistent structures. Motivated by this, we design a simplified controlled probing experiment (details provided in the appendix) to evaluate whether energy functions are sensitive to logical consistency. The results show that, even without prompts or external supervision, a model trained with energy-based loss consistently outperforms its cross-entropy counterpart across all encoder layers, providing initial evidence that EBMs capture latent reasoning logic.

Building on this insight, we propose the E-Logic Prompt framework, which is designed to enhance the ability of prompts to guide stable reasoning in CVQA scenarios and mitigate catastrophic forgetting. The framework consists of three key components. First, we introduce the Query-Prompt Energy Logical Alignment mechanism, which models the interaction between queries and candidate prompts as an energy-based matching process. By assigning lower energy scores to semantically compatible prompts, this mechanism encourages the selection of logically aligned prompts for each query. Second, we extend this objective to a hierarchical structure through Layer-wise Energy Logical Consistency, which aligns the semantics of selected prompts with intermediate encoder representations across multiple reasoning layers, ensuring coherent step-by-step inference. Finally, we propose Cross-Task Prompt Logical Consistency, which regularizes prompt behavior over time by modeling prompt overlap and dynamically adjusting energy constraints based on task similarity. This component promotes more stable prompt reuse and better generalization across tasks in continual learning. Our contributions are three-fold:

- We propose E-Logic Prompt, a unified energy-based prompt framework that enforces implicit logical consistency across tasks and supports coherent reasoning in CVQA scenarios.
- To maintain logical consistency during reasoning, we apply energy-based constraints at queries, layers and tasks. levels, covering Query-Prompt alignment, Layer-wise consistency, and cross-task consistency.
- Extensive experiments show that E-Logic Prompt consistently improves reasoning stability, generalization, and resistance to forgetting across diverse CVQA scenarios.

Related work

Continual Visual Question Answering

Visual Question Answering (VQA) combines visual and textual inputs to answer image-based questions, while Continual Learning (Mai et al. 2017; Ni et al. 2025; Anderson et al. 2018) aims to retain prior knowledge when learning new tasks. Continual VQA (CVQA) unifies these goals by incrementally training models on new image-question-answer pairs. To mitigate forgetting, methods such as regularization, architectural changes, and replay have been explored. More recently, prompt-based approaches offer a lightweight alternative by freezing the backbone and updating only prompts. Notable examples include L2P, which retrieves prompts based on input-query similarity; Dual Prompt, which injects both general-purpose and task-specific expert prompts at different layers to balance stability and plasticity; and CODA, which constructs prompts from weighted combinations of learned components to enhance flexibility. Despite their effectiveness in general continual learning settings, these methods often overlook the need for logical consistency during prompt selection. This can result in misaligned modality interactions and degraded cross-modal reasoning, ultimately limiting performance in CVQA scenarios.

