Towards Lifelong Video Understanding: A Survey on Continual Learning in Video Visual Question Answering

Anonymous EMNLP submission

Abstract

This paper surveys the application of continual learning in Video Visual Question Answering (Video VQA) to advance lifelong video understanding. With the rapid progress in VQA technologies, models perform excellently in static environments but face significant challenges in real-world scenarios, particularly catastrophic forgetting when encountering new tasks or domains. We systematically review the fundamentals of video VQA, including the evolution from image to video, core architectures, and evaluation methods, and thoroughly explore how continual learning techniques are adapted to the video understanding domain. We analyze implementation strategies based on regularization, replay, parameter isolation, and hybrid methods, comparing their performance across different video VQA task streams. The paper discusses experimental evaluation frameworks, spanning task division (by question type, domain, and video style), training protocols, and baseline model selection (joint training, sequential fine-tuning, and independent training). Additionally, we identify current challenges such as long video understanding, modality imbalance, and computational efficiency concerns, while exploring future research directions and potential application scenarios. This survey aims to integrate recent advances, highlight critical trends, and provide guidance for the development of continual video VQA learning.

1 Introduction

Visual Question Answering (VQA) challenges models to answer natural language questions by grounding them in visual content. Early image-based VQA fused convolutional features with attention-based language models to reason over single frames (Antol et al., 2015; Goyal et al., 2017). Yet many applications (from autonomous driving to intelligent tutoring) demand understanding dynamic scenes (Pandey et al., 2025).

Video-based VQA extends this by requiring spatiotemporal reasoning: models must recognize and track objects and actions across frames, capture motion cues, and integrate information over time to answer complex, multi-step questions (Xu et al., 2017; Jang et al., 2017). State-of-the-art architectures typically use frozen visual encoders to extract frame-level features, fine-tuned token embeddings for question semantics, and large pretrained language models to generate answers (see Figure 1). Enhanced temporal attention and cross-modal fusion have boosted performance, but evaluation remains confined to static datasets.

In real-world deployments, video distributions and query types evolve: new objects appear, lighting and viewpoints shift, and question formats change. Retraining VQA models from scratch is costly and risks catastrophic forgetting of prior capabilities. Continual (lifelong) learning addresses this by enabling incremental updates on non-stationary data streams while preserving earlier knowledge (Laal and Salamati, 2012). Applied to video VQA, it must handle heterogeneous modalities, maintain temporal information retention, and operate under memory and compute constraints.

Research at the intersection of continual learning and VQA spans regularization-based methods (e.g. EWC, MAS) (Kirkpatrick et al., 2017; Aljundi et al., 2018), replay-based strategies (experience or generative replay) (Rolnick et al., 2019; Shin et al., 2017), parameter-isolation techniques (adapters, mask-based pruning)(Cheng et al., 2024; Mallya et al., 2018), and emerging meta-learning (Riemer et al., 2019) and prompt-based adaptations (Qian et al., 2023) leveraging large-scale pretrained models. However, these efforts remain fragmented, lacking a unified view of task formulations, benchmark splits, evaluation protocols, and open challenges specific to continual video VQA.

Through this survey, we aim to provide a clear understanding of continual learning applications in

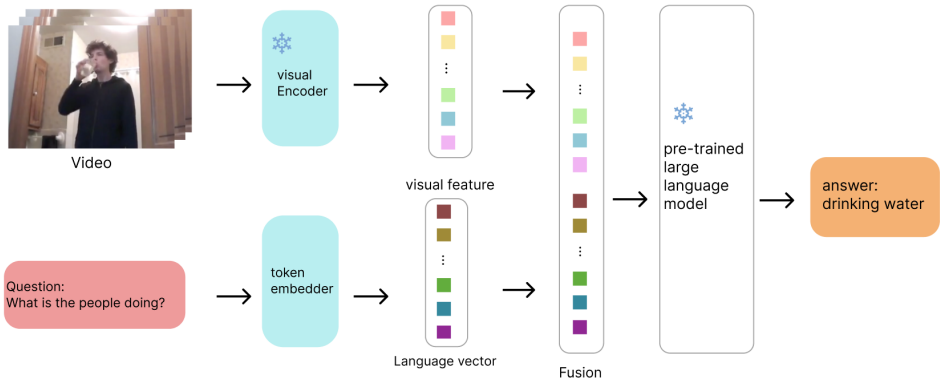


Figure 1: Architecture of a common VQA framework: The visual encoder is frozen to extract robust visual features; the token embedder is unfrozen to fine-tune task-specific linguistic representations; and the pretrained LLM is kept frozen to leverage its strong generative capability for answer production.

video VQA, facilitate knowledge sharing among researchers, and offer valuable guidance for future work. As video data becomes increasingly prevalent in real-world applications, we believe lifelong video understanding will emerge as a crucial frontier in artificial intelligence research.

2 Background

2.1 Visual Question Answering (VQA)

Visual Question Answering (VQA) is a multimodal task in which a model must generate a natural language answer given an image or video and a corresponding question (see Figure 1 for the overall architecture). Early work on *image-based* VQA focuses on spatial reasoning from a single frame, where models learn to associate visual regions with question tokens (e.g., (Antol et al., 2015; Goyal et al., 2017)). In contrast, *video-based* VQA extends this to spatiotemporal reasoning: the model must capture motion cues, frame-to-frame dependencies, and evolving contextual information across time (e.g., (Cai et al., 2024; Cheng et al., 2024; Zhang et al., 2023; Qian et al., 2023; He et al., 2024)).

