Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Memory Retrieval in Visuomotor Policies for Long-Horizon Robot Control

Rutav Shah[1], Yisu Li*, Femi Bello*, Yuke Zhu, Roberto Martín-Martín

The University of Texas at Austin

***Abstract*—General-purpose robots operating in partially observable environments, such as homes, require memory to support autonomy. They must recall diverse information from the past, such as where objects were placed, which tasks a human partner has completed, and when an appliance was turned on, to accomplish a wide range of tasks. Achieving this versatility requires a memory retrieval mechanism that generalizes well across tasks. However, hand-designed or heuristic-based methods rely on task-specific assumptions that may not transfer to different settings. Transformer architectures that use attention over long contexts for memory retrieval provide a promising alternative, as they learn retrieval from data without task-specific assumptions. However, directly incorporating long-context transformer architecture into imitation learning from offline data introduces two key challenges: (1) the policy may learn spurious correlations between the information from the past and predicted actions, and (2) errors accumulate over time in the memory due to prediction inaccuracies and their compounding interactions with the environment, leading to model drift and cascading failures in long-horizon control. To address both challenges, we introduce HALO, a visuomotor policy with an attention-based memory retrieval mechanism for long-horizon control. To suppress spurious correlations, HALO leverages vision-language model (VLM) priors to steer retrieval toward task-relevant information. Concretely, it generates task-relevant, memory-dependent question–answer pairs from demonstration trajectories and trains the policy jointly with a video question-answering objective, transferring VLM priors to the visuomotor policy. Second, to reduce the impact of accumulated errors in memory during closed-loop control, HALO uses sparse attention that restricts retrieval to only the most relevant parts of the history. Together, these components enable more reliable long-horizon control by guiding the policy to retrieve task-relevant information from up to eight minutes of past experience. Project website: https://robin-lab.cs.utexas.edu/HALO**

I. Introduction

General-purpose robots operating in household environments must solve long-horizon manipulation tasks that require recalling information no longer present in current sensory input. The robot may need to remember where an object was placed earlier, when an appliance was turned on, how many pieces of bread were placed in the oven, or which tasks have already been completed (Fig. 1). Such requirements arise frequently in everyday household activities, where successfully completing the task depends on recalling information over extended time horizons. As a result, purely reactive visuomotor policies will fail on these tasks, motivating the need for **memory retrieval** mechanisms that can leverage information gathered earlier [1], [2].

[1] Corresponding author: rutavms@utexas.edu, * Equal contribution

Fig. 1. **Memory retrieval in visuomotor policy learning.** Long-horizon household tasks require robots to act on information no longer present in the current sensory input. Under partial observability, the policy must retrieve relevant information from past observations stored in memory to predict correct low-level actions. The type of information needed can change across task stages, motivating a general, learned retrieval mechanism.

The information that robots must recall from history is diverse. It may include spatial information such as object locations, relational information between objects, numerical quantities, temporal events, or event ordering. Approaches based on hand-designed memory representations or modality-specific retrieval rules, therefore, struggle to scale across the wide range of manipulation tasks required of general-purpose robots [3]–[5]. This motivates the development of a *general* memory retrieval mechanism that can be learned end-to-end, rather than tailored to individual tasks or modalities [6]–[9].

Among candidate mechanisms for general memory retrieval, associative retrieval based on learned query-key similarity, such as attention in transformers, has emerged as a promising mechanism [10], [11]. By learning both queries and keys from data, associative retrieval avoids modality-specific design, and has demonstrated strong empirical performance in domains such as language processing [12]. However, directly applying attention-based memory retrieval to long-horizon robotic imitation learning via offline data exposes two fundamental challenges. First, policies may exploit spurious correlations between past observations and current actions, *i.e.*, attending to information that correlates with expert behavior in the training data but is not causally relevant to task execution. Such correlations often fail to hold at test time, leading to poor performance [13], [14]. Second, in offline imitation learning, small prediction errors compound during closed-loop online execution [15], [16]. Conditioning on large amounts of past

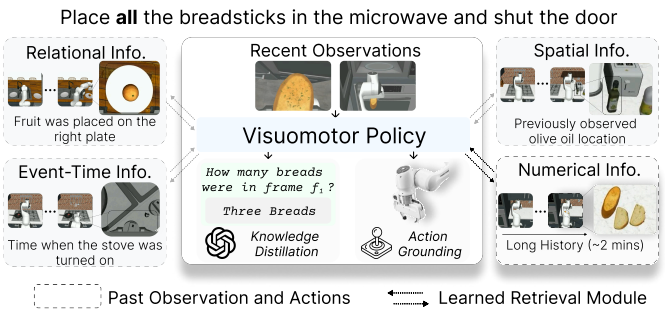

Fig. 2. HALO learns to retrieve diverse forms of task-relevant information from history, guided by priors distilled from vision-language foundation models.

observations can amplify this effect, as the policy repeatedly attends to noisy information, leading to drift and cascading failures over long horizons. Together, these challenges can limit the effective use of memory in imitation learning, a crucial capability for partially observable long-horizon tasks.

To address these challenges, we propose HALO: **H**istory-**A**ware visuomotor policy for **LO**ng-horizon robotic imitation learning. Our key insight is that while attention provides a general retrieval mechanism, it must be guided towards task-relevant information and constrained to avoid conditioning on irrelevant or noisy history (Fig. 2).

To mitigate spurious correlations, we build on the observation that vision-language models (VLMs) encode rich priors about which information from past observations is relevant for a task [17]. Although these priors are not grounded in action and are therefore insufficient on their own for control, they provide useful signals for identifying task-relevant information in history. In HALO, we distill these priors into the policy through automatically generated video question–answer (VQA) supervision. To answer these questions correctly, the retrieval module must attend to specific parts of the history where task-relevant event(s) occurred, providing direct supervision on what should be retrieved from memory. The policy is co-trained to both imitate expert actions and answer questions about past observations. The VQA objective biases memory retrieval towards task-relevant information, whereas the action prediction objective may still access information needed for low-level control. This contrasts with prior work that incorporates VLM priors through text-based summaries [18], which may omit details needed for low-level control.