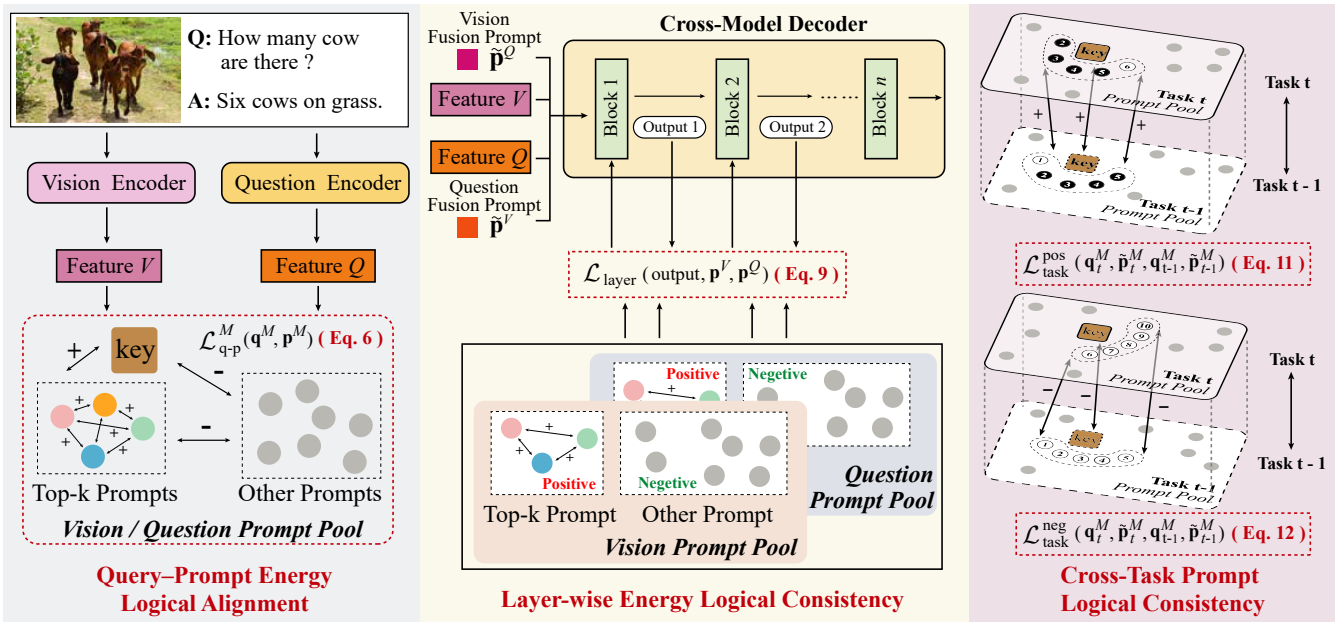


Figure 3: Overview of the E-Logic Prompt framework. (a) Query-Prompt Energy Alignment guides prompt selection via energy-based matching, encouraging semantic compatibility. (b) Layer-wise Energy Consistency aligns encoder representations with prompt pairs to maintain coherent, stepwise reasoning. (c) Cross-Task Prompt Consistency stabilizes prompt behavior across tasks by modeling overlap and semantic drift. Together, these mechanisms enable prompts to capture the model's implicit reasoning logic, enhancing consistency and knowledge retention in CVQA.

Energy-Based Models and Logical Reasoning

Energy-Based Models have recently gained traction for their ability to model implicit constraints over input-output pairs through learned energy functions. Unlike discriminative models that rely on explicit supervision, EBMs assign lower energy to desirable configurations, making them particularly suitable for tasks involving reasoning and alignment. Recent works in neural-symbolic energy modeling (Dickens, Pryor, and Getoor 2024) demonstrate that energy functions can be used to encode logic inspired preferences or constraints, offering an interpretable mechanism for latent logical structure discovery. These models combine deep representations with structured logical forms, where energy minimization leads to consistent outputs under logical rules. Inspired by this perspective, we interpret prompt selection in continual vision-language tasks as a form of latent logical alignment: semantically consistent prompts should have lower energy with respect to a query. Unlike contrastive learning, which relies on margin-based hard negatives, our energy framework provides a soft alignment mechanism that supports both intra-modal and inter-task logic supervision. This allows us to enforce semantic logical consistency during prompt selection without relying on rigid contrastive constraints.

Proposed Method

Problem Formulation

Following (Zhang, Zhang, and Xu 2023), we define CVQA as a sequence of T incremental tasks with training data $\mathcal{D}_1, \cdots, \mathcal{D}_T$, where each $\mathcal{D}t = (x_i^{\mathrm{Q}}, x_i^{\mathrm{V}}, y_i)i = 1^{n_t}$ includes image–question–answer triplets. Models must learn new visual concepts without accessing prior data and maintain performance on previous tasks. We consider three settings: Question Increment (QI) introduces new question types; Class Increment (CI) adds new object classes; and Dual Increment (DI) combines both, where each task contains S subtasks sharing object classes but varying question types.

Existing prompt-based methods (Khattak et al. 2023; Zhang, Zhang, and Xu 2023; Cai and Rostami 2024; Wang et al. 2022a,b; Menabue et al. 2024) adopt prompt pools with key-based retrieval and injection into encoder layers. However, they overlook the underlying reasoning logic between vision and language, leading to illogical or isolated prompts that hinder cross-modal integration and generalization in continual settings.

To address these challenges, we propose E-Logic Prompt, an energy-based prompt-guided framework consisting of three core components: (1) Query-Prompt Energy Logical Alignment, which grounds each cross-modal query in semantically relevant prompts via a Boltzmann style matching loss; (2) Layer-wise Energy Logical Consistency, which extends energy supervision to every encoder block to preserve a coherent reasoning chain across layers; (3) Cross-Task Prompt Energy Logical Consistency, which dynamically regularizes prompt behavior across tasks through overlap driven energy consistency. As shown in Fig. 3.

Query-Prompt Energy Logical Alignment

We aim to guide prompt selection through a logic aware alignment mechanism that softly prioritizes semantically

consistent Query-Prompt pairs. Inspired by recent developments in neural symbolic energy-based modeling (Dickens, Pryor, and Getoor 2024), we treat prompts as latent reasoning components—akin to logical operators that assist in aligning cross-modal inputs with task specific semantics. To model this alignment probabilistically, we adopt an energy-based formulation, where the interaction between a query and a prompt is quantified via a differentiable energy score. Lower energy indicates higher compatibility, thus serving as a soft logical constraint.

Unlike traditional contrastive learning approaches such as InfoNCE (van den Oord, Li, and Vinyals 2018), which rely on hard positive/negative labels and margin-based separation, our method models alignment via Boltzmann distributed likelihoods. This allows the model to learn from soft supervision, offering greater stability in continual learning settings where hard negatives are not always available or meaningful. Formally, for each modality $M \in \{\mathrm{Q}, \mathrm{V}\}$, we maintain a prompt pool $\mathcal{P}^M$ with corresponding keys $\mathcal{K}^M$, where $\mathbf{p}^M, \mathbf{k}^M \in \mathbb{R}^d$. Let $\mathbf{F}^{\mathrm{Q}} \in \mathbb{R}^d$ and $\mathbf{F}^{\mathrm{V}} \in \mathbb{R}^d$ be the encoded question and vision features. The cross-modal queries are computed as:

