PlaceTally: A Test-Time Memory Model for Constructing Place Graphs

Alibay Osmanli Shaogang Gong
Queen Mary University of London
{a.osmanli, s.gong}@qmul.ac.uk

Code, data and supplementary material: https://alibayosmanli.github.io/placetally

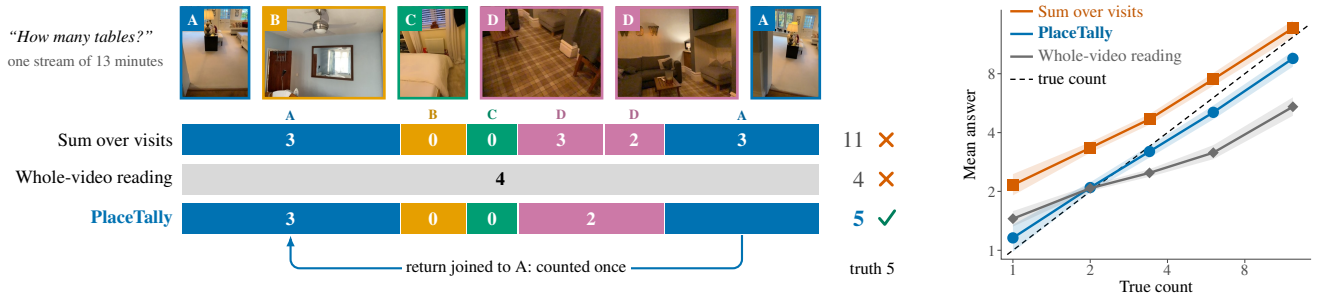


Figure 1. Why do video models miscount when the camera returns? Reading the whole video gives a few small answers, adding up visits counts each return again, and PlaceTally counts each place once. Left: one ARKit-Revisit stream of 13 minutes with six recordings of four rooms, where room D is recorded twice in a row and room A returns at the end; numbers are the same model's counts. Right: mean answer against binned true count on all 460 test questions, with 95% bands. Every question here has an object seen on a return (Sec. 4).

Abstract

Counting distinct objects in a video is hard when the camera returns to places it has seen. A vision-language model that reads the whole video answers with a few small numbers, and one that reads it in parts counts an object again in every part that shows it. Object trackers need the category before the video is read, or a fixed vocabulary. We observe that a memory for counting does not need to know what it is looking for, only where it has been: objects rarely leave a room between visits, so parts of a video grouped by place share almost no objects, and a place can be recognised before any question is asked. Building on this, we introduce PlaceTally, a question-blind test-time memory model that constructs a place graph for each video. Each video fragment joins the place of its best earlier match or opens a new place, because a strong match is rarely wrong while a true return often matches weakly. A frozen vision-language model counts each place once from at most 128 of its frames, so every read has a bounded input and the memory grows with the number of places; the graph is built 15 times faster than real time. On RevisitCount, our benchmark of real room re-recordings, it scores 62.5% mean relative accuracy on test rooms and 61.2% on held-out rooms, against 51.2% and 44.3% for the same model reading the whole video. It stays level as rooms return more or less often, while a fixed correction of summed visits moves from 38.5% to 61.4%.

1. Introduction

Object counting [4, 60] estimates how many instances of a category appear in an image; in video, it must also decide which observations show the same object [2, 33]. This decision is hardest when the camera returns to a place it has already seen, as an assistant or a robot does [42, 76], because every return shows the same objects again. The question also comes later, so the memory has to be constructed before it knows what will be counted.

Existing methods fail for two opposing reasons (Fig. 1): reading the whole video at once under-counts, and adding up parts over-counts. A vision-language model asked about a whole video answers with a few small numbers, a weakness already reported on images [56] and on synthetic videos [13]: on our benchmark, 87% of its answers are 1, 2 or 3, and for nine or more objects it answers 5.4 on average. Reading in parts keeps each count small, but adding up counts an object again in every part that shows it: the counting mode of Cambrian-S [77] works by adding up over surprise-

segmented pieces, which fails when a video repeats [70], and on our videos predicts a median of 8.6 times the true count. Methods that track objects need the objects first. CountVid [2] needs the category known before the video is processed and tracks it for every question, StaMina [45] registers its targets when the stream starts, and VideoAgent [24] and DirectMe [73] keep objects from a detector vocabulary.