To address error accumulation over long horizons, HALO further restricts the policy to condition only on a sparse subset of stored information. Rather than attending to the full history, the policy retrieves the top-k most informative observations or actions via learned query-key matching and conditions exclusively in these elements [19]. This explicit sparsification limits the incorporation of noisy information, reducing closed-loop drift and mitigating cascading failures during execution.

We evaluate our approach on REMEMBENCH [20], a benchmark for long-horizon manipulation tasks that require diverse forms of memory, as well as on five real-world tasks across two robotic platforms, including a fixed-base manipulator and a mobile manipulator. Across these settings, we show that VQA-induced task priors provide a general solution, improving absolute task success by 7% on average across diverse tasks and outperforming prior methods based on hand-designed rules (-12%) and task-specific features (-21%). We further find that while vision-language models encode informative priors about task-relevant information, policies using these priors alone reach only 18% absolute success, compared to 41% with action-grounded retrieval, highlighting the importance of grounding retrieval in action prediction. In addition, we demonstrate that top-k attention sparsification improves absolute performance by 10%, reducing model drift, and outperforms alternative memory mechanisms based on memory compression by 29%. Together, these results highlight the importance of learning to retrieve task-relevant information from memory and grounding retrieval in action prediction for long-horizon control.

II. RELATED WORK

A. Imitation Learning under Partial Observability

Standard imitation learning assumes Markovian observations, causing reactive visuomotor policies to fail under partial observability. Belief-state space methods address this issue by explicitly modeling task state but require task-specific representations and update rules [21]. Recent work instead incorporates explicit memory into policy. Some methods rely on object-centric representations to reduce the amount of information, but often limit retrieval to spatial information [5]. Others retrieve past observations using similarity in learned semantic embeddings, leveraging pretrained vision-language models [4], [22], or apply hand-designed rules for memory filtering [23]. Such assumptions restrict their applicability to settings where task-relevant information is well captured by a fixed representation or retrieval rule (here, a semantic one), and may not scale to more diverse long-horizon tasks.

Several approaches learn compact representations of the history to support long-horizon reasoning [24]. While effective, learned compression can discard granular information, such as low-level action sequences taken by the robot. These limitations motivate memory retrieval mechanisms that can flexibly access diverse task-relevant information from history without relying on task-specific designs.

B. Knowledge Distillation from Foundation Models

Vision-language foundation models encode rich knowledge about scenes, objects, and activities from large-scale multimodal data and have been widely used in robotics for perception, planning, affordance learning, and semantic reasoning [25]–[30]. Prior work has also distilled knowledge from foundation models into robotic systems by improving parametric representations [31], [32]. In contrast, relatively little work has explored using foundation models to guide memory retrieval in visuomotor imitation learning, that is, learning to retrieve information from non-parametric memory [33]. Our work builds on these efforts by distilling task-relevant priors

from vision-language models to supervise learned memory retrieval, while grounding the policy through action imitation.

C. Memory Mechanisms for Long-Context Reasoning

Modeling long-range temporal dependencies is a central challenge in sequential decision making. Recurrent and state-space models maintain memory through compressed latent states but struggle with selective recall and credit assignment over long horizons [9], [34], [35]. Transformer-based models address these limitations by using attention to directly access past inputs [11]. However, dense attention over long contexts can cause irrelevant or noisy information to increasingly influence predictions [36], an issue that is especially pronounced in robotics due to prediction errors and compounding errors during closed-loop execution.

To regulate information flow in long-context models, prior work has proposed combining attention with recurrence [37], a gating mechanism [38], and inducing sparsity in the transformer architecture [39]. While these mechanisms have been extensively studied in language and vision, their effectiveness in long-horizon closed-loop visuomotor control has received less attention. In this work, we find that simple Top-K attention sparsification is effective for reducing the influence of accumulated error in model prediction.

III. HALO

A. Problem Formulation and Overview

Problem formulation. In this work, we address *long-horizon robotic imitation learning under partial observability*, where successful task execution requires recalling information no longer present in the current sensory input. At each time step t, the robot receives an observation o_t (e.g., RGB images and proprioception) and executes a low-level action a_t. Let

$$\tau_t = \{(o_1, a_1), \ldots, (o_{t-1}, a_{t-1}), o_t\}$$

denote the interaction history up to time t. Additionally, we are given a natural-language task instruction l (e.g., "put all the bread in the microwave and close the door"), which specifies the high-level goal and is provided to the policy at every time step. We are given a dataset of expert demonstrations

$$\mathcal{D} = \{\tau^{(i)}\}_{i=1}^{N},$$

where each trajectory

$$\tau^{(i)} = \{(o_1^{(i)}, a_1^{(i)}), \ldots, (o_{T_i}^{(i)}, a_{T_i}^{(i)})\}$$

is collected via teleoperating the robot. Our aim is to learn a closed-loop visuomotor policy

$$\pi_\theta(a_t \mid \tau_t, l),$$

that imitates the expert and performs long-horizon tasks by retrieving and leveraging relevant information from history.

Model architecture overview. We parameterize the visuomotor policy $\pi_\theta(a_t \mid \tau_t, l)$ with three main components: (i) modality-specific encoders consisting of an observation encoder g_θ^{obs} and an action encoder g_θ^{act}, which map heterogeneous inputs (e.g., RGB images, proprioception, and past actions) into a common latent space, (ii) a policy backbone f_θ that performs memory retrieval over the encoded history and processes the information, and (iii) a task-specific head h_θ^{act} and h_θ^{text} that predict low-level control commands and text answers, respectively.