$$\mathbf{q}^{\mathrm{Q}} = \Phi\big(\mathrm{A}(\mathbf{F}^{\mathrm{Q}}, \mathbf{F}^{\mathrm{V}}) + \mathbf{w}^{\mathrm{Q}} \odot \mathbf{F}^{\mathrm{Q}}\big), \tag{1}$$

$$\mathbf{q}^{\mathrm{V}} = \Phi\big(\mathrm{A}(\mathbf{F}^{\mathrm{V}}, \mathbf{F}^{\mathrm{Q}}) + \mathbf{w}^{\mathrm{V}} \odot \mathbf{F}^{\mathrm{V}}\big), \tag{2}$$

where $A(\cdot)$ denotes a cross attention module and Φ is a pooling operation. Based on each query $\mathbf{q}^M$, we retrieve the indices of the top-k most similar prompt keys:

$$\mathcal{I}^M \leftarrow \text{Top-}k\left(\text{sort}\left(\cos(\mathbf{q}^M, \mathbf{k}^M)\right)\right). \tag{3}$$

In our setting, both queries and prompts reside in a shared latent space. We treat cosine similarity as a soft indicator of logical consistency: semantically aligned pairs tend to occupy similar directions. Thus, we define the energy as the negative cosine similarity between query(q) and prompt(p):

$$E_{\text{q-p}} = -\cos(\mathbf{q}^M, \mathbf{p}_i^M), \tag{4}$$

and normalize this into a Boltzmann distribution over all prompts:

$$Z_{\text{q-p}}^M = \sum_{\mathbf{p}_j \in \mathcal{P}} \exp\left(-E_{\text{q-p}}(\mathbf{q}^M, \mathbf{p}_j^M)\right). \tag{5}$$

Given a retrieved prompt $\mathbf{p}^+ \in \mathcal{I}^M$, the training objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{q-p}}^M = \sum_{\mathbf{p}^+ \in \mathcal{I}^M} \left[-\log \frac{\exp\left(-E(\mathbf{q}^M, \mathbf{p}^+)\right)}{Z_{\text{q-p}}^M}\right]. \tag{6}$$

Intuitively, this loss encourages the model to assign higher probability (i.e., lower energy) to semantically compatible prompts while softly penalizing less relevant ones, without requiring explicit negative labels. Compared to contrastive methods, it avoids rigid margin constraints and enables smooth gradient propagation.

Layer-wise Energy Logical Consistency

Recent studies (Abnar and Zuidema 2020; Vig 2019) suggest that attention patterns evolving across layers can reflect a model's latent reasoning trajectory. Motivated by this, we treat multi layer attention as a soft proxy of internal reasoning paths. While Query-Prompt alignment ensures semantic compatibility at the input level, cross-modal reasoning is inherently a multi step process distributed across encoder layers. To preserve logical coherence throughout this hierarchy, we extend our energy-based supervision to a layer-wise setting, encouraging intermediate representations to remain aligned with selected prompts during the reasoning process.

Formally, let the cross-modal encoder consist of $L = N$ hierarchical blocks, each producing an intermediate output $\mathbf{y}^{(l)} \in \mathbb{R}^d$ for $l = 1, \ldots, L$. For each layer l, we construct cross-modal prompt pairs $(\mathbf{p}^V, \mathbf{p}^Q)$ from the top-k selected prompts $\mathcal{I}^V \times \mathcal{I}^Q$ based on visual and question queries, respectively.

We define the energy between a layer representation $\mathbf{y}^{(l)}$ and a fused prompt pair via:

$$E^{(l)}(\mathbf{y}^{(l)}, \mathbf{p}^V, \mathbf{p}^Q) = -\cos\left(\mathbf{y}^{(l)}, f(\mathbf{p}^V, \mathbf{p}^Q)\right), \tag{7}$$

where $f(\cdot)$ is a MLP that projects the prompt pair into the shared representation space. We then compute a softmax-normalized probability over all prompt pairs $(\mathbf{p}_i^V, \mathbf{p}_j^Q)$:

$$P^{(l)}(\mathbf{p}^V, \mathbf{p}^Q) = \frac{\exp\big(-E^{(l)}(\mathbf{y}^{(l)}, \mathbf{p}^V, \mathbf{p}^Q)\big)}{\sum \exp\big(-E^{(l)}(\mathbf{y}^{(l)}, \mathbf{p}_i^V, \mathbf{p}_j^Q)\big)}. \tag{8}$$

Given a positive prompt pair $(\mathbf{p}^V, \mathbf{p}^Q) \in \mathcal{I}^V \times \mathcal{I}^Q$, we define the layer-wise energy consistency loss as the negative log-likelihood:

$$\mathcal{L}_{\text{layer}} = \sum_{l=1}^{L} \sum_{(\mathbf{p}^V, \mathbf{p}^Q) \in \mathcal{I}^V \times \mathcal{I}^Q} \Big[-\log P^{(l)}(\mathbf{p}^V, \mathbf{p}^Q)\Big]. \tag{9}$$

This hierarchical supervision encourages intermediate encoder states to align with logically coherent prompt pairs across reasoning layers, thereby reinforcing multi-step semantic consistency. Unlike contrastive learning, which typically supervises only input-level representations, our formulation enables prompt-driven alignment to persist throughout the model's depth.

