

Affordance-First Decomposition for Continual Learning in Video–Language Understanding

Mengzhu Xu^{1†}, Hanzhi Liu^{2†}, Ningkan Peng³, Qianyu Chen⁴, Canran Xiao^{5*}

¹University of Sydney ²University of California, Santa Barbara

³Nanjing Normal University ⁴Nanyang Technological University

⁵Shenzhen Campus of Sun Yat-sen University

* xiaocr3@mail.sysu.edu.cn

Abstract

Continual learning for video–language understanding is increasingly important as models face non-stationary data, domains, and query styles, yet prevailing solutions blur what should stay stable versus what should adapt, rely on static routing/capacity, or require replaying past videos. We aim to explicitly specify where stability lives and where plasticity should be focused under realistic memory and privacy constraints. We introduce Affordance-First Decomposition (AFD): videos are mapped to slowly varying affordance tokens that form a shared, time-aligned substrate, while a lightweight, query-routed, conflict-aware scheduler concentrates adaptation and grows capacity only when needed. The substrate is stabilized via weak alignment and teacher consistency, and training uses question-only replay. AFD achieves state-of-the-art across protocols: 51.6% average accuracy with -1.8% forgetting on domain-incremental VideoQA, ViLCo $R@1@0.5$ of 29.6% (MQ) and 20.7% (NLQ) with 18.4% $stAP@0.25$ (VQ), and 39.5% accuracy with -1.6% forgetting on time-incremental iVQA. Overall, AFD offers an explicit, interpretable split between a stable interaction-centered substrate and targeted adaptation.

1. Introduction

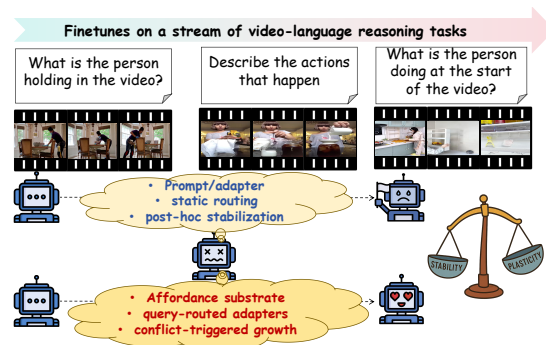


Figure 1. Under a stream of video–language reasoning tasks, existing methods rely on prompt/adaptor add-ons with static routing and post-hoc stabilization, leaving the stability–plasticity trade-off implicit. AFD instead anchors evidence in a slowly varying affordance substrate and applies query-routed, conflict-triggered adaptor updates, explicitly separating stability from plasticity.

Video understanding [1, 57, 60] underpins assistive analytics, retrieval, and embodied agents, where models must parse long, multi-event sequences and align visual dynamics with natural language [17, 21, 28, 53, 61]. Yet real deployments seldom operate in a stationary world: data, domains, and query styles evolve over time [18, 23, 24, 32, 35, 41]. Continual video–language reasoning is therefore central to assistants that must

[†] Equal contribution

* Corresponding author

keep learning from non-stationary streams while answering open-form queries, and step reasoning in videos. Recent studies underscore both the opportunity and difficulty: long egocentric protocols mix heterogeneous tasks and language forms [5, 43], and time-continual pretraining shows that naïve fine-tuning quickly drifts while replay-heavy strategies raise cost and privacy concerns [10, 19, 44, 52, 55]. Robust solutions must retain prior skills, acquire new ones, and remain efficient without relying on storing old videos [3, 14, 27, 46, 47, 51].

Despite rapid progress, two shortcomings persist: (i) *Objectives for stability vs. adaptability are under-specified.* Existing lines either specialize with prompts/adapters or preserve geometry via distillation/topology constraints, but rarely articulate *which structures should remain task-invariant and which should adapt* along a stream—making stability largely incidental and hard to diagnose [2, 6, 41, 65, 68]. (ii) *Plasticity is budgeted heuristically.* Capacity and routing are commonly fixed or task-indexed, while interference is mitigated post hoc by merging or global regularizers. Few approaches use online signals to decide *when/where* to change [12, 40, 42, 48, 58].

This paper asks: *Can continual video–language learning be anchored in a slowly varying, interaction-centered substrate that separates stability from adaptation?* As shown in Fig. 1, we adopt an *affordance-first decomposition* (AFD): videos are mapped into affordance evidence, while adaptation is concentrated in a lightweight, query-routed reasoning module. The affordance evidence provides time-aligned, reusable signals that remain stable across tasks. The routed module focuses plastic updates only where conflict arises, preserving past capabilities without broad parameter drift.

Our contributions are as follows: (i) We introduce an *affordance-first decomposition* that separates a *slowly varying* shared substrate from a *plastic* routed scheduler, clarifying where stability vs. adaptability should live in continual video–language learning. (ii) We operationalize conflict-aware adaptation by query-conditioned per-layer routing and selective capacity growth, and adopt a question-only replay strategy that

is privacy- and memory-friendly. (iii) Across ViLCo[43] and standard VideoQA suites, our approach achieves state-of-the-art results with substantially lower forgetting compared to strong baselines and recent SOTAs, while remaining order-robust and compute-efficient.

2. Related Work

2.1. Continual video-language learning.

Early multimodal CL revealed strong order sensitivity and forgetting in linguistically structured VQA and captioning [8, 11]. ViLCo-Bench later standardized long-video continual protocols and shifted evaluation toward open-form reasoning [43]. Recent SOTAs adapt LLM/VLM backbones via prompting or adapters—ColPro injects collaborative prompts [2], DAM merges dataset-wise adapters at inference [6], and Bisele couples binding with separation to reduce interference [41]. Yet these approaches often specialize by dataset or prompt banks and model interference only implicitly. AFD departs by separating a slowly varying affordance head from a query-routed, conflict-aware scheduler and by relying on question-only replay.