At time step t, the current observation is encoded as $x_t = g_\theta^{\text{obs}}(o_t)$ and past actions as $e_i = g_\theta^{\text{act}}(a_i)$ for $i < t$. The encoded history is

$$\mathcal{M}_t = \{(x_1, e_1), \ldots, (x_{t-1}, e_{t-1})\}.$$

Given $\mathcal{M}_t$, the current embedding x_t, and the task instruction l, the policy backbone f_θ produces a latent state

$$z_t = f_\theta(\mathcal{M}_t, x_t, l),$$

This latent state is passed to two prediction heads: an action head $h_\theta^{\text{act}}(z_t)$ for control and a text prediction head $h_\theta^{\text{text}}(z_t)$ for auxiliary supervision.

Method overview. A common approach for retrieving information from history is to equip the policy backbone (f_θ) with an *associative memory retrieval mechanism*, typically implemented via learned query–key similarity (*i.e.*, attention) [11]. Such mechanisms provide a general, modality-agnostic way to retrieve information from history, avoiding task-specific designs.

Concretely, the policy backbone f_θ implements retrieval using a query–key–value formulation. At each time step t, a query vector is computed from the current context,

$$q_t = f_q(x_t, l),$$

while each memory element $m_i \in \mathcal{M}_t$ is mapped to a key–value pair,

$$k_i = f_k(m_i), \qquad v_i = f_v(m_i),$$

where f_q, f_k, and f_v are learned functions parameterized by $\theta_q \subset \theta$, $\theta_k \subset \theta$, and $\theta_v \subset \theta$, respectively. Retrieval is performed by computing similarities between q_t and $\{k_i\}$ and aggregating the corresponding values $\{v_i\}$. The retrieved features are integrated with the current representation and processed by subsequent model layers.

However, as explained before, directly applying such an attention-based retrieval in long-horizon imitation learning introduces two *key limitations*. First, because attention aggregates information from all stored history $\mathcal{M}_t$, the policy may attend to task-irrelevant details and incorporate them into decision-making, leading to spurious correlations between past observations and actions. Second, during closed-loop execution, small prediction errors, compounded by environmental interactions, accumulate over time. These errors introduce noise into the stored representations, which can degrade latent representation quality, leading to model drift and cascading failures over long horizons.

HALO addresses these limitations with two components (Fig. 3). First, we distill task-relevant priors from vision-language models (VLMs) into the policy using automatically

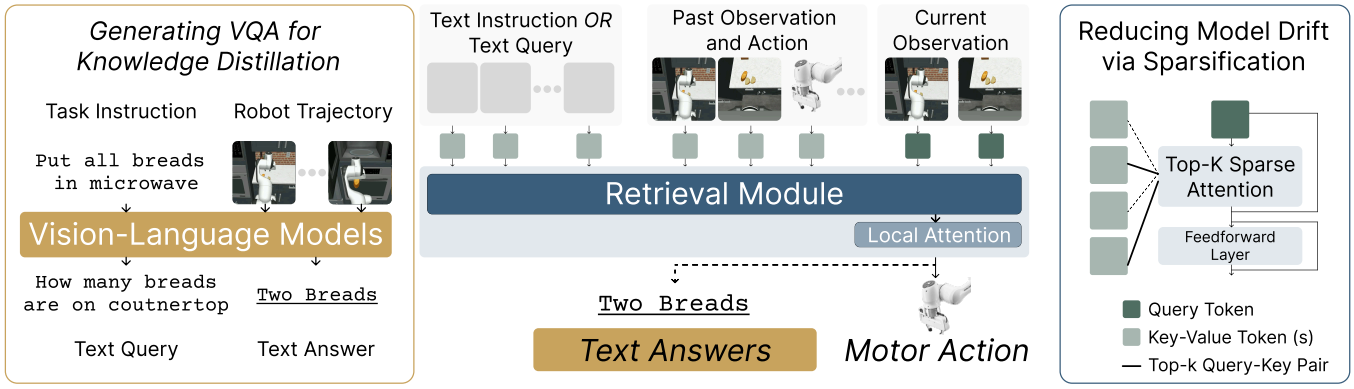

Fig. 3. **Overview.** HALO learns a visuomotor policy that retrieves information from the past observations and actions to predict low-level robot actions (middle), guided by priors from vision-language models to retrieve task-relevant information through VQA (left) and uses sparse attention to reduce model drift (right).

generated video question–answer (VQA) supervision. This supervision encourages the retrieval mechanism to focus on task-relevant information, reducing reliance on irrelevant history (Sec. III-B, III-C). Second, we restrict retrieval to the top-k most relevant memory entries at each time step. This sparsification limits the influence of accumulated error during closed-loop execution, leading to robust long-horizon control (Sec . III-D).

Concretely, HALO trains a single visuomotor policy that jointly (i) imitates expert actions and (ii) answers VQA queries about the trajectory history. Both objectives share the same modality-specific encoders g_θ and a common backbone f_θ that contains the memory retrieval mechanism, differing only in their prediction heads. Top-k sparsification is applied to retrieval for both objectives, encouraging the policy to condition only on a subset of relevant memory entries. This shared training and policy weights encourages the retrieval mechanism to learn representations that are useful for both answering questions about the past and predicting actions. As a result, supervision from VQA transfers to action prediction, guiding the policy to retrieve information relevant for control.

B. Distilling Vision-Language Model Priors via Video Question–Answering

Vision-language models encode rich priors about scenes, objects, activities, and semantics to understand task instructions, making them a useful source of guidance for identifying task-relevant information in trajectories. However, because VLMs are not grounded in action prediction, these priors alone do not capture all the information required for effective low-level control. HALO therefore distills priors from VLMs into the policy while grounding them through imitation learning.