Cross-Task Prompt Logical Consistency

In continual multi-task settings, prompts often suffer from semantic drift and role ambiguity as tasks evolve, weakening their effectiveness as reusable logical anchors. To address this issue, we introduce a Cross-Task Prompt Logical Consistency mechanism that regularizes prompt semantics over time. This component ensures that prompts with shared logical meaning across tasks retain similar energy characteristics, while unrelated prompts are energy-separated, thereby improving prompt stability and generalization.

Method	Type	VQA v2							NExT-QA						
		DI			CI		QI		DI			CI		QI	
		$A(\uparrow)$	$F_{\text{inter}}(\downarrow)$	$F_{\text{intra}}(\downarrow)$	$A(\uparrow)$	$F_{\text{inter}}(\downarrow)$	$A(\uparrow)$	$F_{\text{inter}}(\downarrow)$	$A(\uparrow)$	$F_{\text{inter}}(\downarrow)$	$F_{\text{intra}}(\downarrow)$	$A(\uparrow)$	$F_{\text{inter}}(\downarrow)$	$A(\uparrow)$	$F_{\text{inter}}(\downarrow)$
Dual Prompt	CL	33.601	2.660	10.574	36.063	4.449	14.146	23.518	16.163	9.591	9.974	10.693	1.663	12.393	11.803
L2P	CL	31.186	2.112	12.541	33.706	4.323	13.720	23.807	14.181	9.441	9.263	9.903	1.635	10.904	10.081
L2P*	CL	31.343	2.201	12.723	33.428	4.150	13.853	23.651	14.452	9.316	9.473	10.207	1.641	11.204	10.152
CODA	CL	35.138	0.902	10.735	37.102	4.527	15.438	22.561	15.115	9.282	9.361	11.479	1.556	22.416	6.960
Triplet	CVQA	32.826	1.134	12.244	37.102	4.527	14.382	22.697	18.781	8.674	9.633	11.362	1.511	18.719	9.587
Maple	CL	35.187	1.054	11.148	37.450	4.606	15.371	22.164	17.877	8.538	9.394	11.596	1.641	16.501	10.588
VQACL	CVQA	34.224	0.867	10.626	36.950	4.431	15.525	22.452	20.472	8.772	9.041	10.787	1.818	14.870	13.500
CluMo	CVQA	35.079	1.580	11.300	36.462	4.039	13.208	25.016	18.451	9.183	8.904	11.519	1.559	17.971	11.005
Star-prompt	CL	34.437	1.784	10.584	36.516	4.172	15.104	22.837	16.695	10.347	9.218	11.030	1.742	15.571	11.750
MISA	CL	34.875	1.613	10.184	36.923	3.954	15.641	22.032	17.854	9.323	9.671	11.683	1.623	15.888	11.231
TPPT	CL	34.908	1.641	10.223	36.889	3.978	15.607	21.994	17.813	9.298	9.697	11.670	1.651	19.903	9.198
E-Logic Prompt	**CVQA**	**36.963**	**0.672**	**10.366**	**39.943**	**3.793**	**18.021**	**20.976**	**21.922**	**8.991**	**8.783**	**13.731**	**1.193**	**24.714**	**6.911**

Table 1: Performance comparison under DI, CI, and QI settings on VQA v2 and NExT-QA. L2P* means L2P with a single prompt pool. The best performance is highlighted in bold, and the second-best is underlined.

Our motivation stems from the observation that semantically consistent prompts should yield similar energy interactions with task queries, even across task boundaries. Instead of hard prompt re-use or task-specific prompt reinitialization, we adopt an energy-based objective that softly aligns prompt-query pairs over time. Formally, for each modality $M \in \{\mathrm{Q}, \mathrm{V}\}$ at tasks t and $t-1$, we aggregate the selected prompts into a fused vector:

$$\tilde{\mathbf{p}}_t^M = \frac{1}{K} \sum_{i \in \mathcal{I}_t^M} \mathbf{p}_i^M, \quad \tilde{\mathbf{p}}_{t-1}^M = \frac{1}{K} \sum_{i \in \mathcal{I}_{t-1}^M} \mathbf{p}_i^M. \tag{10}$$

We then define two energy-based losses. The first, a positive alignment loss, minimizes the squared energy distance between queries and their corresponding fused prompts when a significant prompt overlap exists ($|\mathcal{I}_t^M \cap \mathcal{I}_{t-1}^M| \geq 3$):

$$\mathcal{L}_{\text{task}}^{\text{pos}} = \left(E(\mathbf{q}_t^M, \tilde{\mathbf{p}}_t^M) - E(\mathbf{q}_{t-1}^M, \tilde{\mathbf{p}}_{t-1}^M)\right)^2, \tag{11}$$