The core problem is that a model can count only a few objects at a time; a long video must be counted in parts, and parts can be added up only if they share no objects. How much they share depends on which rooms this camera came back to and must be worked out for each video. Finding parts that share no objects appears to need the objects themselves: on images, LVLM-Count [56] first segments the target so that no object is cut. We observe that in video it does not. Objects rarely leave a room between visits, so parts grouped by place share almost no objects, and "have I seen this object?" becomes "have I been here before?". This question names no category. Answering it is a computational problem of its own: the video must be partitioned online into an unknown number of places, with no depth or camera path, when two views of one room can share no visible surface and different rooms look alike. Each decision is final: a split counts a room's objects twice, and a wrong merge asks the model to count two rooms as one.

To overcome these problems, our idea is to introduce a test-time memory model that is built for each test video instance: a place graph, constructed before any question is asked. As the category is discovered only afterwards, the same memory must serve any category (open-vocabulary); we call this question-blind memory. It does not need to know what it is looking for, only where it has been. It is built per video and not accumulated over a dataset, because only this video's returns decide what can be added up, and rooms in other videos add look-alikes; it is cheap enough to rebuild for every video. Our main contributions are:

- We show that counting across returns can be posed as partitioning a video into places. Earlier work groups video by place to map a kitchen or a building [35, 51], or sums counts over image parts found by segmenting the target [56]; with the place as the unit of counting identity, no category, vocabulary or per-question tracking is needed. With the same frozen model, this lifts mean relative accuracy (MRA [76]) from 51.2% to 62.5% on real returns, within 2.6% of the score with true visits, from 44.3% to 61.2% on held-out rooms evaluated once, and from 10.5% to 56.3% at nine or more objects, where whole-video reading is weakest.
- We introduce a test-time memory model of where the camera has been: a place graph constructed per video before any question. It must recognise a place online with no category, depth or camera path, although one room can look different and two rooms alike. We find the appearance evidence one-sided (a strong match is rarely wrong, a true return often matches weakly), so one decision per fragment on its strongest link is enough, about nine on an 11-minute video, while summing all links over-splits places. With at most K frames per place, K sets the memory, each read's input and the answer time, and the graph is built 15 times faster than real time (Sec. 5.5).
- We introduce RevisitCount. The eight benchmarks we inspected have videos of up to two hours, up to 10,000 questions and, in five cases, a blind check (Table 2), but none reports a return to a room as a separate recording, so adding up parts is not penalised and a model can score without recognising a place: at 60 minutes, a prior from the video's length reaches 80.8% on VSI-SUPER-Count but 17.5% on ours. RevisitCount records every return separately, counts each object once and reports blind baselines on every split; one split of 248 rooms was held out and evaluated once.

Method	Identity of	Before question	Named in advance	Depth, path
CountVid [2]	target objects	no	the target	no
StaMina [45]	objects, events	yes	targets	no
VideoAgent [24]	objects	yes	detector	no
Emb. VideoAgent [25]	3D objects	yes	detector	yes
DirectMe [73]	objects, places	yes	Objects365 [62]	est.
Hydra [35]	objects, rooms	yes	labels	yes
Ego-Topo [51]	zones	yes	none	no
PlaceTally	places	yes	none	no

Table 1. How methods decide identity. Depth, path: depth or camera path used (est.: estimated). Cells come from the primary papers. None of them uses place identity to stop double counting.

2. Related Work

Counting with vision-language models. Class-agnostic counters use exemplars, detection or open-world prompts [3, 4, 22, 43, 53, 54, 60]. Vision-language models count poorly without grounding [20, 23, 52]: they concentrate their answers on preferred integers [13] and weaken as counts grow. Pointing at objects first helps [1, 17, 19], though not inside our places (Supp. D.9). In video, DR.VIC [33] and CountVid [2] match objects across frames, StaMina [45] learns when a count should change, and benchmarks report failures on long videos [69] and on objects that reappear [14]. LVLM-Count [56] divides an image around the segmented target and sums the parts. We share its logic but find parts differently: places need no category or segmentation, so a question-blind memory can come first.