Video question–answering as retrieval supervision. In standard end-to-end imitation learning, the retrieval mechanism is trained solely through the action prediction objective. HALO additionally supervises the retrieval process using the video question–answering objective. Specifically, we generate VQA pairs (u, v), where u is a text-based question probing task-relevant information from the trajectory history (e.g., object locations, object relations, event ordering, or subgoal progress), and v is the corresponding answer. For VQA supervision, the policy backbone is conditioned on the encoded history $\mathcal{M}_t$, the current observation embedding x_t, and the question u, and the answer is predicted via a VQA head:

$$\hat{v} \sim \pi_\theta^{\text{vqa}}(v \mid \mathcal{M}_t, x_t, u), \tag{1}$$

where π_θ^{vqa} consists of the shared observation g_θ^{obs}, action encoder g_θ^{act}, policy backbone f_θ (including the retrieval mechanism) with the action prediction objective, and a text prediction head h_θ^{text}.

To answer u, the policy must retrieve the specific information from the history that supports the answer. Since the questions are generated using vision-language models and tailored to the task, this objective injects task-relevant priors into the retrieval mechanism, providing direct supervision on what information should be retrieved.

Joint grounding through imitation learning. While VQA supervision guides retrieval toward task-relevant information, it does not by itself ensure that the retrieved information is sufficient for low-level control. We therefore jointly train the policy using an imitation learning objective:

$$\mathcal{L}_{\text{IL}}(\theta) = \mathbb{E}_{(\tau_t, s, a_t^*)\sim\mathcal{D}} \left[-\log \pi_\theta^{\text{act}}(a_t^* \mid \tau_t, l) \right], \tag{2}$$

where a_t^* denotes the expert action. We also define a VQA loss over the generated dataset $\widetilde{\mathcal{D}}$:

$$\mathcal{L}_{\text{VQA}}(\theta) = \mathbb{E}_{(\tau_t, s, u, v)\sim\widetilde{\mathcal{D}}} \left[-\log \pi_\theta^{\text{vqa}}(v \mid \tau_t, u) \right]. \tag{3}$$

The final training objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{IL}}(\theta) + \lambda \mathcal{L}_{\text{VQA}}(\theta), \tag{4}$$

where λ balances imitation and VQA supervision. This joint training ensures that VQA provides informative priors about what information from the history should be retrieved, while imitation learning grounds retrieval in expert action prediction.

C. Generating VQA Data from Long-Horizon Trajectories

Generating accurate VQA supervision for long-horizon robot trajectories is challenging. Single VLM calls over high-dimensional visual information often produce incorrect or

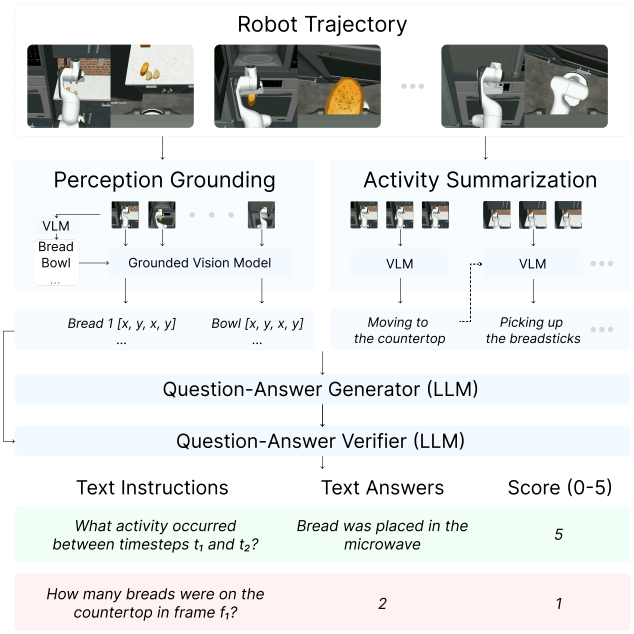

Fig. 4. **Overview of video question-answer generation for knowledge distillation.** Images are first converted to text using a grounded vision model, and trajectories are summarized with each subtask and corresponding frame numbers. An LLM generates task-relevant question-answer pairs from these summaries and the task description, which are then verified for correctness and relevance, filtering out about 20% of the data.

hallucinated question-answer pairs, likely because state-of-the-art VLMs struggle to process long-context visual information coherently. To address this, we employ a multi-stage generation pipeline that produces high-quality supervision (Fig. 4).

(i) Trajectory-to-text summarization. Given a trajectory segment, we construct a textual description that captures scene elements (e.g., object identities and locations) and activity descriptions (e.g., "the robot opened the drawer and placed the mug inside"). To do so, we combine information from different specialized vision models, such as object detectors for spatial cues and VLMs for higher-level activity descriptions, and consolidate them into a single text representation of the trajectory. This consolidation is performed only to generate questions and answers that provide retrieval priors; the information itself may not be sufficient to capture all the information required by the policy to predict low-level actions.

(ii) Question and answer synthesis. From the consolidated text, we synthesize questions that probe memory requirements relevant to the task. We observe that language models tend to bias questions toward task-relevant information present in the initial portion of a trajectory. To mitigate this bias, we explicitly prompt the language model to generate a question corresponding to a specific frame index, randomly sampled from the trajectory. The model is also allowed to return "N/A" if no task-relevant information is present at the specified frame. We ask the language model to also generate the answer corresponding to the question.

(iii) Automatic filtering for correctness and relevance. Generated VQA pairs may contain incorrect or be weakly task-relevant or labeled as "N/A". To improve data quality, we use an additional language model to rate each pair on correctness with respect to the trajectory description and relevance to the task. Pairs with low scores are filtered out. The remaining set forms $\widetilde{\mathcal{D}}$, which is used for training.