2.2. Parameter-efficient routing and prompting.

Prompt/adaptor methods enable rehearsal-free selectivity for CLIP-style models [40, 48, 54, 67], while adapter/MoE variants improve transfer/retention via selective gating or consolidation [12, 13, 25, 31, 33, 42, 56, 58, 59]. However, capacity is typically fixed (e.g., prompt count or LoRA rank), routers are often task/domain-driven, and the stabilized representation is not made explicit [37]. Our scheduler instead performs per-layer query-conditioned routing over LoRA experts and grows rank only when measured conflict exceeds a threshold, concentrating plasticity while bounding capacity.

2.3. Continual multimodal learning.

Geometry-preserving CL for VLMs aligns cross-/intra-modal similarity or momentum topologies to protect zero-shot ability [36, 65, 66, 68], prag-

matic training shows that warm-start plus replay approaches full re-training at far lower cost [10]. Privacy-aware/data-free directions use structured or synthetic replay and rectify teacher noise [7, 22, 34, 39, 50, 62, 64]. These streams still blur what should remain shared and where to place plasticity. AFD contributes an interpretable, slow-varying affordance substrate.

3. Method

We address continual video–language reasoning where tasks arrive as a stream and both domains and query formats evolve over time. Our approach is **Affordance-First Decomposition (AFD)**. A shared head converts a video into temporally grounded affordance tokens. A plastic LLM-backbone scheduler consumes the query tokens together with these affordance tokens and performs event-level reasoning through per-layer routed low-rank adapters. Stability acts only on the shared head. Plasticity and task specialization are absorbed by the LLM scheduler. Two compact memories enable practical rehearsal. Figure 2 presents the pipeline.

3.1. Problem setup

We work in a streaming setting where tasks arrive over time. Each task provides labeled triples (V, q, y) and an unlabeled clip pool for stability. A video $V = \{F_t\}_{t=1}^T$ is the visual input. A query q is free text. The target y is either an open answer or a temporal span or a step sequence.

The model has two parts: the *shared affordance head* h_ψ converts the video into continuous affordance tokens that are linearly projected into the LLM hidden space to form keys and values (K, V) . The *LLM-backbone scheduler* g_ϕ^{LLM} embeds the query with the same LLM to obtain U and attends to (K, V) to produce the task-appropriate prediction. Overall predictor is

$$f_\Theta(V, q) = g_\phi^{\text{LLM}}(U, K, V) \quad (1)$$

where $U = E_{\text{LLM}}[\text{Tok}(q)]$, $(K, V) = \Pi(h_\psi(V))$. We maintain two small memories that serve training only. $\mathcal{M}_Q$ stores diverse past questions for replay distillation. $\mathcal{M}_A$ stores affordance prototypes

for diagnostics. By design, stability constraints are applied to the shared head h_ψ and task plasticity is absorbed by the scheduler g_ϕ^{LLM} .

3.2. Architecture

Shared affordance head h_ψ Affordances are object–action regularities that vary slowly across domains and tasks. A stable affordance space reduces gradient conflict for downstream reasoning.

Let $X_{1:T} = \text{Enc}_v(V)$ with $X_t \in \mathbb{R}^{d_v}$. The head produces an affordance distribution

$$\begin{aligned} z_t &= f_{\text{st}}(X_t), \\ s_t(a) &= \langle w_a, z_t \rangle, \\ P_t(a) &= \text{softmax}_{a \in \mathcal{V}_A}(s_t(a)/\tau). \end{aligned} \quad (2)$$

We form a sparse renormalized distribution on the top L categories,

$$q_t(a) = \frac{\mathbf{1}[a \in \text{Top-}L] \cdot P_t(a)}{\sum_{a' \in \text{Top-}L} P_t(a')}, \quad (3)$$

then build a continuous token with an embedding table $E_A \in \mathbb{R}^{|\mathcal{V}_A| \times d_a}$,

$$A_t = \sum_{a \in \mathcal{V}_A} q_t(a) E_A[a] \in \mathbb{R}^{d_a}. \quad (4)$$

To interface with the LLM hidden space we project

$$K_t = W_K A_t, \quad V_t = W_V A_t, \quad (5)$$

where stack $K = \text{stack}_t(K_t)$ and $V = \text{stack}_t(V_t)$. $W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times d_a}$.

LLM tokenizer and text interface The scheduler is an LLM. Queries must be embedded in its native space for unified conditioning and generation. Let Tok be the LLM tokenizer with vocabulary $\mathcal{V}_{\text{LLM}}$ and $E_{\text{LLM}} \in \mathbb{R}^{|\mathcal{V}_{\text{LLM}}| \times d_q}$ its input embedding matrix. We compute

$$\begin{aligned} [w_1, \dots, w_L] &= \text{Tok}(q), \\ U &= [u_1, \dots, u_L] \text{ with } u_\ell = E_{\text{LLM}}[w_\ell], \\ u &= \text{Pool}(U). \end{aligned} \quad (6)$$

where $U \in \mathbb{R}^{L \times d_q}$ are token embeddings. $u \in \mathbb{R}^{d_q}$ is a pooled query state used by the router.