To discourage semantic collapse for unrelated prompts ($|\mathcal{I}_t^M \cap \mathcal{I}_{t-1}^M| < 1$), we apply a negative margin loss:

$$\mathcal{L}_{\text{task}}^{\text{neg}} = \max\left(0,\ m + E(\mathbf{q}_t^M, \tilde{\mathbf{p}}_{t-1}^M) - E(\mathbf{q}_t^M, \tilde{\mathbf{p}}_t^M)\right), \tag{12}$$

where $m > 0$ is a margin. This formulation enforces that unrelated prompt–query pairs are less energetically compatible than current-task aligned prompts. Unlike triplet-based contrastive losses that impose hard boundaries on representations, our energy-based cross-task loss models logical alignment probabilistically, leveraging soft energy scores to preserve semantic structure over time. This facilitates smoother task transitions and more consistent prompt reuse in continual learning scenarios.

Theoretical Insight: Energy-Based Modeling as Soft Logical Alignment

Let $\mathbf{q}$ and $\mathbf{p}$ denote query and prompt embeddings. Their compatibility is defined via an energy function $E(\mathbf{q}, \mathbf{p})$, where lower energy implies higher semantic alignment. We model the prompt distribution (simplified Eq.6):

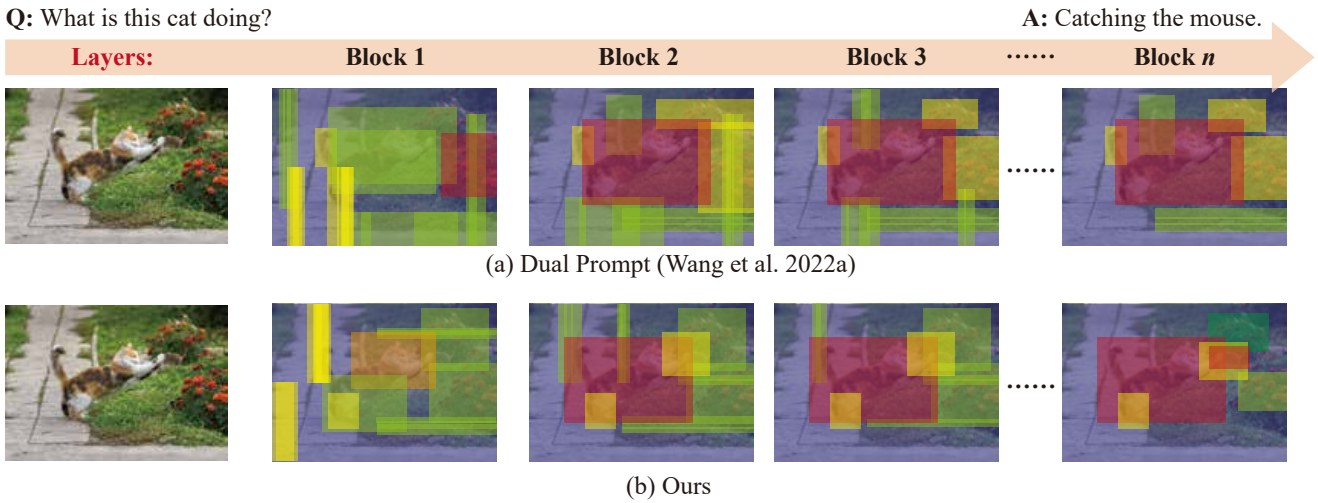
$$P(\mathbf{p} \mid \mathbf{q}) = \frac{\exp(-E(\mathbf{q}, \mathbf{p}))}{Z(\mathbf{q})}, \tag{13}$$

where, $Z(\mathbf{q}) = \sum_{\mathbf{p}' \in \mathcal{P}} \exp(-E(\mathbf{q}, \mathbf{p}'))$. Unlike contrastive methods relying on hard negatives, this formulation enables smooth, differentiable supervision better suited for continual learning. The energy function implicitly serves as a logical selector, assigning lower energy to semantically consistent prompt–query pairs. We extend this principle across model layers and tasks: layer-wise energy promotes multi-step reasoning consistency, while task-level regularization mitigates semantic drift. This unified framework enables stable and interpretable prompt selection, enhancing generalization in CVQA.