D. Reducing Model Drift with Top-K Attention Sparsification

Even with improved retrieval supervision, long-horizon closed-loop execution remains susceptible to error accumulation. In offline imitation learning, policies are trained on expert trajectories but must execute actions based on their own predictions at test time. Small prediction errors can lead to future observations that deviate from the expected observations, leading to compounding errors over time. Conditioning on large amounts of information from memory can further exacerbate this effect by repeatedly incorporating noisy information, resulting in cascading failures over long horizons.

To mitigate this issue, HALO explicitly limits how much information from the history the policy can condition on at each time step. Rather than attending to all available information, we first identify the information most relevant to the current decision and restrict retrieval to only these elements. This allows the policy to focus on important information while reducing the effect of accumulated error in memory.

Concretely, let $\{m_i\}_{i=1}^{|\mathcal{M}_t|}$ denote candidate information extracted from the history, with keys $k_i = f_k(m_i)$ and values $v_i = f_v(m_i)$. Given a query q_t, standard attention computes scores $s_i = \langle q_t, k_i \rangle$ and forms a weighted sum over all values. In top-k sparsification, we instead select

$$\mathcal{I}_k = \mathrm{TopK}(\{s_i\}_{i=1}^{|\mathcal{M}_t|}, k), \tag{5}$$

corresponding to the k most relevant elements, and compute attention only over this subset with all elements outside $\mathcal{I}_k$ receive zero weight:

$$c_t = \sum_{i \in \mathcal{I}_k} \frac{\exp(s_i)}{\sum_{j \in \mathcal{I}_k} \exp(s_j)} v_i. \tag{6}$$

Since the top-k operation is non-differentiable, we employ a straight-through estimator during training. Intuitively, at each training step, the policy retrieves the top-k elements based on the query–key similarity using the current model weights. If the retrieved information improves action prediction or VQA accuracy, gradients flowing through the attention output increase the corresponding query–key similarities, reinforcing the selection of useful information. Conversely, elements that do not contribute meaningfully receive weaker updates and are less likely to be selected in future steps.

To summarize, HALO combines VQA-based distillation of informative VLM priors, which encourages task-relevant retrieval and reduces spurious correlations, with top-k sparse attention, which limits error accumulation and mitigates model drift, yielding a policy that is effective for long-horizon, partially observable tasks.

IV. Experiments

A. Evaluation Tasks.

We evaluate our approach on long-horizon manipulation tasks that require retrieving diverse information types—spatial,

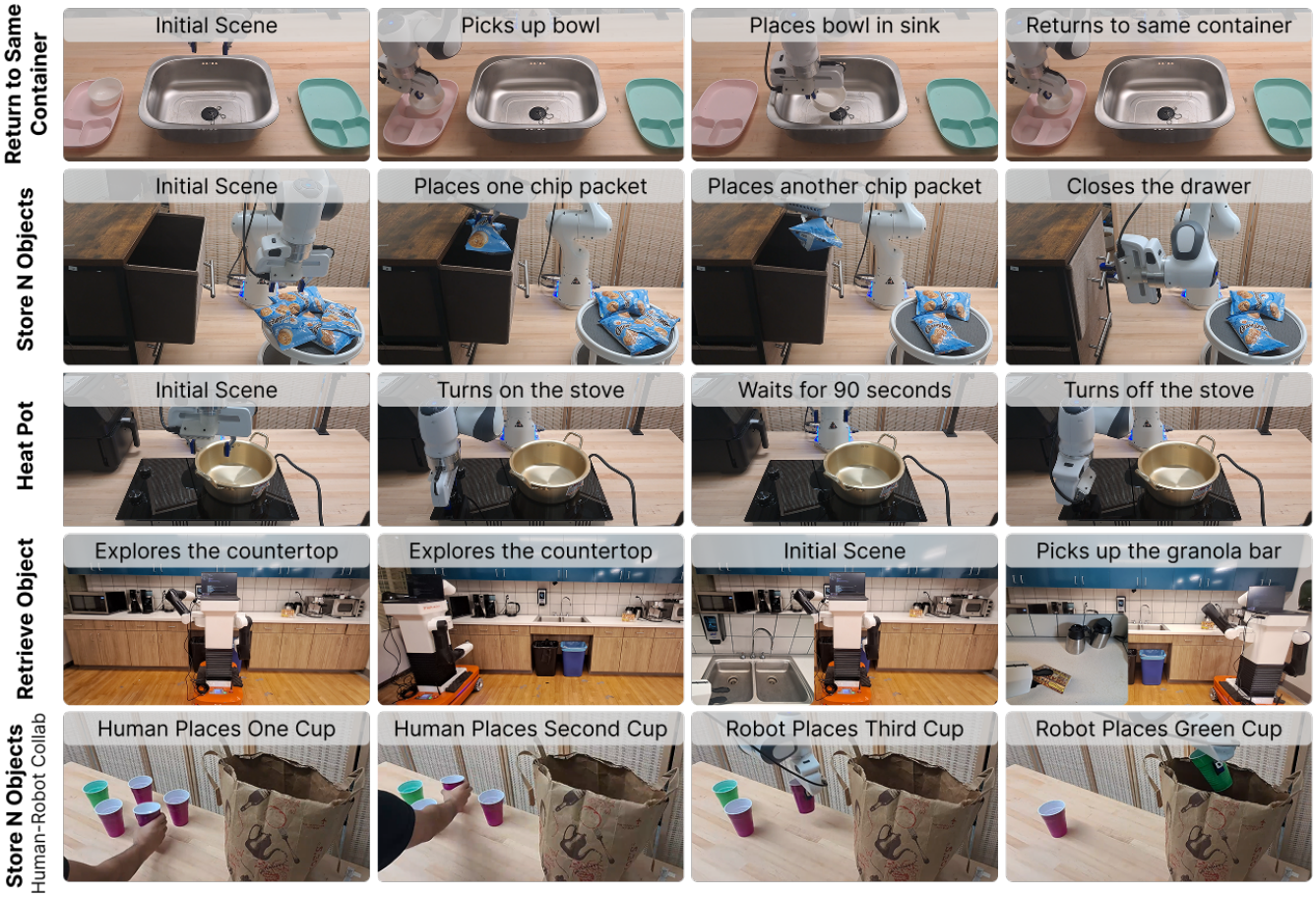

Fig. 5. **Visualization of real-world tasks.** We evaluate HALO in four stationary and mobile manipulation tasks, including a human-robot collaborative task (*rows*) with strong partial observability, where the agent needs to retrieve different types of information from memory. The pictures in each column depict a different phase of the task.