2.1.1 Core model architecture

State-of-the-art video VQA models typically consist of several key components designed to handle the spatiotemporal nature of videos (P.J. and Kovoor, 2024):

Multimodal Fusion Strategies. To combine textual and visual information, multimodal fusion is critical. Common approaches include:

- **Early Fusion:** Concatenate features from different modalities (Barnum et al., 2020) at the input level to learn joint representations.
- **Late Fusion:** Independently process each modality and fuse outputs or embeddings in a subsequent stage (Shankar et al., 2022).
- **Cross-Modal Attention:** Use attention mechanisms for dynamic interactions between vision and language streams, improving context-aware reasoning (Nagrani et al., 2022).

Spatiotemporal Reasoning Mechanisms. Effective reasoning over both space and time is essential:

- **Temporal Attention and Graphs:** Focus on key frames or build temporal graphs to model interactions across time (Khan et al., 2023).
- **Recurrent Layers and Transformers:** Integrate LSTM/GRU units or temporal transformer blocks (Lei et al., 2019; Gao et al., 2022) to capture long-range dependencies.
- **Memory Modules:** Use external-memory or memory-augmented networks to store and retrieve relevant past information for answering questions (Bai et al., 2024; He et al., 2024).

2.1.2 Evaluation Metrics for VQA

VQA tasks can be broadly divided into open-ended and multiple-choice settings. The evaluation methods differ depending on the question type, and additional metrics may be applied based on the task's specifics.

Open-Ended VQA. Evaluation of open-ended VQA frequently uses n-gram overlap metrics like BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) to compare model outputs with reference answers, though these metrics have limitations for truly open-ended responses where various phrasings can be equally correct.

More sophisticated approaches employ semantic similarity measures using embedding models (like BERT) (Reimers and Gurevych, 2019; Zhang* et al., 2020), where correctness is determined by the cosine similarity between embeddings of predicted and reference answers, with a threshold defining correct responses.

Multiple-Choice VQA: In multiple-choice settings, the task is treated as a standard classification problem. The accuracy is computed as:

$$\text{Accuracy} = \frac{\text{Number of correctly selected answers}}{\text{Total number of questions}} \quad (1)$$

Additionally, ranking metrics such as Mean Reciprocal Rank (Zhong et al., 2017) can be used if candidate answer ranking is important.

2.1.3 Current Research Status

Current research in video VQA has diversified along several fronts: a wide range of model architectures has emerged, from unified end-to-end networks that jointly learn spatiotemporal features (Gao et al., 2018) to modular pipelines in which individual components (e.g., feature extraction, temporal reasoning, question encoding) are optimized separately (Park et al., 2021); performance has steadily improved as newer methods leverage sophisticated temporal attention mechanisms and richer multimodal fusion strategies, often outperforming simple extensions of image-based VQA models; nonetheless, researchers continue to grapple with key challenges, including the high computational cost and memory footprint of processing long video sequences, the difficulty of maintaining robustness under occlusions or rapid scene transitions, and the need to scale effectively to large, diverse video corpora.

2.2 Continual Learning

2.2.1 Basics of Continual (Lifelong) Learning

Continual learning (Grossberg, 2012) is the paradigm in which a model ingests a non-stationary stream of tasks or data and incrementally updates its parameters without retraining from scratch. A central obstacle is catastrophic forgetting, where new learning overwrites representations critical for earlier tasks. Closely related is the plasticity–stability trade-off (Grossberg, 1987): a model must remain plastic enough to acquire new knowledge yet stable enough to preserve old knowledge. Other emerging challenges include handling unknown task boundaries, scaling to large numbers of tasks under limited memory and compute budgets, and devising evaluation schemes that faithfully capture both learning and retention over time.

2.2.2 Multimodal Continual Learning

When applying continual learning to multimodal tasks, such as Video VQA, the model must jointly update across heterogeneous inputs (e.g., images/video frames and text) while avoiding modality-specific forgetting (Yu et al., 2024a). In video QA settings, three extra requirements arise:

- **Temporal Information Retention**: preserving frame-to-frame dynamics and motion cues so that reasoning over sequences does not degrade when new video tasks are learned.
- **Balanced Multimodal Retention**: ensuring that neither the visual stream nor the language stream disproportionately forgets past knowledge.
- **Efficiency under Lengthy Inputs:** maintaining a small memory footprint (e.g., via selective buffering or compressed replay) and low computational overhead (e.g., lightweight adapters) to feasibly handle long, high-resolution video clips.

2.2.3 Current Solution Strategies

Researchers have developed a variety of strategies to address the challenges in continual learning. These strategies can be broadly categorized as follows:

- **Regularization-Based Methods:** Methods that add constraints to prevent important parameters from changing drastically, such as EWC (Kirkpatrick et al., 2017) and MAS (Aljundi et al., 2018). Recent extensions include Transformer Calibration (Yang et al., 2022) and LPC (Li et al., 2022).
- **Replay-Based Methods:** Techniques that mitigate forgetting by revisiting past samples, e.g., Experience Replay (Rolnick et al., 2019) and generative replay (Shin et al., 2017).