Memory for long and streaming video. Token memories compress what the model has seen [34, 64, 79, 81] or retrieve cached keys when a question arrives [21]; both ac-

cumulate without deciding which parts show the same place, so a count over them must still find every duplicate. Agent memories keep re-identified tracks [24], persistent 3D objects with depth and pose [25], image snapshots [78] or captions indexed by position and time [5]. Spatial models add 3D encoders [26, 75] or keep 3D-aware maps and memories [31, 47, 84], and Spatial-TTT [44] stores evidence in weights updated at test time. The closest counting memory is DirectMe [73], which builds object and place nodes before the question and reports object-quantity accuracy (Table 1); its counting identity still comes from object tracks, with estimated depth and pose and a fixed vocabulary, and on our streams its map holds far more entries than objects (Sec. 5.3). In the place graph of PlaceTally, a place is the unit of identity, and the memory holds no objects at all.

Place recognition and re-identification. Visual place recognition retrieves views of one place [6, 9, 37], and loop closure recognises a return to a mapped place [27]. FAB-MAP [18] adds an explicit new-place option. One place can look different while different places look alike [48]. Robotics work grouped images into rooms by graph clustering [85] and segmented image streams into places by change-point detection [59], and Ego-Topo [51] turns egocentric video into visits to the zones of a kitchen to learn what each zone affords. Topological maps [12, 61, 66, 72] and scene graphs [32, 35, 74, 82] link views, objects, places or rooms for navigation. Person re-identification matches people across cameras and times [29, 80], and sports analytics re-identifies players and locates them on the pitch [63]. For counting, re-identifying a returning room replaces object identity. It is harder than re-identifying a person, as two views of one room can share no visible surface, and unlike the zones of a pitch, places must be found in each video.

Graph partitioning. Tracking casts association as graph partitioning [10, 11], and correlation clustering [7] gives the signed objective; its online form merges clusters on summed evidence [50]. Coreference resolution links each mention to one best antecedent [41]. GraphThinker [16] builds an event graph of events and entities for grounded video reasoning; our nodes are places, built before any question. We keep the graph and replace the sum with one best link per fragment.

3. Methodology

In this section, we define the task and then describe how PlaceTally constructs and reads a question-blind place graph for each video at test time (Fig. 2).

Counting as a sum over parts. A stream V arrives frame by frame, and a category q is asked afterwards. The target is the number of distinct annotated instances of q in V. Whole-video reading computes $\hat{y} = R(V, q)$ with a reader R, whose answer stays small as the truth grows (Sec. 5.1). Reading in parts computes $\hat{y} = \sum_s R(s, q)$, which is right only if no object appears in two parts; we call such a partition object-disjoint. Object memories compute $|\{o \in \mathcal{O} : \mathrm{cat}(o) = q\}|$, which needs the object set $\mathcal{O}$. Our working assumption is that objects stay in one place between visits. Parts grouped by place are then object-disjoint for every q, and the task reduces to finding the places. PlaceTally therefore constructs, for each video, a memory model $M_V = W_\theta(V)$, a place graph G_V with the frames of each place, where W_θ is the construction below with parameters θ. It answers with

$$\hat{y}(q) = \sum_{P \in \mathcal{P}} \mathbf{1}[g_P \geq \tau_g]\, c_P, \tag{1}$$

where $\mathcal{P}$ are the places of G_V, c_P is the reader's count from at most K frames of P, and g_P, the presence gate, is the probability that the reader answers yes to "is q visible?", which keeps a place without the object out of the sum. θ is fixed on training data, which share no recording with any test stream; G_V is inferred per test video.

One memory per video. What the parts of a video share depends on which rooms this camera came back to; no number fixed across videos can stand in for it, and a rate estimated per video needs its places anyway (Sec. 5.2). Rooms in other videos only add look-alikes (Sec. 5.1), so nothing carries over between videos except θ. The design handles two risks: look-alike rooms inside a video, with a strict threshold and no merging of old places, and memory that grows with the video, by keeping at most K frames per place. A third remains: a θ fitted on one camera may not hold on another (Sec. 5.6).

Fragments. The nodes of G_V are fragments $f_1, \ldots, f_n$, cut from the stream at 5 fps. A missed cut joins two rooms for good, while a false cut, inside one recording, can be repaired, so the detector cuts generously. It is a logistic regression on six similarities between neighbouring frames, among them SigLIP 2 [68] (Supp. B), fitted on streams joined from other ARKitScenes [8] recordings, and it cuts at peaks where its log-odds exceed 3.792.