relational, numerical, and event-time—from past interactions, including human-robot collaborative scenarios. These tasks span five scenarios reflecting diverse memory demands often faced in household settings (Fig. 5):

Retrieve Object (*Spatial information*). This task requires recalling object locations observed earlier in the episode; the policy must fetch a granola bar from a location previously observed on the kitchen countertop, which is no longer visible at the time of action. Across episodes, the object's initial location varies between the left and right sides of the countertop.

Return to Same Container (*Object relations*). This task requires recalling earlier relations among objects; after placing a bowl in a sink, the policy must later retrieve the plate on which the bowl was originally placed. Across episodes, the involved objects and their relations vary.

Store N objects (*Numerical information*). This task requires recalling a numerical count over time; the robot should place n chip packets in a drawer before closing it, with the packets becoming occluded once placed. Across episodes, both the initial number of chip packets and the target count n vary.

Heat Stove (*Event-time information*). This task requires recalling when an event occurred to act after a specified duration; the robot must heat a stove for a fixed duration. Across episodes, the required heating duration varies between 4, 6, and 8 minutes.

Store N objects (*Spatial information, Human-robot collaboration*). This task requires tracking a numerical count in a collaborative setting: a human places some purple cups first, then the robot completes the total of three purple cups in the shopping bag, and then adds a green cup to it. Across episodes, the number of initial purple cups in the scene ranges from 3 to 4, and the human places 1 to 3 cups.

B. Baselines.

With the goal of reducing memory size or introducing inductive priors about what should be stored, prior work proposes several alternatives:

SAM2Act++ [5] uses object-centric memory by storing object center pixel coordinates. This induces a strong spatial prior and reduces memory size by storing only object locations.

ReMemBer [18] stores text summaries of past observations generated by a VLM. This introduces semantic priors while compressing history into text.

Hand-designed Features uses human-designed rules to select which task-relevant features to store or discard, similar to prior work MemER [23]. It encodes human priors about what information is important while keeping memory compact.

TABLE I
EVALUATION ON SIMULATION BENCHMARK WITH 50 ROLLOUTS PER TASK.

Method	**Retrieve Object** *(Spatial Info.)*	**Return to Same Container** *(Relational Info.)*	**Store N Objects** *(Numerical Info.)*	**Heat Stove for T mins.** *(Event-time Info.)*	**Average**
Standard Transformer	0.26	0.23	0.12	0.27	0.22
Scene Memory Transformer	0.53	0.25	0.17	**0.40**	0.34
SAM2Act++	0.39	0.11	**0.27**	0.04	0.20
ReMemBer	0.36	0.13	0.21	0.00	0.18
Hand-Designed Features	**0.68**	0.15	0.21	0.13	0.29
Token Merging	0.29	0.24	0.25	0.38	0.29
HALO w/o VQA	0.42	0.29	0.17	0.37	0.31
HALO	0.64	**0.32**	0.26	**0.40**	**0.41**

TABLE II
EVALUATION ON REAL-WORLD TASKS WITH 20 ROLLOUTS PER TASK

Method	**Retrieve Obj.** *(Spatial Info.)*	**Return to Container** *(Relational Info.)*	**Store N Objs.** *(Numerical Info.)*	**Heat Stove for T mins.** *(Event-time Info.)*	**Store N Objs.** Human-Robot *Spatial Info.*	**Average**
Standard Transformer	0.40	0.30	0.20	0.40	0.50	0.36
HALO	**0.55**	**0.40**	**0.55**	**0.60**	**0.65**	**0.55**

Scene Memory Transformer (SMT) [24] compresses the entire history into a single scene-level embedding via attention. This reduces memory footprint but forces all past information into a fixed-size representation.

Token Merging [40] compresses the history by merging tokens with similar embeddings that are temporally adjacent to maintain a fixed budget of memory.

C. Results

How well does HALO perform compared to prior visuomotor imitation learning methods that incorporate alternative inductive priors? (Table I)

Comparison to object-centric memory. Compared to SAM2Act++, HALO improves average task success by 21% points. SAM2Act++ performs well when tasks align with its object-centric prior, such as counting objects, but struggles with tasks that require non-spatial memory, particularly recalling event-time information. This highlights the limitation of relying on a single type of prior.

Comparison to text-based summarization. HALO outperforms ReMemBer by 23% points on average. Text summaries generated by VLM often capture coarse semantic information, but may omit details needed for control. Textual summaries may capture object locations but omit action information for navigating to the object, explaining poor 'Retrieve Object' performance. It suggests that while VLM priors are useful, text-based compression alone is insufficient for control.

Comparison to hand-designed rule-based filtering. Compared to hand-designed features, HALO achieves an absolute improvement of 12%. While this performs on par with HALO on spatial tasks, it remains less competitive than HALO on other tasks. This suggests that learning what to retrieve directly from data using HALO not only removes manually designed task-specific priors but also improves performance, possibly by enabling the policy to exploit additional information for decision-making (App. Sec. VII-A).