- **Parameter Isolation Methods:** Methods that allocate dedicated parameters for different tasks, such as task-specific adapters (see e.g., TL-CL (Satapara and Srijith, 2024)) and mask-based methods like EXSSNET (Yadav and Bansal, 2023).

- **Hybrid Methods:** Approaches that combine the above strategies to balance stability and plasticity (e.g., CLIF (Jin et al., 2022) and Mixture-of-Experts adapters (Yu et al., 2024b)).

2.2.4 Evaluation Metrics

Continual Learning (Lifelong Learning) involves sequentially training on multiple tasks while retaining knowledge from previous tasks. The evaluation metrics are designed to capture both the performance on new tasks and the retention of past knowledge.

- **Average Accuracy (ACC):** After learning T tasks sequentially, the overall performance is measured by:

$$\text{ACC} = \frac{1}{T}\sum_{i=1}^{T} R_{T,i} \tag{2}$$

where $R_{T,i}$ denotes the test accuracy on task i after training up to the final task T.

- **Average Forgetting (AF):** Forgetting quantifies the degradation in performance on earlier tasks after learning new tasks. With overall forgetting computed as the average for the first $T-1$ tasks:

$$F = \frac{1}{T-1}\sum_{i=1}^{T-1} F_i \tag{3}$$

- **Forward and Backward Transfer (FWT & BWT)** Following Lopez–Paz (Lopez-Paz, 2022), let $R \in \mathbb{R}^{T\times T}$ be the accuracy matrix where $R_{i,j}$ denotes the test accuracy on task j after training sequentially up to task i, let $\bar{b}_j$ be the baseline (random-init) accuracy on task j, and let $R_{i,i}$ be the accuracy on task i immediately after learning it. Then:

$$\text{FWT} = \frac{1}{T-1}\sum_{j=2}^{T}\left(R_{j-1,j} - \bar{b}_j\right) \tag{4}$$

$$\text{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left(R_{T,i} - R_{i,i}\right) \tag{5}$$

3 Task Formulations & Paradigms in Continual Learning for Video VQA

Continual learning in Video Visual Question Answering (Video VQA) (Zhang et al., 2023) can be cast under several complementary paradigms. Each paradigm defines a different way in which new data or tasks arrive over time, and places distinct requirements on the model's ability to retain past knowledge and transfer to new scenarios.

3.1 Task-Incremental Learning

In the task-incremental setting, the model is presented with a sequence of distinct tasks, each associated with its own question–answer distribution (for example, counting questions, action-recognition questions, or object-tracking questions) (Zhang et al., 2023). At training time, each task arrives with a unique identifier, and during inference the model is told which task it should perform. The goal is to optimize performance on each task in turn while avoiding catastrophic forgetting of previously learned tasks. Task-incremental Video VQA thus requires mechanisms for task-specific parameter isolation or task-conditioned routing, so that representations for new question types do not overwrite those learned for earlier tasks.

3.2 Domain-Incremental Learning

Domain-incremental learning assumes a single underlying task (e.g. the same question types and answer vocabulary) but with the input video distribution shifting over time (Cheng et al., 2024). For Video VQA, this could correspond to different camera viewpoints, lighting conditions, or video genres (sports, surveillance, movies). The model is not given explicit domain labels at test time, and must maintain invariance to domain shifts while continually adapting its visual and temporal features. Effective domain-incremental strategies leverage domain-robust feature extraction, normalization layers that adapt to new contexts, and replay or alignment losses to preserve performance on earlier video domains.

3.3 Class-Incremental Learning

In the class-incremental scenario, the set of possible answers grows as new classes are introduced over time (Marouf et al., 2025; Chen et al., 2025). For Video VQA, this might mean gradually exposing the model to novel action verbs, object categories, or compositional phrases. Critically, at

inference the model must classify among the entire union of all seen answer classes, without being told which subset applies. Class-incremental Video VQA demands dynamic output expansions, balanced replay of old answer examples, and bias-correction techniques to prevent the model from over-favoring newly introduced classes.

3.4 Online Streaming Learning

Online streaming learning represents the most demanding paradigm, where individual video–question pairs arrive sequentially in a single pass and the model must update on each example under strict memory and compute constraints (Cheng et al., 2024; Zhang et al., 2024). There is no clear task or domain boundary, and the model cannot revisit past data except via a limited buffer. In Video VQA, streaming learning challenges include efficient sampling of key frames, continual alignment of multimodal features, and lightweight update rules (e.g. parameter regularization or prototype-based updates) that minimize interruption while preserving long-term knowledge.

4 Methodologies in Continual Learning for Video VQA

Continual learning for Video VQA seeks to equip models with the ability to assimilate new information (e.g. new video domains, question types, or answer categories) while preserving previously acquired capabilities. Over the past decade, researchers have pursued several complementary strategies to strike this balance between stability and plasticity (Grossberg, 1987; Kirkpatrick et al., 2017). Broadly, these approaches fall into replay-based, regularization-based, and parameter-isolation methods, with recent advances in meta-learning and prompt-based adaptation using large pre-trained models (see detailed in Appendix Table A2). In the following subsections, we survey each of these paradigms, highlighting their core mechanisms, strengths, and challenges in the context of Video VQA.

4.1 Replay-based Methods

Replay-based methods mitigate forgetting by revisiting past experiences, either by storing real examples or by generating synthetic ones.