Experiment

Set Up

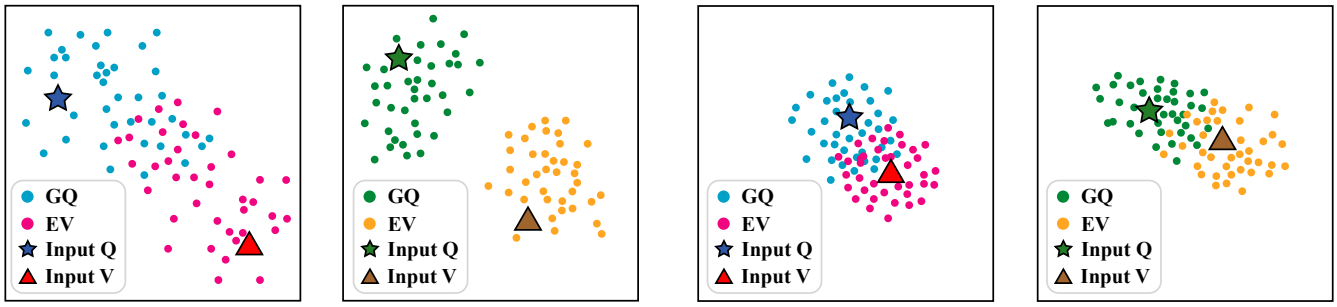
Dataset. Following (Zhang, Zhang, and Xu 2023), we conduct experiments on two widely-used datasets: VQA v2 (Goyal et al. 2017) and NExT-QA (Xiao et al. 2021). VQA v2 consists of 1.1 million image-question pairs, while NExT-QA contains 52K video-based question-answer pairs. For VQA v2, the continual learning experiments are organized into 8 sequential tasks for Domain Incremental Learning (DI) and Class Incremental Learning (CI), and 10 tasks for Question Incremental Learning (QI). In QI, each task introduces a new question type. For CI, each task involves 10 unique object classes, and under DI, each task consists of 5 subtasks, with each subtask focusing on two different question types about the same 10 object classes. For NExT-QA, the experiments are structured into 7 sequential tasks. In QI, each task introduces a new question type. For CI, 4 tasks involve 11 object classes each, and 3 tasks involve 12 object classes each. In DI, each task contains unique question types and 5 subtasks, each addressing the same 16 object classes.
Evaluation Metrics. We evaluate model's performance



(a) Dual Prompt (Wang et al. 2022a)

(b) Ours

Figure 4: Comparison of attention evolution across layers. E-Logic Prompt maintains coherent and focused attention relevant to the question, while Dual Prompt produces fragmented attention, reflecting weaker reasoning consistency.



(a) Dual Prompt (Wang et al. 2022a) (A : 33.601, F_{inter}: 2.660)

(b) E-Logic Prompt (A : 36.223, F_{inter}: 0.447)

Figure 5: Visualization of injected prompts and input feature using t-SNE (Van der Maaten and Hinton 2008).

based on two matrices: Average Performance (A), calculated as $A = \frac{1}{T}\sum_{t=1}^{T} \text{Acc}_{T,t}$, where $\text{Acc}_{T,t}$ means accuracy on task t after training on task T, measures overall accuracy after the training; Inter Task Forgetting (F_{inter}), computed as $F_{\text{inter}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\max_{j\in\{1,\ldots,t\}} \text{Acc}_{j,t} - \text{Acc}_{T,t}\right)$ quantifies the average gaps between the best performance and the current performance across tasks. In addition, for the DI setting, we introduce Inner Task Forgetting (F_{intra}), which is logically similar to the forgetting, and it specifically measures performance degradation across subtasks within a task.

Experimental Main Results

In this section, we compare **E-Logic Prompt** with 9 state-of-the-art methods, including 6 general continual learning (CL) approaches (Wang et al. 2022a; Smith et al. 2023; Wang et al. 2022b; Khattak et al. 2023; Menabue et al. 2024) and 3 CVQA-specific models (Qian et al. 2023; Zhang, Zhang, and Xu 2023; Cai and Rostami 2024). All models use both visual and textual prompts (except L2P*) and a comparable prompt budget. As shown in Table 1, E-Logic Prompt consistently outperforms all baselines across all settings (DI, CI, QI) and datasets (VQA v2, NExT-QA). For example, it achieves the highest accuracy (36.963%) and lowest forgetting (0.672) on VQA v2-DI, outperforming MaPLe (35.187%, 1.054). On NExT-QA-DI, it reaches 21.922% accuracy with $F_{\text{inter}} = 8.991$, showing strong resilience to domain shift. In CI and QI settings, E-Logic Prompt also achieves the best scores, e.g., 39.943% (VQA v2-CI), 13.731% (NExT-QA-CI), and 24.714% (NExT-QA-QI), all with lower forgetting than competing methods. Overall, these results confirm the superior effectiveness, generalization, and robustness of E-Logic Prompt in continual vision-language tasks.

Energy-based Logical Validation

t-SNE-Based Prompt Space Validation. To qualitatively verify the effectiveness of our Query-Prompt Energy Logical Alignment, we conduct a visualization analysis on the learned prompt space using t-SNE. Specifically, we project all visual and textual prompt embeddings into a two-dimensional space after training, and visualize their distribu-

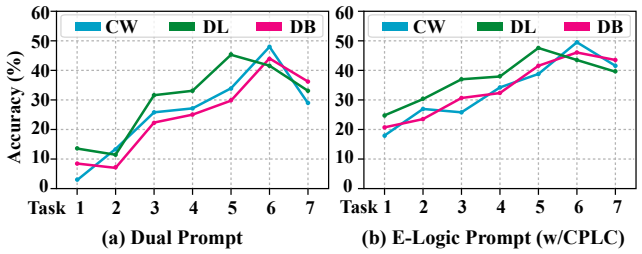


Figure 6: Different task orders in NExt QA under DI setting.

tion. As shown in Fig.5, the prompts are no longer randomly scattered but form distinct semantic clusters, indicating that the model has successfully captured the underlying logical structure among prompts. Notably, prompts across modalities (visual and textual) tend to co-locate when they share similar semantic roles, demonstrating that the energy-based alignment encourages cross-modal consistency.