Comparison to compressed memory representations. We compare HALO against two approaches that compress history: Scene Memory Transformer (SMT), which uses a learned pooled attention layer to compress history into a single embedding, and Token Merging, which merges nearby tokens to maintain a fixed memory budget. HALO outperforms SMT by 7% and Token Merging by 12% points. While compression reduces redundancies and noise in memory, it can lose granular information necessary for low-level action prediction [41]. In contrast, HALO stores all information in memory, enabling access to fine-grained details and performing selective retrieval to reduce the impact of noise.

Overall, we observe that no single method wins across all tasks. However, HALO remains competitive without task-specific assumptions or hand-designed rules, making it versatile with less engineering effort across information types. This versatility is reflected in average performance across tasks, alongside competitive performance on individual ones.

How effective is HALO in real-world tasks? (Table II)

We observe a similar trend in real-world settings, where HALO consistently outperforms the standard Transformer baseline by 19%. Notably, this improvement persists on challenging tasks, including storing multiple objects, tasks completed by the human partner, and long-horizon tasks like heating a pot for up to 8 minutes, which demand diverse types of information and long-context memory.

In addition, we measure manipulation and memory failures in real-world evaluations, finding that HALO reduces them by 8% and 25% absolute over full attention in the 'Retrieve Object' task, respectively. These results support our hypothesis that HALO reduces model drift (fewer manipulation failures)

TABLE III
EVALUATION ON SIMULATION WITH 50 ROLLOUTS PER TASK.

Method	Retrieve Object	Return to Container
LSTM	0.14	0.12
Mamba	0.20	0.18
TransformerXL	0.12	0.20
Window Attention	0.13	0.16
Strided Attention	0.20	0.28
Hierarchical Attention	0.28	0.22
Gated Attention	0.45	0.28
HALO	**0.64**	**0.32**

and improves memory retrieval (fewer memory failures).

What are the effects of VLM-induced priors and Top-K sparsification on performance? (Table I)

Effect of VLM-induced priors. We compare HALO against a variant trained without VQA supervision. Adding VLM-induced priors improves average success by 10%, suggesting that VQA supervision helps the policy identify relevant history, particularly in tasks that require recalling past object placements or numerical information. Without it, the retrieval mechanism may often attend to irrelevant history, leading to poor performance at test time.

Effect of Top-K sparsification. We compare Top-K retrieval to full attention over memory. Top-K improves performance by 9%. Restricting retrieval reduces the amount of irrelevant or noisy information incorporated into decision-making, which is particularly important in long-horizon tasks where only a few past events are relevant to the current action.

See App. Sec. VII-C and Sec. VII-D for more analysis.

How does HALO compare to alternative sparsification and memory mechanisms?

Comparison to alternative sparsification methods. We compare Top-K sparsification to window attention [19], strided attention [19], hierarchical attention [42], and gated attention [38]. HALO outperforms the best alternative, gated attention, by 11%. The sparsification methods relying on fixed access patterns may miss task-relevant events occurring at unpredictable times. In contrast, Top-K selects memory entries based on content relevance, allowing the policy to retrieve information tied to the task rather than position.

Comparison to non-associative memory mechanisms. We compare against models that compress history into latent states, including LSTM [9] (-20%), Mamba [34] (-14%), and Transformer-XL [37] (-12%). While compressed-state models summarize past information efficiently, they may discard details needed for recalling specific past events. Selective retrieval from explicit memory allows the policy to access such details when required.

What are the effects of different design choices in HALO?

Number of retrieved entries (k). We ablate the number of retrieved history entries (k) used by the attention mechanism. A moderate value ($k = 8$) achieves the best performance (52% success). Smaller k (*e.g.*, $k = 4$, 40% success) omits useful context, while larger k incorporates irrelevant history (48% for $k = 12$, 33% for $k = 16$, 44% for $k = 24$). We fix $k = 8$ across all tasks. Developing adaptive strategies that retrieve only the necessary amount of information at each step is a promising direction for future work.

Co-training vs. pretraining. We compare joint co-training of VQA and action prediction against a two-stage approach (VQA pretraining followed by action finetuning). Co-training VQA and action prediction achieves 64% success, outperforming pretrain-then-finetune (44%) and no-VQA training (42%) by 20 and 22 points, respectively. The minimal gain from separate pretrain-then-finetune compared to no-VQA suggests that VQA knowledge is lost during fine-tuning, whereas co-training effectively shapes retrieval toward task-relevant information.

Stage-wise VQA generation vs. single video prompt generation. We compare HALO's stage-wise pipeline against a single-prompt baseline that generates QA pairs directly from long videos. HALO achieves $\mathbf{60}\%$ points more accurate QAs and $\mathbf{20}\%$ points fewer hallucinations. The results suggest SOTA VLMs struggle with direct long-video reasoning for question-answer generation, highlighting the importance of the proposed stage-wise pipeline for generating high-quality VQA (Examples in App. Sec. VII-B).

V. CONCLUSION AND FUTURE WORK

In this work, we study imitation learning in partially observable domains where retrieving information from past observations and actions is essential. We present HALO, a method for training a visuomotor policy with a memory-retrieval mechanism for long-horizon imitation learning. To address spurious correlations learned during training, HALO co-trains action prediction with a video question-answering objective using VLM-generated pairs that provide priors on task-relevant information. To mitigate error accumulation from large memory, HALO employs a top-k sparsification strategy that restricts retrieval to the most relevant parts of the history. Through comprehensive evaluation in simulation and real-world experiments across two robotic platforms, we demonstrate that our approach, which stores all information in memory and selectively retrieves relevant content, achieves significant improvements over alternative methods that reduce memory size through task-specific priors or learned compression. We also find that VLM-induced priors are useful for selecting task-relevant information from memory, but, by themselves, can be insufficient for low-level control. In the future, we plan to investigate mechanisms to switch from fixed to active querying, adapting to each task. Moreover, we will explore broadening the scope of querying to cross-episodic information acquired by the robot in other runs. Additionally, because retrieval from large memory adds retrieval latency, especially as we expand the context to longer time periods, we will explore speculative retrieval techniques to hide it.