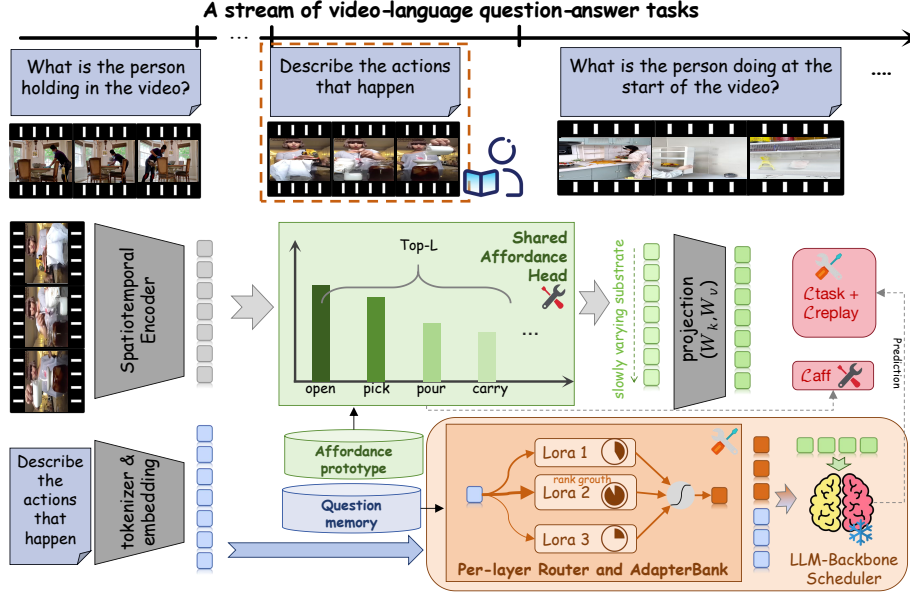


Figure 2. Overview of the proposed Affordance-First Decomposition (AFD) framework for continual video-language question answering. A stream of video-language tasks arrives over time, each video is encoded and mapped by a shared affordance head into slowly varying affordance tokens and prototypes, while questions are embedded and stored for replay to route per-layer LoRA adapters in the LLM-backbone scheduler. Stability loss $\mathcal{L}_{\text{aff}}$ acts only on the affordance head, whereas task and replay losses ($\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{replay}}$) act only on the routed adapters, explicitly separating a stable affordance substrate from a plastic reasoning module.

Per-layer routing and adapter injection Heterogeneous tasks require different reasoning skills. Instance-wise routing focuses plastic capacity and reduces interference. At each adapterized LLM layer $\ell \in \mathcal{S}$ the router computes mixture weights

$$\alpha^{(\ell)} = \text{softmax}(W_r^{(\ell)} u) \in \Delta^{m-1}, \quad (7)$$

and injects a mixture of LoRA experts into the linear map $W^{(\ell)}$

$$\widetilde{W}^{(\ell)} = W^{(\ell)} + \sum_{j=1}^m \alpha_j^{(\ell)} \frac{B_j^{(\ell)} A_j^{(\ell)}}{s_j^{(\ell)}}, \quad (8)$$

where $A_j^{(\ell)} \in \mathbb{R}^{r_j^{(\ell)} \times d_{\text{in}}}$ and $B_j^{(\ell)} \in \mathbb{R}^{d_{\text{out}} \times r_j^{(\ell)}}$ are low-rank factors with rank $r_j^{(\ell)}$ and $s_j^{(\ell)}$ is a scale.

We measure conflict by a clamped negative co-

sine with numerical stabilization

$$c_j^{(k)} = \left[-\frac{\langle g_j^{(k)}, \bar{g}_j^{(1:k-1)} \rangle}{\|g_j^{(k)}\|_2 \|\bar{g}_j^{(1:k-1)}\|_2 + \varepsilon} \right]_+, \quad \varepsilon > 0. \quad (9)$$

The LoRA rank grows discretely by the excess conflict above a threshold and is capped

$$\begin{aligned} \Delta r_j^{(k)} &= \min \left\{ r_{\text{max}} - r_j^{(k-1)}, \lfloor \gamma (c_j^{(k)} - \tau_c)_+ \rfloor \right\}, \\ r_j^{(k)} &= r_j^{(k-1)} + \Delta r_j^{(k)}. \end{aligned} \quad (10)$$

Here $\tau_c \in [0, 1)$ is the threshold, $\gamma > 0$ is a gain, and $(\cdot)_+ = \max\{\cdot, 0\}$. Initialization of new columns can follow a truncated SVD of the projected gradient with Tikhonov-regularized inverses, detailed in the Supplementary Materials.

LLM-backbone scheduler with affordance cross-attention The LLM composes language

evidence with affordance evidence and outputs the final reasoning result while insulating the shared head from frequent changes.

At layers $\ell \in \mathcal{S}$ the LLM attends to affordances

$$\begin{aligned} Q &= UW_Q, \\ H &= \text{Attn}(Q, K, V), \\ r &= \text{Pool}(H, U). \end{aligned} \quad (11)$$

and task heads support three query formats with a unified supervision

$$\begin{aligned} \mathcal{L}_{\text{task}} = & \mathbb{I}_{\text{gen}} \left[- \sum_m \log p(y_m \mid y_{<m}, U, K, V) \right] \\ & + \mathbb{I}_{\text{span}} \left[- \log p_s(t_s) - \log p_e(t_e) + \lambda_u(1 - \text{tIoU}) \right] \\ & + \mathbb{I}_{\text{step}} \left[- \sum_m \log p(\pi_m \mid \pi_{<m}, U, K, V) \right]. \end{aligned} \quad (12)$$

Here $t_s, t_e \in \{1, \dots, T\}$ and tIoU [20] is computed on discrete frame intervals. Each sample activates exactly one head indicated by the selector $\mathbb{I}_{\text{gen}}$ or $\mathbb{I}_{\text{span}}$ or $\mathbb{I}_{\text{step}}$.