Episodic Replay A fixed-size buffer of video–question–answer triplets is maintained, often via reservoir sampling to ensure unbiased coverage of all past tasks (Lopez-Paz, 2022). During training on a new task, mini-batches are sampled jointly from the current stream and the buffer, interleaving past examples to reinforce previously learned associations between video features and question semantics (Rolnick et al., 2019; Chaudhry et al., 2019).

Generative Replay Instead of storing raw examples, a generative model (e.g. VAE, GAN, or transformer) is trained alongside the VQA model to approximate the joint distribution of past video–QA pairs. On each update, the generator produces pseudo–examples for rehearsal, enabling the VQA model to rehearse earlier tasks without explicit memory of raw data.

4.2 Regularization-based Methods

Regularization methods constrain parameter updates to preserve knowledge deemed important for past tasks.

Parameter Importance Regularization Each weight's importance is estimated (e.g. via the Fisher information matrix in EWC or path-integral measures in SI) after each task (Zenke et al., 2017; Aljundi et al., 2018). When learning a new task, changes to highly important parameters incur a quadratic penalty, thus preventing drastic overwriting of critical features for earlier video question types .

Distillation-based Methods Knowledge distillation preserves the behavior of the model on previous tasks by matching its output distributions (soft logits) on either stored examples or generated pseudo-samples. A distillation loss encourages the updated model to mimic the "teacher" snapshot before learning the new task, maintaining temporal reasoning or object-counting capabilities acquired earlier (Hinton et al., 2015).

4.3 Parameter-Isolation Methods

These methods allocate disjoint subsets of network parameters to different tasks, avoiding interference at the cost of increased capacity.

Dynamic Architectures The network dynamically grows by adding new modules, heads, or layers for each incoming task. For Video VQA, this might involve task-specific attention blocks or question–type adapters, while sharing a common

backbone for video feature extraction (Yu et al., 2024b).

Mask-based Pruning A single overparameterized network is trained once, then for each new task a binary mask selects a subnetwork (e.g. via iterative pruning or learned mask-generators). Masks can be switched at inference to recover task-specific parameters without interference (Mallya et al., 2018; Yadav and Bansal, 2023).

4.4 LLM-based & Prompting Methods

Recent advances leverage large pre-trained multimodal models (e.g. CLIP, VideoBERT, Flamingo) with prompt tuning. Instead of fine-tuning all weights, only a small set of soft or discrete prompts is learned per task or domain (Qian et al., 2023; Cai et al., 2024). This allows continual expansion to new question formats or video genres by appending new prompts to guide the frozen backbone, drastically reducing catastrophic forgetting and memory footprint.

5 Datasets & Evaluation Protocols

To rigorously assess continual learning methods in Video VQA, it is essential to select representative video QA benchmarks, define appropriate task sequences, and adopt standardized evaluation procedures. In this section, we first review the most commonly used Video VQA datasets (as summarized in Appendix Table A1 and in Appendix Figure A1), then describe how to derive continual-learning splits and task orders, and finally point to the evaluation metrics defined earlier.

5.1 Typical Video VQA Datasets

Several large-scale datasets have become standard testbeds for Video VQA research. We summarize the key properties below:

- **MSVD-QA** (Xu et al., 2017) Based on the MSVD video description corpus, it contains ~2k short clips and ~50k QA pairs covering object, action, and temporal reasoning questions.

- **MSRVTT-QA** (Xu et al., 2017) Built on the MSR Video-to-Text dataset, it includes ~10k diverse web videos and ~243k question–answer pairs, focusing on counting, appearance, and transition questions.

- **TGIF-QA** (Jang et al., 2017) Derived from GIF clips, it provides three QA tasks (FrameQA, CountQA, and ActionQA) designed to evaluate both spatial and temporal understanding over short, looped animations.

- **ActivityNet-QA** (Yu et al., 2019) Extends the ActivityNet dataset with open-ended questions about complex, untrimmed videos, requiring multi-step inference across longer time horizons.

5.2 Continual Learning Splits & Task Sequence Design

To transform a static Video VQA dataset into a continual-learning benchmark, the video–QA pairs must be partitioned and ordered into a sequence of tasks. Common strategies include:

- **Question-Type Split:** Group QA pairs by semantic category (e.g. "counting" vs. "object recognition" vs. "action inference") and present each category as a separate task (Zhang et al., 2023; Cai et al., 2024).

- **Answer-Vocabulary Split:** Divide the answer space into disjoint subsets (e.g. colors, numbers, verbs) and introduce each subset incrementally to simulate class-incremental learning (Greco et al., 2019).

- **Domain Split:** Partition videos by domain factors (e.g. indoor vs. outdoor, sports vs. cooking) to create domain-incremental tasks where the question types remain the same but the visual appearance shifts (Zhang et al., 2025; Cheng et al., 2024).

- **Curriculum vs. Random Ordering:** Tasks may be arranged in increasing difficulty (curriculum learning) (Yuan et al., 2022; Akl et al., 2024) or in a randomized sequence to evaluate robustness to task order.

5.3 Evaluation Metrics

To quantitatively assess continual learning in Video VQA, we summarize the following metrics defined in section 2.2.4, which jointly capture a model's ability to acquire new knowledge (plasticity) and to retain past knowledge (stability).