VI. Acknowledgments

We thank Huihan Liu, Dvij Kalaria, and Yifeng Zhu for providing valuable feedback on the manuscript. We also thank all members of the Robot Perception and Learning Lab and the Robot Interactive Intelligence Lab at UT Austin for their insightful discussions. This work was partially supported by the National Science Foundation (FRR-2145283, EFRI-2318065), the Office of Naval Research (N0001424-1-2550), the DARPA TIAMAT program (HR0011-24-90428), and the Army Research Lab (W911NF-25-1-0065). It was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS2024-00457882, National AI Research Lab Project).

References

[1] E. Tulving *et al.*, "Episodic and semantic memory," *Organization of memory*, vol. 1, no. 381-403, p. 1, 1972.

[2] R. C. Atkinson and R. M. Shiffrin, "Human memory: A proposed system and its control processes," in *Psychology of learning and motivation*. Elsevier, 1968, vol. 2, pp. 89–195.

[3] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell, "Open-vocabulary object retrieval." in *Robotics: science and systems*, vol. 2, no. 5, 2014, p. 6.

[4] V. S. Dorbala, G. Sigurdsson, R. Piramuthu, J. Thomason, and G. S. Sukhatme, "Clip-nav: Using clip for zero-shot vision-and-language navigation," *arXiv preprint arXiv:2211.16649*, 2022.

[5] H. Fang, M. Grotz, W. Pumacay, Y. R. Wang, D. Fox, R. Krishna, and J. Duan, "Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation," *arXiv preprint arXiv:2501.18564*, 2025.

[6] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," *Advances in neural information processing systems*, vol. 28, 2015.

[7] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[8] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou *et al.*, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] G. A. Carpenter, "Neural network models for pattern recognition and associative memory," *Neural networks*, vol. 2, no. 4, pp. 243–257, 1989.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and machines*, vol. 30, no. 4, pp. 681–694, 2020.

[13] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in neural information processing systems*, vol. 32, 2019.

[14] C. Wen, J. Lin, T. Darrell, D. Jayaraman, and Y. Gao, "Fighting copycat agents in behavioral cloning from observation histories," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2564–2575, 2020.

[15] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[16] J. Spencer, S. Choudhury, A. Venkatraman, B. Ziebart, and J. A. Bagnell, "Feedback in imitation learning: The three regimes of covariate shift," *arXiv preprint arXiv:2102.02872*, 2021.

[17] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[18] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, "Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 2838–2845.

[19] R. Child, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[20] Anonymous, "Scaling short-term memory of visuomotor policies for long-horizon tasks," 2026. [Online]. Available: https://openreview.net/forum?id=5SMNtmJFGa

[21] K. J. Åström, "Optimal control of markov processes with incomplete state information i," *Journal of mathematical analysis and applications*, vol. 10, pp. 174–205, 1965.

[22] P. Liu, Y. Orru, J. Vakil, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint arXiv:2401.12202*, 2024.

[23] A. Sridhar, J. Pan, S. Sharma, and C. Finn, "Memer: Scaling up memory for robot control via experience retrieval," 2025. [Online]. Available: https://arxiv.org/abs/2510.20328

[24] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 538–547.

[25] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*. PMLR, 2022, pp. 894–906.

[26] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2998–3009.

[27] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[28] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu *et al.*, "Rt-trajectory: Robotic task generalization via hindsight trajectory sketches," *arXiv preprint arXiv:2311.01977*, 2023.

[29] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao, "Rt-affordance: Affordances are versatile intermediate representations for robot manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8249–8257.

[30] R. Shah, A. Yu, Y. Zhu, Y. Zhu, and R. Martín-Martín, "Bumble: Unifying reasoning and acting with vision-language models for building-wide mobile manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 13 337–13 345.

[31] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, Y. Peng, C. Shen, F. Feng *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 5377–5395.

[32] D. Driess, J. T. Springenberg, B. Ichter, L. Yu, A. Li-Bell, K. Pertsch, A. Z. Ren, H. Walke, Q. Vuong, L. X. Shi *et al.*, "Knowledge insulating vision-language-action models: Train fast, run fast, generalize better," *arXiv preprint arXiv:2505.23705*, 2025.

[33] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[34] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[35] Y. Bengio and P. Frasconi, "Credit assignment through time: Alternatives to backpropagation," *Advances in neural information processing systems*, vol. 6, 1993.

[36] K. Hong, A. Troynikov, and J. Huber, "Context rot: How increasing input tokens impacts llm performance," Chroma Research, Tech. Rep., July 2025, technical Report. [Online]. Available: https://research.trychroma.com/context-rot

[37] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[38] Z. Qiu, Z. Wang, B. Zheng, Z. Huang, K. Wen, S. Yang, R. Men, L. Yu, F. Huang, S. Huang, D. Liu, J. Zhou, and J. Lin, "Gated attention for

large language models: Non-linearity, sparsity, and attention-sink-free," 2025. [Online]. Available: https://arxiv.org/abs/2505.06708

[39] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17 283–17 297, 2020.

[40] H. Shi, B. Xie, Y. Liu, L. Sun, F. Liu, T. Wang, E. Zhou, H. Fan, X. Zhang, and G. Huang, "Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation," *arXiv preprint arXiv:2508.19236*, 2025.

[41] K. Shao, K. Tao, K. Zhang, S. Feng, M. Cai, Y. Shang, H. You, C. Qin, Y. Sui, and H. Wang, "When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios," *arXiv preprint arXiv:2507.20198*, 2025.

[42] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[44] Y. Guo, S. Sun, S. Ma, K. Zheng, X. Bao, S. Ma, W. Zou, and Y. Zheng, "Crossmae: Cross-modality masked autoencoders for region-aware audio-visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 721–26 731.

[45] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[46] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[47] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.