3.3. Training objective

Affordance stability on h_ψ We blend weak alignment with teacher consistency

$$\begin{aligned} \mathcal{L}_{\text{aff}} = & \beta \left[- \sum_\ell \log \left(\sum_{t \in S_\ell} \sum_{a \in \mathcal{C}_\ell} P_t(a) \right) \right] \\ & + (1 - \beta) \frac{1}{T} \sum_{t=1}^T \text{KL}(\bar{P}_t \parallel P_t), \end{aligned} \quad (13)$$

where S_ℓ is an ASR span with verb candidates $\mathcal{C}_\ell$. $P_t(\cdot)$ is the current affordance distribution and $\bar{P}_t(\cdot)$ is the frozen teacher from the previous task. The scalar $\beta \in [0, 1]$ balances the two terms. Gradients of $\mathcal{L}_{\text{aff}}$ update ψ only.

Question-only replay distillation on g_ϕ^{LLM} We store diverse past questions and distill on current clips with temperature $T_{\text{kd}} > 0$ and optional confidence masking $\rho \in (0, 1)$

$$\begin{aligned} \mathcal{L}_{\text{replay}} = & \mathbb{E}_{q^{(u)}, V} \text{KL}(\bar{p}_T(\cdot \mid V, q) \parallel p_T(\cdot \mid V, q)), \\ & \text{where } p_T = \text{softmax}(z/T_{\text{kd}}). \end{aligned} \quad (14)$$

and we include only pairs whose teacher maximum probability exceeds ρ to suppress noisy supervision. Here, the expectation is over $q^{(u)} \in \mathcal{M}_Q$ and V , and $p_T = \text{softmax}(z/T_{\text{kd}})$ denotes the temperature-scaled probability distribution. Gradients of $\mathcal{L}_{\text{replay}}$ update ϕ .

Full objective At task k we minimize the three-term objective

$$\mathcal{L}^{(k)} = \mathcal{L}_{\text{task}}^{(k)} + \lambda_{\text{aff}} \mathcal{L}_{\text{aff}}^{(k)} + \lambda_{\text{rep}} \mathcal{L}_{\text{replay}}^{(k)} \quad (15)$$

with positive scalars λ_{aff} and λ_{rep} . Gradients of $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{replay}}$ update the LLM scheduler with routed adapters and rank growth.

4. Experimental Results

4.1. Experimental Setup

Datasets (i) ViLCo-Bench. To evaluate continual video-language reasoning across heterogeneous tasks, we adopt **ViLCo-Bench** with its three continual tracks built from Ego4D: *Moment Query (MQ)*, *Natural Language Query (NLQ)*, and *Visual Query (VQ)* [43]. We follow its official query-incremental protocols: MQ (5 tasks, 110 actions), NLQ (13 tasks, open-vocabulary queries), and VQ (5 tasks with vision queries). **(ii) Continual VideoQA suites.** For domain- and time-incremental VideoQA, we adopt the datasets used by recent continual VL methods [6]. We use the *dataset-incremental* protocol (train adapters sequentially per dataset, evaluate on all test sets) and the *time-incremental* protocol on iVQA by partitioning videos by upload time, as in [6]. **(iii) Complex reasoning VideoQA.** To probe multi-step reasoning and planning, we include **CVQA** from VQAGuider [4] and the **11-VideoQA** benchmark set from LTR [30] (we report the overlapped common sets with our training budget). These tracks are not used for pure continual metrics. Instead, we report zero-shot/finetuned generalization and sequence-trained robustness. **(iv) Long-video understanding (stress test).** To measure long-range temporal robustness of our AFD, we include **VideoMME** and **MLVU** used by LongVU [38] and the long-context evaluations in VideoLLaMB [46].

Table 1. **Domain-Incremental VideoQA on 6 datasets.** Metric: top-1 accuracy (%). Best, second-best, and third-best cells are shaded in **dark** gray, **medium** gray, and **light** gray, respectively.

Method	iVQA	MSVD	MSRVTT	LSMDC	ANet	TGIF	Avg.↑	Forget↓
<i>Upper bounds</i>								
Adapters (Multitask)	39.7	56.6	46.7	62.9	42.2	67.8	52.6	–
Prompt Tuning (Multitask)	35.0	49.0	37.1	57.4	33.9	59.2	45.3	–
<i>Continual methods</i>								
Zero-Shot	26.8	33.0	15.0	51.5	25.5	41.9	32.3	–
Seq-FT	28.4	36.0	23.7	52.1	31.2	67.6	39.8	–
EWC [16]	29.9	39.3	25.5	54.9	32.4	68.5	41.6	–10.9
LwF [26]	28.3	38.2	25.8	56.4	33.6	67.7	41.8	–10.7
L2P [48]	32.8	43.3	32.1	54.8	27.2	54.4	40.8	–4.6
CODA-Prompt [40]	32.9	44.8	28.7	50.7	23.9	54.7	39.6	–5.7
S-Prompts [45]	31.8	45.5	30.2	54.9	27.9	56.1	41.1	–4.2
MoE (adapters)	31.7	37.1	23.9	57.7	28.9	66.8	41.0	–11.6
ColPro [2]	35.3	49.6	36.7	58.4	32.1	61.0	45.5	–3.9
LAE [9]	36.1	50.2	37.5	58.8	32.7	61.5	46.1	–3.4
Biseclle [41]	38.9	52.1	41.3	62.1	35.4	66.3	49.4	–2.7
DAM [6]	39.1	53.6	42.2	63.0	36.3	66.8	50.2	–2.3
AFD (ours)	40.7	55.8	43.7	63.6	38.0	68.1	51.6	–1.8

Notes. The six datasets follow the sequence iVQA → MSVD → MSRVTT → LSMDC → ActivityNet (ANet) → TGIF. “Adapters/Prompt Tuning (Multitask)” jointly train on all datasets and are shown only as non-continual ceilings.