Analysis of Layer-wise Logical Focus. To investigate the internal effect of Layer-wise Energy Logical Consistency, we visualize the attention distributions over object bounding boxes at each encoder block. As shown in Figure 4, the attention maps reveal that with E-Logic Prompt, the model progressively focuses on semantically relevant image regions as the reasoning depth increases. In contrast, existing methods such as Dual Prompt exhibit scattered and inconsistent attention patterns across layers. By introducing energy-based hierarchical constraints, our method leads to more stable layer-wise behavior and significantly improves reasoning accuracy throughout the network. This indicates a coherent logical progression across layers, further validating that our layer-wise energy mechanism not only aligns intermediate representations with the selected prompts but also enhances semantic consistency and multi-step reasoning quality across the model depth.

Impact of Task Order. To validate the effectiveness of Cross-Task Prompt Logical Consistency (CPLC), we integrate it into the Dual Prompt (Wang et al. 2022a) and conduct an order sensitivity analysis under the DI protocol on NExT-QA. Specifically, we evaluate the model's performance under different task orders (e.g., starting from CW, DL, or DB). As shown in Fig. 6, the original Dual Prompt exhibits noticeable performance fluctuations across task orders, indicating prompt instability and the accumulation of task-specific biases. In contrast, our enhanced version with CPLC yields much smoother and more stable accuracy curves. This order-invariant behavior demonstrates that our method effectively preserves the semantic coherence of prompts across tasks and mitigates cascading effects caused by task order variation during continual learning.

Computational Efficiency. As shown in Fig. 7, E-Logic Prompt achieves competitive inference efficiency. Processing 100 samples takes 0.179 seconds, which is faster than MAPLE (0.203s) and TRIPLET (0.218s), and comparable to CODA (0.176s). While slightly slower than Dual Prompt (0.151s), E-Logic Prompt introduces only minimal overhead despite its multi-level energy-based consistency design. This modest cost is justified by the notable gains in performance,

$\mathcal{L}_{\text{q-p}}$	$\mathcal{L}_{\text{layer}}$	$\mathcal{L}_{\text{task}}$	A (↑)	F_{inter} (↓)
			33.172	2.503
✓			34.420	1.972
	✓		33.885	1.638
		✓	34.062	1.407
	✓	✓	34.510	1.263
✓	✓		35.142	1.098
✓		✓	35.306	0.982
✓	✓	✓	**36.223**	**0.447**

Table 2: Ablation study on VQA v2 under DI setting showing the effects of $\mathcal{L}_{\text{q-p}}$, $\mathcal{L}_{\text{layer}}$, $\mathcal{L}_{\text{task}}$.

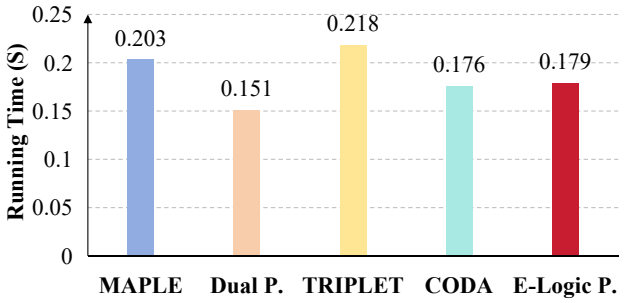


Figure 7: Time spent processing 100 samples.

stability, and interpretability, confirming that enforcing logical consistency across prompts can be achieved efficiently for practical continual learning applications.

Ablation Study

To evaluate the contribution of each proposed loss component, we conduct an ablation study on VQA v2 under the DI setting by incrementally adding the three objectives: $\mathcal{L}_{\text{q-p}}$, $\mathcal{L}_{\text{layer}}$, and $\mathcal{L}_{\text{task}}$. As shown in Table 2, all components consistently improve accuracy (A) and reduce forgetting (F_{inter}). Individually, $\mathcal{L}_{\text{q-p}}$ brings the largest single gain, improving accuracy from 33.17% to 34.42% (+1.25) and reducing forgetting by 0.53. Adding $\mathcal{L}_{\text{layer}}$ or $\mathcal{L}_{\text{task}}$ separately also yields noticeable improvements, demonstrating their respective benefits for hierarchical alignment and cross-task stability. Combining $\mathcal{L}_{\text{q-p}}$ with either $\mathcal{L}_{\text{layer}}$ or $\mathcal{L}_{\text{task}}$ further increases accuracy beyond 35%.