- **Average Accuracy (ACC)**: This metric computes the mean test accuracy across all tasks

once the model has finished learning the entire sequence. It reflects the model's overall ability to both acquire new knowledge and retain previously learned skills, serving as a single-value summary of final performance.

- **Forgetting**: For each earlier task, record its highest accuracy at any point during training and compare it to its accuracy at the end of the sequence. The average drop across all tasks quantifies how much performance deteriorated due to learning subsequent tasks, directly measuring catastrophic forgetting.

- **Backward Transfer (BWT)**: This measures how learning later tasks affects performance on prior tasks. Positive backward transfer indicates that training on new tasks has improved earlier task accuracy (beneficial synergy), while negative backward transfer reveals interference and forgetting caused by the new tasks.

- **Forward Transfer (FWT)**: This captures the influence that knowledge from earlier tasks has on learning new tasks. By comparing the model's performance on each new task before and after any training on that task, forward transfer quantifies how much prior experience accelerates or boosts learning of future tasks, indicating zero-shot or few-shot transfer ability.

These metrics jointly measure a model's stability (resistance to forgetting) and plasticity (ability to learn new tasks efficiently), providing a comprehensive evaluation of continual learning performance in Video VQA.

6 Case Studies & Applications

To demonstrate the practical utility of continual learning in Video VQA, we highlight three representative application scenarios.

6.1 Real-time Video Question Answering Systems

In interactive settings such as live sports commentary or multimedia customer support, models must answer questions on streaming video frames with minimal latency. Continual learning enables these systems to update on new content (e.g., emerging players, novel camera angles, or newly defined question types), without retraining from scratch. Techniques like lightweight adapter modules (Cheng et al., 2024) or prompt tuning (Qian et al., 2023) ensure quick adaptation while preserving core QA capabilities on previously seen video domains.

6.2 Autonomous Driving & Security Surveillance

In autonomous vehicles and surveillance platforms, video streams evolve over time due to changing environments, weather conditions, or novel event categories (e.g., new traffic signs, unusual behaviors) (Lin et al., 2025). Continual Video VQA allows models to incorporate these shifts, answering queries such as "Is there a pedestrian in the crosswalk?" or "Has the traffic light changed color?" while retaining accuracy on earlier learned scenarios. Replay-based methods with bounded memory buffers and domain-robust feature extractors are particularly effective at minimizing catastrophic forgetting in safety-critical deployments.

6.3 Educational Tools & Human–Computer Interaction

In intelligent tutoring systems and assistive interfaces, Video VQA can support interactive learning by answering student questions about video lectures, laboratory demonstrations, or instructional animations (Du et al., 2024). As curricula evolve (new concepts, visual experiments, or updated teaching styles), continual learning ensures that the QA model stays current without losing proficiency on foundational topics. Regularization-based strategies (e.g., EWC) and prompt-based adaptation facilitate seamless updates in educational platforms with minimal human intervention.

7 Challenges & Open Issues

Although continual learning for Video VQA has advanced significantly, several fundamental challenges and open research directions remain.

7.1 Long term Dependencies and Temporal Modeling

Videos often contain long range dependencies spanning hundreds of frames, making it difficult for models to maintain coherent reasoning over time. Temporal modeling architectures such as transformers or memory modules can mitigate this issue, but they may scale poorly and tend to forget early frames when updated on new tasks. Open questions

include designing efficient architectures that capture extended context without excessive compute, and developing temporal abstraction mechanisms to compress video information while preserving crucial long term cues.

7.2 Balancing Memory and Computational Costs

Replay buffers or generative replay introduce storage and compute overhead that grows with the number of tasks or video length. Parameter isolation methods trade parameter efficiency against task coverage, but may not scale as the task sequence grows. Future research should focus on adaptive memory management strategies (for example dynamic coreset selection or compression) and lightweight update rules (for example low rank adapters or prompt tuning) to achieve scalable continual learning in resource constrained environments.

7.3 Multimodal Feature Alignment

Video VQA requires aligning visual features with linguistic representations across tasks, yet continual updates can disrupt previously learned alignment and degrade cross task performance. Domain or task shifts exacerbate this misalignment, leading to poor retention of earlier question types. Open issues include developing stable multimodal embedding spaces, contrastive objectives that resist forgetting, and calibration techniques to maintain semantic consistency across evolving video and language distributions.

7.4 Explainability and Safety

As Video VQA systems are deployed in critical applications (for example surveillance, autonomous driving, or healthcare), interpretability of model decisions and robustness to adversarial inputs become paramount. Existing continual learning methods focus primarily on accuracy and forgetting, with limited attention to explainable reasoning or safety constraints. Future work must integrate interpretability modules (for example attention visualization or causal reasoning traces) and certify safety properties (for example bounded error under domain shift or secure memory access) to ensure trustworthy Video VQA in dynamic environments.

8 Future Directions

Looking forward, several promising directions can further enhance continual learning for Video VQA.

The proliferation of video–language foundation models (e.g. VideoCLIP, Flamingo, GPT-4V) offers rich spatiotemporal and semantic priors. Future work should investigate tighter coupling between these backbones and continual learning strategies, for example by dynamically selecting which layers to freeze or adapt, by designing task-specific soft prompts that evolve with new data, or by leveraging cross-task contrastive objectives to maintain shared representation quality across sequential updates.