Evaluation Metrics (i) **ViLCo-Bench.** We follow the official metrics [43]: average recall $R@1/R@5$ at IoU thresholds (MQ/NLQ), with *Backward Forgetting* (BwF) as the continual metric; for VQ we report $tAP@0.25$, $stAP@0.25$, average recall, and success rate. We also compute the average performance P across tasks as in [43]. (ii) **VideoQA.** For single-answer VideoQA, we report *top-1 accuracy*. For multiple-choice datasets, we use *MC Acc.* For open-ended datasets with standard VQA-style processing we additionally report *EM/F1* when applicable. (iii) **Continual learning diagnostics.** In addition to BwF on ViLCo, we report *Average Accuracy* after each task and the standard *Backward Transfer* (BWT) and *Forgetting* where defined in the compared papers. For time-incremental iVQA, we report per-slice accuracy and average across time slices [6].

4.2. Main results

VideoQA. Table 1 shows that AFD attains the best average accuracy among continual methods

Table 2. **ViLCo-Bench (Ego4D) under query-incremental protocols.**

Method	MQ $R@1@0.5$ ↑	NLQ $R@1@0.5$ ↑	VQ $stAP@0.25$ ↑
ViLCo [43]	21.2	12.6	13.4
ColPro [2]	26.2	17.8	15.9
DAM [6]	27.1	16.9	16.5
Biseclle [41]	26.8	18.2	16.1
AFD (ours)	29.6	20.7	18.4

(51.6%), surpassing DAM by +1.4 points while maintaining the lowest forgetting (–1.8). Gains are consistent on five of six datasets.

ViLCo-Bench. As summarized in Table 2, AFD achieves the strongest performance on all three query types: +2.5 $R@1$ on MQ, +2.5 $R@1$ on NLQ, and +1.9 $stAP$ on VQ over the best competing baseline.

Beyond the domain-incremental and ViLCo results in Tables 1–2, we further report the time-incremental iVQA protocol and the two non-continual settings (complex reasoning and long-video stress tests) to verify the robustness.

Table 3. **Additional evaluations.** Left: *time-incremental iVQA* (4 temporal slices by upload time). Right: *complex reasoning* (non-continual) and *long-video stress tests* (non-continual). Best, second-best, and third-best cells are shaded in **dark** / **medium** / **light** gray, respectively.

(a) Time-Incremental iVQA (top-1 %, higher is better)							(b) Non-continual references (higher is better)		
Method	S1	S2	S3	S4	Avg.↑	Forget↓	<i>Complex reasoning</i>		
Zero-Shot	26.9	26.6	26.3	26.0	26.5	–	Method	CVQA EM	11-VideoQA
Seq-FT	29.1	27.4	27.0	26.3	27.5	–	ColPro [2]	55.8	60.2
EWC [16]	31.5	30.6	30.0	29.4	30.4	−6.7	LTR [30]	58.7	63.1
LwF [26]	31.2	30.3	29.8	29.1	30.1	−6.9	VQAGuider [4]	61.3	66.5
SMoE [63]	35.3	33.6	31.5	34.2	33.7	−3.5	AFD (ours)	62.8	67.4
DIKI [42]	31.4	28.6	30.2	31.9	30.5	−5.3	<i>Long-video stress test</i>		
DMNSP [15]	34.0	33.6	30.6	33.9	33.0	−4.2	Method	VideoMME	MLVU
SMoLoRA [49]	33.2	31.0	29.4	32.5	31.5	−5.6	VideoLLaMB [46]	60.1	55.9
L2P [48]	34.9	34.0	33.1	32.5	33.6	−3.9	LongVU [38]	61.2	57.3
ColPro [2]	36.1	35.2	34.1	33.5	34.7	−3.3	AFD (ours)	61.7	57.9
Bisecler [41]	39.8	37.7	36.8	36.1	37.6	−2.5			
DAM [6]	40.2	38.1	37.2	36.9	38.1	−2.2			
AFD (ours)	41.8	39.6	38.5	38.1	39.5	−1.6			

Time-Incremental iVQA. As shown in Table 3(a), AFD achieves the best average accuracy (39.5%) with the lowest forgetting (−1.6), outperforming DAM by +1.4 points and Bisecler by +1.9. Per-slice gains are consistent (S1–S4), indicating that affordance-stabilized video tokens alleviate temporal distribution drift.

Complex Reasoning. In Table 3(b, top), AFD attains the highest scores on CVQA and 11-VIDEOQA, slightly surpassing VQAGuider and LTR. These results suggest that AFD’s scheduler composes affordance evidence effectively for multi-step reasoning without explicit tool calls.

Long-Video Stress Tests. Table 3(b, bottom) shows that AFD is better than specialized long-video systems under the same backbone family, while our approach remains architecture-light (no dedicated memory bridges or heavy token pruning).

4.3. Ablation and Analysis

Ablation results We perform single-factor ablations under the same setup as Section 4.1. Variants are : ❶ w/o affordance tokens (direct frame tokens to LLM); ❷ w/o router (uniform adapter mixing); ❸ fixed LoRA rank $r=8$ (no rank growth); ❹ w/o

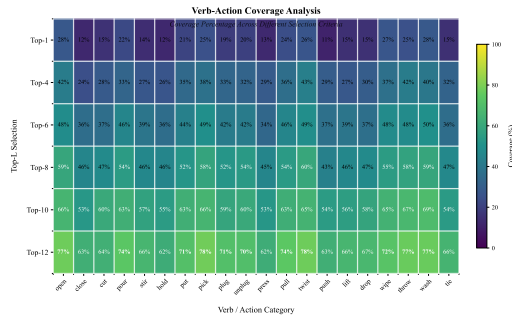
question-only replay ($\lambda_{\text{rep}}=0$); ❺ w/o ASR weak-alignment term in $\mathcal{L}_{\text{aff}}$; ❻ w/o teacher-consistency (KL) in $\mathcal{L}_{\text{aff}}$; ❼ hard sparsity Top- $L=1$; ❽ smaller memory budgets ($B_Q=2k$, $B_A=256$). From Table 4, ❶ produces the largest drop and increases forgetting by 1.5, highlighting the central role of a stable affordance space. ❷ and ❸ further confirm that instance-wise routing and conflict-triggered capacity are both important. Other modules also have a positive impact on performance.