Conclusion

We propose an energy-based prompt framework that enhances logical consistency in continual cross-modal learning. Our approach introduces three components: Query-Prompt alignment, layer-wise consistency, and cross-task prompt regularization. Together, they enable more stable, interpretable, and semantically aligned prompt selection across tasks and layers. Experiments on NExT-QA demonstrate improved performance and robustness under varying task orders. **Limitation**: Our method assumes prompt overlap reflects task similarity, which may be unreliable in highly entangled settings. Future work may explore adaptive prompt disentanglement and dynamic query refinement to better capture evolving task semantics.

Acknowledgments

Our work was supported by the following institutions and projects: National Natural Science Foundation of China (Grant No. 62572349, No. 62476196, 62476189); Suzhou Science and Technology Project (No. SYG2024149); Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center.

References

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *ACL*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6077–6086.

Cai, Y.; and Rostami, M. 2024. CluMo: Cluster-Based Modality Fusion Prompt for Continual Learning in Visual Question Answering. *arXiv preprint arXiv:2408.11742*.

Dickens, C.; Pryor, C.; and Getoor, L. 2024. Modeling patterns for neural-symbolic reasoning using energy-based models. In *AAAI*, volume 3, 90–99.

Douillard, A.; Ramé, A.; Couairon, G.; and Cord, M. 2022. DyTox: Transformers for Continual Learning With Dynamic Token Expansion. In *CVPR*, 9285–9295.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, 6904–6913.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. MAPLE: Multi-Modal Prompt Learning. In *CVPR*, 19113–19122.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; Huang, F.; et al. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Lei, S. W.; Gao, D.; Wu, J. Z.; Wang, Y.; Liu, W.; Zhang, M.; and Shou, M. Z. 2023. Symbolic Replay: Scene Graph as Prompt for Continual Learning on VQA Task. In *AAAI*, 139–149.

Lin, Y.; Xie, Y.; Chen, D.; Xu, Y.; Zhu, C.; and Yuan, L. 2022. REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering. In *NeurIPS*, 10560–10571.

Lyu, F.; Sun, Q.; Shang, F.; Wan, L.; and Feng, W. 2023. Measuring asymmetric gradient discrepancy in parallel continual learning. In *ICCV*, 11411–11420.

Lyu, F.; Wang, S.; Feng, W.; Ye, Z.; Hu, F.; and Wang, S. 2021. Multi-Domain Multi-Task Rehearsal for Lifelong Learning. In *AAAI*, 8819–8827.

Mai, Z.; Li, R.; Jeong, J.; Quispe, D.; Kim, H.; and Sanner, S. 2017. Online continual learning in image classification: An empirical survey. In *Neurocomputing*, 1–6.

Menabue, M.; Frascaroli, E.; Boschini, M.; Sangineto, E.; Bonicelli, L.; Porrello, A.; and Calderara, S. 2024. Semantic Residual Prompts for Continual Learning. In *ECCV*, 1–18.

Ni, C.; Lyu, F.; Tan, J.; Hu, F.; Yao, R.; and Zhou, T. 2025. Maintaining consistent inter-class topology in continual test-time adaptation. In *CVPR*, 15319–15328.

Nikandrou, M.; Pantazopoulos, G.; Konstas, I.; and Suglia, A. 2024. Enhancing Continual Learning in Visual Question Answering With Modality-Aware Feature Distillation. *arXiv preprint arXiv:2406.19297*.

Qian, Z.; Wang, X.; Duan, X.; Qin, P.; Li, Y.; and Zhu, W. 2023. Decouple Before Interact: Multi-Modal Prompt Learning for Continual Visual Question Answering. In *ICCV*, 2953–2962.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming Language Priors in Visual Question Answering With Adversarial Regularization. In *NeurIPS*, 152–164.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. CODA-Prompt: Continual Decomposed Attention-Based Prompting for Rehearsal-Free Continual Learning. In *CVPR*, 11909–11919.

Sun, Q.; Lyu, F.; Shang, F.; Feng, W.; and Wan, L. 2022. Exploring Example Influence in Continual Learning.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. In *arXiv preprint arXiv:1807.03748*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 2579–2605.

Vig, J. 2019. A multiscale visualization of attention in the transformer model. *ACL*.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022a. DualPrompt: Complementary Prompting for Rehearsal-Free Continual Learning. In *ECCV*, 631–648. Springer.

Wang, Z.; Zhang, Z.; Lee, C.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to Prompt for Continual Learning. In *CVPR*, 139–149.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T. 2021. NExT-QA: Next Phase of Question Answering to Explaining Temporal Actions. In *CVPR*, 9777–9786.

Xingsi Dong, S. W., Xiangyuan Peng. 2025. Predictive Learning in Energy-based Models with Attractor Structures. *arXiv preprint arXiv:2501.13997*.

Zhang, X.; Zhang, F.; and Xu, C. 2023. VQACL: A Novel Visual Question Answering Continual Learning Setting. In *CVPR*, 19102–19112.

Zhou, D.; Cai, Z.; Ye, H.; Zhan, D.; and Liu, Z. 2024. Revisiting Class-Incremental Learning With Pre-Trained Models: Generalizability and Adaptivity Are All You Need. *International Journal of Computer Vision*, 1–21.