Real-world Video VQA systems demand low-latency adaptation under limited memory and compute budgets. Research should focus on event-driven updates (only updating on uncertain or novel examples), adaptive buffer schemes (prioritizing samples with high forgetting risk), and lightweight optimization techniques (such as sketching gradients or low-rank parameter updates) to enable scalable, real-time continual learning on edge devices.

9 Conclusion

In this survey, we have provided a systematic overview of continual learning techniques applied to Video Visual Question Answering. We began by contrasting video VQA with its image-based counterpart, highlighting the unique challenges posed by temporal dynamics and multimodal integration. We then categorized existing continual learning methods into replay-based, regularization-based, parameter-isolation, meta-learning, and prompt-based approaches, and analyzed their strengths and weaknesses in the context of Video VQA. We reviewed common benchmark datasets and outlined strategies for constructing continual-learning splits and evaluation protocols. Finally, we discussed practical applications, identified open challenges, and suggested future research directions. By synthesizing recent advances and organizing them into a coherent framework, we aim to guide researchers toward the development of robust, scalable lifelong video understanding systems.

10 Limitations

While this survey covers a broad spectrum of methodologies and applications, several limitations remain. First, our discussion focuses primarily on supervised continual learning and does not deeply explore unsupervised or semi-supervised paradigms. Second, the benchmarks reviewed are mostly medium-scale datasets; we do not evaluate performance on very large or real-world video cor-

pora. Third, empirical comparisons between methods are based on published results under heterogeneous protocols, which may limit direct fairness. Finally, rapidly evolving foundation models and hardware accelerators are changing the practical feasibility of some approaches; our analysis may become outdated as new architectures and compute paradigms emerge.

References

Ahmed Akl, Abdelwahed Khamis, Zhe Wang, Ali Cheraghian, Sara Khalifa, and Kewen Wang. 2024. Task progressive curriculum learning for robust visual question answering.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Ziyi Bai, Ruiping Wang, and Xilin Chen. 2024. Glance and focus: Memory prompting for multi-event video question answering.

George Barnum, Sabera Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning.

Chen Cai, Zheng Wang, Jianjun Gao, Wenyang Liu, Ye Lu, Runzhong Zhang, and Kim-Hui Yap. 2024. 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.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019. Efficient lifelong learning with a-gem.

Tieyuan Chen, Huabin Liu, Chern Hong Lim, John See, Xing Gao, Junhui Hou, and Weiyao Lin. 2025. Csta: Spatial-temporal causal adaptive learning for exemplar-free video class-incremental learning.

Feng Cheng, Ziyang Wang, Yi-Lin Sung, Yan-Bo Lin, Mohit Bansal, and Gedas Bertasius. 2024. Dam: Dynamic adapter merging for continual video qa learning.

Yuyang Du, Kexin Chen, Yue Zhan, Chang Han Low, Tao You, Mobarakol Islam, Ziyu Guo, Yueming Jin, Guangyong Chen, and Pheng-Ann Heng. 2024. Llm-assisted multi-teacher continual learning for visual question answering in robotic surgery.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2022. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering.

Stephen Grossberg. 1987. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23–63.

Stephen T Grossberg. 2012. *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*, volume 70. Springer Science & Business Media.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering.

Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. 2022. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning.

Aisha Urooj Khan, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels Lobo, and Mubarak Shah. 2023. Learning situation hypergraphs for video question answering.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Marjan Laal and Peyman Salamati. 2012. Lifelong learning; why do we need it? *Procedia - Social and Behavioral Sciences*, 31:399–403. World Conference on Learning, Teaching Administration - 2011.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. Tvqa: Localized, compositional video question answering.

Xiaodi Li, Zhuoyi Wang, Dingcheng Li, Latifur Khan, and Bhavani Thuraisingham. 2022. LPC: A logits and parameter calibration framework for continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7142–7155, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yuxin Lin, Mengshi Qi, Liang Liu, and Huadong Ma. 2025. Vlm-assisted continual learning for visual question answering in self-driving.

David Lopez-Paz. 2022. Gradient episodic memory for continual learning.

Arun Mallya, Dillon Davis, and Svetlana Lazebnik. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights.

Imad Eddine Marouf, Enzo Tartaglione, Stephane Lathuiliere, and Joost van de Weijer. 2025. No images, no problem: Retaining knowledge in continual vqa with questions-only memory.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2022. Attention bottlenecks for multimodal fusion.

Anupam Pandey, Deepjyoti Bodo, Arpan Phukan, and Asif Ekbal. 2025. The quest for visual understanding: A journey through the evolution of visual question answering.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 2021. Bridge to answer: Structure-aware graph interaction network for video question answering.

Jeshmol P.J. and Binsu C. Kovoor. 2024. Video question answering: A survey of the state-of-the-art. *Journal of Visual Communication and Image Representation*, 105:104320.

Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. 2023. Decouple Before Interact: Multi-Modal Prompt Learning for Continual Visual Question Answering . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2941–2950, Los Alamitos, CA, USA. IEEE Computer Society.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. 2019. Experience replay for continual learning.

Shrey Satapara and P. K. Srijith. 2024. TL-CL: Task and language incremental continual learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12123–12142, Miami, Florida, USA. Association for Computational Linguistics.