Are affordance tokens stable? We validate the structural division of labor (*stable* affordance head vs. *plastic* LLM scheduler) from a representation perspective. Across tasks $k=1 \dots 6$ (iVQA→MSVD→MSRVTT→LSMDC→ANet→TGIF), we (i) track prototype drift as cosine distance per affordance prototype between consecutive tasks ($k-1 \rightarrow k$), and (ii) compute coverage of verb/action clusters captured by Top- L mixtures ($L \in \{1, 4, 6, 8, 10, 12\}$). From Fig. 3, prototype drift remains small and narrowly distributed across all task transitions, while adjacent-task CKA remains high, evidencing a stable, reusable affordance space. Verb/action coverage rises monotonically with Top- L and plateaus around $L=8$, sup-

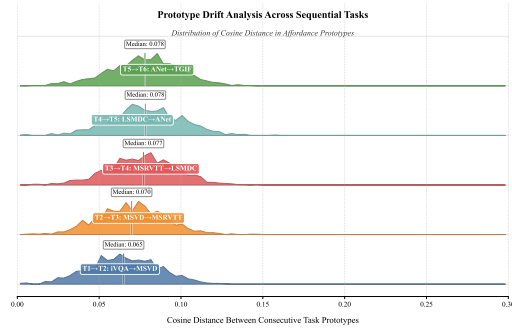
Table 4. Single-factor ablations.

Variant	Domain-Incremental VideoQA				ViLCo-Bench (Ego4D)		
	Avg. Acc. $\uparrow$	MSRVTT $\uparrow$	ANet $\uparrow$	Forget $\downarrow$	MQ R@1@0.5 $\uparrow$	NLQ R@1@0.5 $\uparrow$	VQ stAP@0.25 $\uparrow$
Full AFD	51.6	43.7	38.0	-1.8	29.6	20.7	18.4
1	48.7 (-2.9)	41.1 (-2.6)	35.4 (-2.6)	-3.3 (+1.5)	28.0 (-1.6)	19.1 (-1.6)	16.9 (-1.5)
2	49.8 (-1.8)	42.3 (-1.4)	36.9 (-1.1)	-2.6 (+0.8)	28.5 (-1.1)	19.7 (-1.0)	17.5 (-0.9)
3	50.5 (-1.1)	42.9 (-0.8)	37.3 (-0.7)	-2.3 (+0.5)	28.9 (-0.7)	20.0 (-0.7)	17.8 (-0.6)
4	50.2 (-1.4)	43.0 (-0.7)	36.8 (-1.2)	-2.8 (+1.0)	28.6 (-1.0)	19.8 (-0.9)	17.6 (-0.8)
5	50.8 (-0.8)	43.2 (-0.5)	37.4 (-0.6)	-2.2 (+0.4)	29.0 (-0.6)	19.9 (-0.8)	17.9 (-0.5)
6	50.6 (-1.0)	43.1 (-0.6)	37.2 (-0.8)	-2.4 (+0.6)	28.8 (-0.8)	19.8 (-0.9)	17.7 (-0.7)
7	50.7 (-0.9)	43.0 (-0.7)	37.3 (-0.7)	-2.4 (+0.6)	28.9 (-0.7)	20.1 (-0.6)	17.8 (-0.6)
8	51.0 (-0.6)	43.4 (-0.3)	37.6 (-0.4)	-2.1 (+0.3)	29.2 (-0.4)	20.3 (-0.4)	18.0 (-0.4)

porting the design choice that soft, sparse mixtures encode co-occurring affordances without destabilizing the head.



(a) Verb-action coverage analysis at different Top-L thresholds



(b) Prototype drift distributions across task transitions

Figure 3. **Affordance stability and coverage.** (a) Soft Top- L mixtures increase verb/action coverage without requiring more scheduler capacity. (b) Drift distributions concentrate near zero with small spread across tasks, consistent with a slowly varying shared space.

Case study In Fig. 4, AFD correctly anticipates the interaction, while Bisele[41] focuses on incidental cues. This indicates that our affordance-first decomposition prioritizes object-action regularities over local appearance, improving causal anticipation.

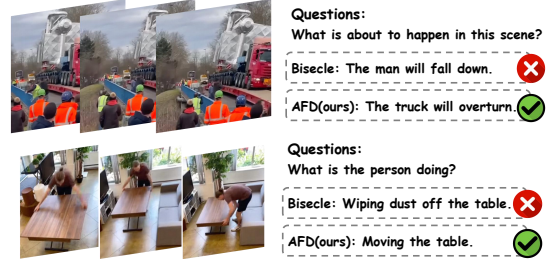


Figure 4. Case study.

5. Conclusion

We addressed the central question of *where stability should live and where plasticity should adapt* in continual video-language learning, proposing an *affordance-first decomposition* with a slowly varying substrate and a query-routed, conflict-aware scheduler. Across ViLCo and domain/time-incremental VideoQA, this split yields state-of-the-art accuracy with substantially lower forgetting, and the diagnostics support the “slowly varying” substrate hypothesis. Future work will explore online affordance discovery, and multi-sensor extensions[29].