Shiv Shankar, Laure Thompson, and Madalina Fiterau. 2022. Progressive fusion for multimodal integration.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*.

Prateek Yadav and Mohit Bansal. 2023. Exclusive supermask subnetwork training for continual learning.

Peng Yang, Dingcheng Li, and Ping Li. 2022. Continual learning for natural language generations with transformer calibration. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 40–49, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S. Yu, and Irwin King. 2024a. Recent advances of multimodal continual learning: A comprehensive survey.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. 2024b. Boosting continual learning of vision-language models via mixture-of-experts adapters.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering.

Zhenghang Yuan, Lichao Mou, Qi Wang, and Xiao Xiang Zhu. 2022. From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024. Flash-vstream: Memory-based real-time understanding for long video streams.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xi Zhang, Feifei Zhang, and Changsheng Xu. 2023. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yao Zhang, Haokun Chen, Ahmed Frikha, Denis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. 2025. CL-Cross VQA: A Continual Learning Benchmark for Cross-Domain Visual Question Answering . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6269–6278.

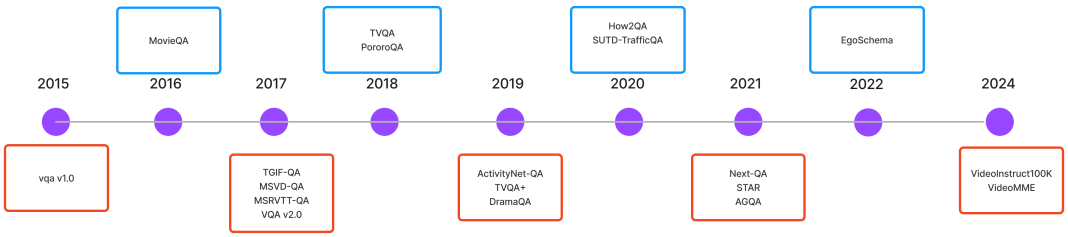
Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327.

Appendix Table A1: Comprehensive Comparison of Visual Question Answering Datasets

Dataset	Year	# Images/Videos	# QA Pairs	Answer Format	Temporal	Media Type	Source	Key Features
Image-based VQA Datasets								
VQA v1.0	2015	204,721 images	614,163	Open-ended	×	Image	MS COCO	First large-scale VQA dataset; strong language priors
TDIUC	2017	170,000 images	1.6M+	Open-ended	×	Image	MSCOCO	A dataset for diverse, balanced, and detailed visual question answering.
VQA v2.0	2017	265,016 images	1.1M+	Open-ended	×	Image	MS COCO	Balanced with complementary images; reduced language bias
iVQA	2018	200,000 images	1.1M+	Open-ended	×	Image	MS COCO	Inverse VQA; generates questions from images and answers
VQuAD	2022	7,000	1.3M+	Open-ended	×	Image	Synthetic	Video frames with question answering in document format
Video-based VQA Datasets								
MovieQA	2016	408	14,944	Multiple-choice	✓	Video	Movies	Includes subtitles, scripts, and plots; story understanding
TGIF-QA	2017	71,741	165,165	Multiple-choice	✓	GIFs	Tumblr GIFs	Repetition count, action, transition, and frame QA tasks
MovieFIB	2017	128,085	348,998	Fill-in-blank	✓	Video	Movies	Movie clips with fill-in-the-blank task
MarioQA	2017	13 hours	187,757	Multiple-choice	✓	Game video	Mario game	Gaming environment; rule-based automatic generation
MSVD-QA	2017	1,970	50,505	Open-ended	✓	Video	YouTube	Derived from MSVD dataset; 5 question types
MSRVTT-QA	2017	10,000	243,680	Open-ended	✓	Video	YouTube	Larger scale than MSVD-QA; diverse visual content
TVQA	2018	21,793	152,545	Multiple-choice	✓	Video	TV shows	Combines video frames and subtitles; 6 TV shows
PororoQA	2018	16,066	27,328	Open-ended	✓	Animation	Cartoon	Children's cartoon; character-centric questions
LifeQA	2018	275	2,326	Multiple-choice	✓	Video	Daily life	Everyday activities; realistic situations
ActivityNet-QA	2019	5,800	58,000	Open-ended	✓	Video	ActivityNet	Long videos (avg 180s); human activities focus
Social-IQ	2019	1,250	7,500	Multiple-choice	✓	Video	Social videos	Social intelligence; human behavior understanding
CLEVRER	2019	20,000	305,000	Multiple-choice	✓	Synthetic	Rendered physics	Physical reasoning in synthetic scenes; causal understanding
TVQA+	2019	4,200	29,383	Multiple-choice	✓	Video	TV shows	TVQA extension with spatial bounding box annotations
DramaQA	2020	23,928	17,983	Multiple-choice	✓	Video	Korean drama	Four levels of reasoning difficulty; character-centric
KnowIT VQA	2020	12,087	24,282	Multiple-choice	✓	Video	TV sitcom	Requires external knowledge beyond video content
How2QA	2020	22,000	44,007	Multiple-choice	✓	Video	Instructional	Based on instructional videos; multimodal learning
Tutorial VQA	2020	76	6,195	Open-ended	✓	Video	Tutorials	Instructional videos with detailed explanations
V2C-QA	2020	1,500	37,000	Open-ended	✓	Video	MSRVTT	Video-to-commonsense QA; requires world knowledge
NExT-QA	2021	5,440	52,044	Multiple-choice	✓	Video	YouTube	Causal and temporal reasoning; complex questions
AGQA	2021	9,595	192,000	Multiple-choice	✓	Video	Charades	Compositional reasoning; systematically generated QA
SUTD-TrafficQA	2021	10,080	62,535	Multiple-choice	✓	Video	Traffic scenes	Focus on traffic scenarios and accident analysis
ENV-QA	2021	23,261	85,072	Open-ended	✓	Video	Environment	Environmental scenes; domain-specific knowledge
Value	2021	152,600	252,400	Open-ended	✓	Video	Various	Human value understanding; normative reasoning
YouTube2Text-QA	2021	1,987	122,708	Open-ended	✓	Video	YouTube	Based on YouTube2Text corpus; natural language QA
Charades-SRL-QA	2021	9,513	71,735	Open-ended	✓	Video	Charades	Semantic role labeling for action understanding
ASRL-QA	2021	35,805	162,091	Open-ended	✓	Video	ActivityNet	Action semantic role labeling for QA
Pano-AVQA	2022	5,400	51,700	Multiple-choice	✓	360° Video	Panoramic	360-degree videos; audio-visual reasoning
Music-AVQA	2022	9,288	45,867	Multiple-choice	✓	Video+Audio	Music videos	Musical understanding; audio-visual integration
WebVidVQA3M	2022	3M	3M	Open-ended	✓	Video	Web videos	Web-scale pretraining for open-domain video QA
WildQA	2022	369	916	Open-ended	✓	Video	In-the-wild	Uncontrolled environments; practical use cases
HowToVQA69M	2022	69,000	69M	Open-ended	✓	Video	HowTo100M	Large-scale weakly supervised dataset from instructional videos
EgoSchema	2023	250 hours	5,000+	Multiple-choice	✓	Video	Egocentric	First-person perspective; procedural understanding
Video-ChatGPT	2023	100,000	100,000	Open-ended	✓	Video	Various	Instruction-response pairs; conversational format
FIBER	2023	2.9M	11.2M	Multiple-choice	✓	Video	Web videos	Large-scale weakly supervised; benchmark for video reasoning
VideoInstruct100K	2024	100,000	100,000	Open-ended	✓	Video	Various	Diverse instruction-tuning data for video LLMs
STAR	2024	22,000	60,000+	Multiple-choice	✓	Video	Various	Situated reasoning in diverse scenarios