References

- [1] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the” video” in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022. 1
- [2] Chen Cai, Zheng Wang, Jianjun Gao, Wenyang Liu, Ye Lu, Runzhong Zhang, and Kim-Hui Yap. Empowering large language model for continual video question answering with collaborative prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3921–3932, 2024. 2, 6, 7
- [3] Yankai Chen, Hongkun Du, and Yutong Zhou. Lightweight network-based semantic segmentation for uavs and its risc-v implementation. *Journal of Technology Innovation and Engineering*, 1(2), 2025. 2
- [4] Yuyan Chen, Jiyuan Jia, Jiaxin Lu, Siyue Li, Yu Guan, Ming Yang, and Qingpei Guo. Vqaguide: Guiding multimodal large language models to answer complex video questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7821–7834, 2025. 5, 7
- [5] Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Haokun Wen, and Weili Guan. Hud: Hierarchical uncertainty-aware disambiguation network for composed video retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6143–6152, 2025. 2
- [6] Feng Cheng, Ziyang Wang, Yi-Lin Sung, Yan-Bo Lin, Mohit Bansal, and Gedas Bertasius. Dam: Dynamic adapter merging for continual video qa learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6805–6817. IEEE, 2025. 2, 5, 6, 7
- [7] Zhenyu Cui, Yuxin Peng, Xun Wang, Manyu Zhu, and Jiahuan Zhou. Continual vision-language retrieval via dynamic knowledge rectification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11704–11712, 2024. 3
- [8] Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew Bagdanov, and Joost Van de Weijer. Ratt: Recurrent attention to transient tasks for continual image captioning. *Advances in Neural Information Processing Systems*, 33:16736–16748, 2020. 2
- [9] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11483–11493, 2023. 6
- [10] Saurabh Garg, Mehrdad Farajtabar, Hadi Pouransari, Raviteja Vemulapalli, Sachin Mehta, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. Tic-clip: Continual training of clip models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3
- [11] Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, 2019. 2
- [12] Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19608–19617, 2025. 2
- [13] Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in neural information processing systems*, 37:129146–129186, 2024. 2
- [14] Linling Jiang, Xin Wang, Fan Zhang, and Caiming Zhang. Transforming time and space: efficient video super-resolution with hybrid attention and deformable transformers. *The Visual Computer*, 41(9):6879–6890, 2025. 2
- [15] Borui Kang, Lei Wang, Zhiping Wu, Tao Feng, Yawen Li, Yang Gao, and Wenbin Li. Dynamic multi-layer null space projection for vision-language continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2077–2086, 2025. 7
- [16] 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. 6, 7
- [17] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization. In *Chinese con-*

- ference on biometric recognition*, pages 180–190. Springer, 2023. 1
- [18] Yingxin Lai, Guoqing Yang, Yifan He, Zhiming Luo, and Shaozi Li. Selective domain-invariant feature for generalizable deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2335–2339. IEEE, 2024. 1
- [19] Yingxin Lai, Cuijie Xu, Haitian Shi, Guoqing Yang, Xiaoning Li, Zhiming Luo, and Shaozi Li. Font-agent: Enhancing font understanding with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19670–19680, 2025. 2
- [20] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–33, 2023. 5
- [21] Gal Lavee, Ehud Rivlin, and Michael Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(5):489–504, 2009. 1
- [22] Yiming Lei, Qiannan Shen, and Junhao Song. A multi-scale graph learning framework with temporal consistency constraints for financial fraud detection in transaction networks under non-stationary conditions. *arXiv preprint arXiv:2603.14592*, 2026. 3
- [23] Haoang Li, Ji Zhao, Jean-Charles Bazin, Pyojin Kim, Kyungdon Joo, Zhenjun Zhao, and Yun-Hui Liu. Hong kong world: Leveraging structural regularity for line-based slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13035–13053, 2023. 1
- [24] Jinming Li and Yutong Zhou. Bideeplab: An improved lightweight multi-scale feature fusion deeplab algorithm for facial recognition on mobile devices. *Computer Simulation in Application*, 3(1):57–65, 2025. 1
- [25] Yuqi Li, Junhao Dong, Chuanguang Yang, Shiping Wen, Piotr Koniusz, Tingwen Huang, Yingli Tian, and Yew-Soon Ong. Mmt-ard: Multimodal multi-teacher adversarial distillation for robust vision-language models. *arXiv preprint arXiv:2511.17448*, 2025. 2
- [26] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 6, 7
- [27] Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5101–5109, 2025. 2
- [28] Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23373–23381, 2026. 1
- [29] Bangyan Liao, Zhenjun Zhao, Haoang Li, Yi Zhou, Yingping Zeng, Hao Li, and Peidong Liu. Convex relaxation for robust vanishing point estimation in manhattan world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15823–15832, 2025. 8
- [30] Zhaohe Liao, Jiangtong Li, Siyu Sun, Qingyang Liu, Fengshun Xiao, Tianjiao Li, Qiang Zhang, Guang Chen, Li Niu, Changjun Jiang, et al. Divide and conquer: Exploring language-centric tree reasoning for video question-answering. In *Forty-second International Conference on Machine Learning*, 2025. 5, 7
- [31] Jiakai Lin, Jinchang Zhang, and Guoyu Lu. Key-point detection and description for raw bayer images. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11736–11742. IEEE, 2025. 2
- [32] Kuangcong Liu, Shini Yang, and Jiayi Xia. Research and practice of advertisement recommendation algorithm based on graph neural network. In *Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems*, pages 210–215, 2025. 1
- [33] Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *The Thirteenth International Conference on Learning Representations*, 2025. 2
- [34] Imad Eddine Marouf, Enzo Tartaglione, Stéphane Lathuilière, and Joost Van De Weijer. Ask and remember: A questions-only replay strategy for continual visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18078–18089, 2025. 3
- [35] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu,

- Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1559–1569, 2025. 1
- [36] Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual vision-language representation learning with off-diagonal information. In *International Conference on Machine Learning*, pages 26129–26149. PMLR, 2023. 2
- [37] Qirui Ning, Jinlai Zhang, Yuhang Xie, Kaifeng Liu, Kai Gao, Bin Chen, Gengbiao Chen, Qing Fan, Hui Liu, and Ronghua Du. Multi-resolution context augmentation and dual channel attention for 3d lane detection. *IEEE Internet of Things Journal*, 2025. 2
- [38] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 5, 7
- [39] James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14994–15004, 2023. 3
- [40] 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*, pages 11909–11919, 2023. 2, 6
- [41] Yue Tan, Xiaoqian Hu, Hao Xue, Celso De Melo, and Flora D Salim. Bisecle: Binding and separation in continual learning for video language understanding. *arXiv preprint arXiv:2507.00469*, 2025. 1, 2, 6, 7, 8
- [42] Longxiang Tang, Zhuotao Tian, Kai Li, Chunming He, Hantao Zhou, Hengshuang Zhao, Xiu Li, and Jiaya Jia. Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. In *Euro-pean conference on computer vision*, pages 346–365. Springer, 2024. 2, 7
- [43] Tianqi Tang, Shohreh Deldari, Hao Xue, Celso De Melo, and Flora Salim. Vilco-bench: Video language continual learning benchmark. *Advances in Neural Information Processing Systems*, 37:70213–70229, 2024. 2, 5, 6
- [44] Vishaal Udandaraao, Karsten Roth, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie, Zeynep Akata, and Matthias Bethge. A practitioner’s guide to real-world continual multimodal pretraining. *Advances in Neural Information Processing Systems*, 37:133801–133845, 2024. 2
- [45] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam’s razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022. 6
- [46] Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24170–24181, 2025. 2, 5, 7
- [47] Yongyao Wang, Ziqi Miao, Lu Yang, Haonan Jia, Wenting Yan, Chen Qian, and Lijun Li. Tabsieve: Explicit in-table evidence selection for tabular prediction. *arXiv preprint arXiv:2602.11700*, 2026. 2
- [48] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 2, 6, 7
- [49] Ziqi Wang, Chang Che, Qi Wang, Yangyang Li, Zenglin Shi, and Meng Wang. Smolora: Exploring and defying dual catastrophic forgetting in continual visual instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 177–186, 2025. 7
- [50] Bin Wu, Wuxuan Shi, Jinqiao Wang, and Mang Ye. Synthetic data is an elegant gift for continual vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2813–2823, 2025. 3
- [51] Jiachun Wu, Jinlai Zhang, Jihong Zhu, Yijian Duan, Youyang Fang, Jingyu Zhu, Lairong Yin, Jiahui Jiang, Zhiyong He, Yi Huang, et al. Multi-

- scale convolution and dynamic task interaction detection head for efficient lightweight plum detection. *Food and Bioprocess Processing*, 149:353–367, 2025. [2](#)
- [52] Sheng Wu and Jinlai Zhang. Spatiotemporal multi-view continual dictionary learning with graph diffusion. *Knowledge-Based Systems*, 316:113388, 2025. [2](#)
- [53] Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. Vmt-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In *Proceedings of the AAAI conference on artificial intelligence*, pages 16085–16093, 2024. [1](#)
- [54] Yi Xin, Siqi Luo, Xuyang Liu, Yuntao Du, Haodi Zhou, Xinyu Cheng, Christina Lee, Junlong Du, Haozhe Wang, Mingcai Chen, et al. V-petl bench: A unified visual parameter-efficient transfer learning benchmark. *Advances in neural information processing systems*, 37:80522–80535, 2024. [2](#)
- [55] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025. [2](#)
- [56] Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vision-language models. *Advances in Neural Information Processing Systems*, 37:51552–51576, 2024. [2](#)
- [57] Jingchun Yang and Jinchang Zhang. Interpretable traffic responsibility from dashcam video via legal multi agent reasoning. *arXiv preprint arXiv:2603.17930*, 2026. [1](#)
- [58] 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*, pages 23219–23230, 2024. [2](#)
- [59] Yu-Chu Yu, Chi-Pin Huang, Jr-Jen Chen, Kai-Po Chang, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In *European Conference on Computer Vision*, pages 219–236. Springer, 2024. [2](#)
- [60] Fan Zhang, Gongguan Chen, Hua Wang, Jinjiang Li, and Caiming Zhang. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4496–4506, 2023. [1](#)
- [61] Jinchang Zhang and Guoyu Lu. Vision-language embodiment for monocular depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29479–29489, 2025. [1](#)
- [62] Jinchang Zhang, Zijun Li, Jiakai Lin, and Guoyu Lu. Adaptive event stream slicing for open-vocabulary event-based object detection via vision-language knowledge distillation. *arXiv preprint arXiv:2510.00681*, 2025. [3](#)
- [63] Liang Zhang, Ziyao Lu, Fandong Meng, Hui Li, Jie Zhou, and Jinsong Su. Advancing smoe for continuous domain adaptation of mllms: Adaptive router and domain-specific loss. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26584–26602, 2025. [7](#)
- [64] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19102–19112, 2023. [3](#)
- [65] 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*, pages 19125–19136, 2023. [2](#)
- [66] Heng Zhou, Jing Tang, Yanshu Li, Canran Xiao, Liwei Hou, Zong Ke, Jiawei Yao, et al. Comem: Compositional concept-graph memory for vision-language adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026. [2](#)
- [67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. [2](#)
- [68] Hongguang Zhu, Yunchao Wei, Xiaodan Liang, Chunjie Zhang, and Yao Zhao. Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22257–22267, 2023. [2](#)