Appendix Table A2: Comparison of Continual Learning Methods for Visual Question Answering.

Method	Innovation	Methodology	Pros	Cons
MAFED	Modality-aware feature distillation	Applies weighted distillation losses on visual and textual modalities	Mitigates modality-specific forgetting; improves stability	Requires careful weight tuning; increased computation
Symbolic Replay	Scene graph as replay prompt	Extracts scene graphs to serve as symbolic replay for past tasks	Low memory/storage; privacy-friendly	Dependent on scene graph quality; may lose fine-grained details
QUAD	Query-based Interpretable Neural Motion Planning for Autonomous Driving	Query occupancy at sparse points for planning	Efficient, interpretable, safer driving	Depends on good trajectory sampling
TRIPLET	Decoupled multi-modal prompt learning	Uses decoupled prompts (across modalities and layers) with prompt interaction strategies	Enhances multi-modal fusion and modality interaction	Complex prompt design and training dynamics
LLM-Assisted Multi-Teacher	Multi-teacher guidance using LLMs	Leverages teacher models to guide incremental learning in surgical VQA	Utilizes strong LLM performance; robust in high-stakes domains	High computational cost; domain-specific tuning
VQACL	Dual-level continual learning setting for VQA	Constructs outer (language-driven) and inner (vision-driven) tasks with SS/SI feature learning	Comprehensive benchmark; tests compositional generalization	Complex task partitioning and setup
One VLM to Keep it Learning	Data-free continual learning via pseudo-data generation	Uses VLM to generate pseudo-data and balances old vs. new knowledge	Eliminates data storage issues; privacy-friendly	Variable pseudo-data quality; delicate balancing mechanism
VLM-Assisted (Self-Driving)	Continual VQA tailored for autonomous driving	Combines VLM, selective memory replay, distillation, and projection layers	Suitable for safety-critical self-driving; effective balance	Domain-specific adjustments needed; scalability challenges
ViLCo-Bench	Continual learning benchmark for video-language tasks	Provides standard protocols and diverse scenarios for evaluation	Facilitates fair comparison; diverse task coverage	Complex benchmark setup; may have limited scope
ColPro	Collaborative prompt optimization	Optimizes prompts collaboratively across tasks	Improves prompt efficiency and adaptation	Requires extensive tuning; details less documented
DAM	Merge adapters dynamically for continual VidQA learning	Efficient, less forgetting, strong domain generalization	Alleviates forgetting via distillation	Relies on good router prediction and adapter quality
MA-LMM	Store past video info in memory for efficient long-term understanding	Handles long videos, low GPU memory, plug-and-play	Needs careful memory compression to avoid redundancy	



Appendix Figure A1: This timeline graph clearly illustrates the evolution of Video VQA datasets from 2015–2024, showing progression from basic image VQA to sophisticated video understanding benchmarks across diverse domains.