Continuation evidence. Generous cuts over-split a recording: 496 of 1,791 cuts on the training streams are false. Appearance cannot repair this, because two fragments of one recording often look at different walls: on the training streams, 93% of fragment pairs from one recording score below the return threshold defined next. Time can. At each cut, a continuation model, a logistic regression on six statistics fitted on training streams, gives p_t, the probability that the

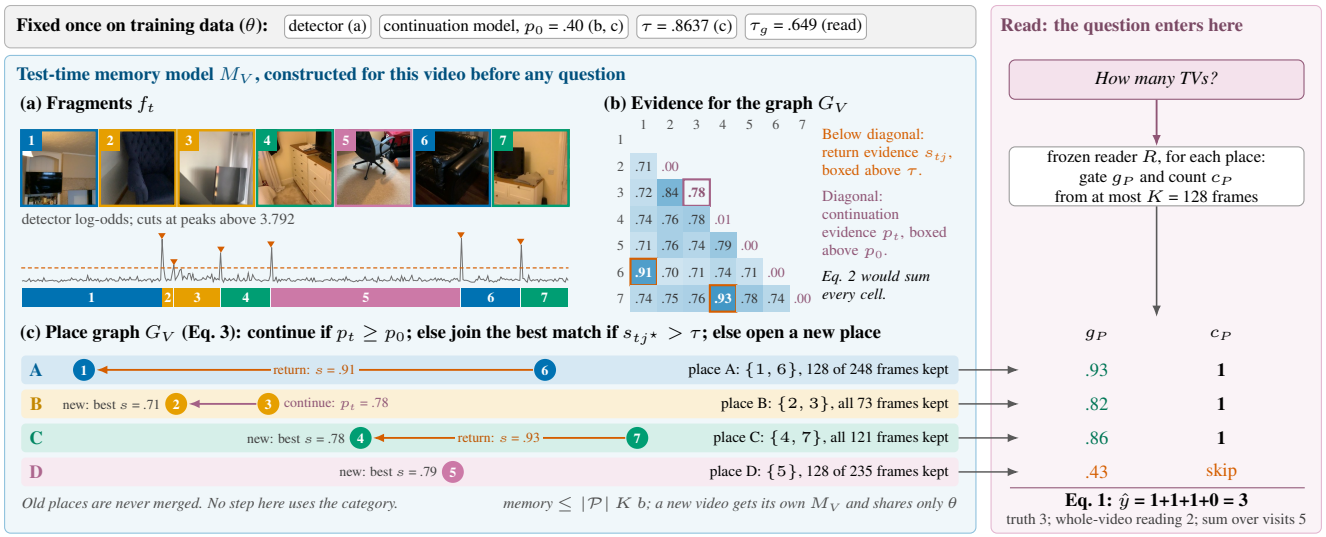


Figure 2. A PlaceTally memory is constructed per video instance before the question and read after it, shown on one real ARKit-Revisit test stream. (a) The detector cuts the stream into fragments. (b) Continuation and return evidence between fragments, which Eq. 2 would sum. (c) The place graph G_V: Eq. 3 takes one decision per fragment (node), the kept links are its edges, and each lane is a place. Read: the category enters only here; each place is gated and counted, and the counts are summed once per place (Eq. 1).

cut before f_t is false. One statistic (Supp. B) is a learned similarity: SigLIP 2 features under a linear map, trained on other recordings to match 8-s windows of one recording.

Return evidence. Time cannot connect a return, which is a new recording, so appearance has to. The return evidence s_{tj} is the SigLIP 2 chamfer similarity between f_t and an earlier fragment f_j (each frame's best cosine match, averaged both ways), and it is not learned. It is one-sided: a high score is reliable and a low one is not, as a room looks different from another side; on the training streams 40 to 50% of true-return pairs score below the threshold τ, while 95 to 97% of pairs above it show the same place (Sec. 5.4).

The best-link rule. The standard formulation of identity across frames partitions a graph [10, 11]. With weights w_{ij} that calibrate the two kinds of evidence as log-odds of the same place, the places would be

$$\mathcal{P}^\star = \arg\max_{\mathcal{P}} \sum_{P\in\mathcal{P}} \sum_{i<j;\ f_i,f_j\in P} w_{ij}, \qquad (2)$$

the objective of correlation clustering [7]. This sum counts every low score against the same place, although many true returns score low, and it needs the whole graph, which a stream does not have. PlaceTally keeps the graph and scores each fragment by its strongest link instead of the sum of all its links. Each fragment continues the current place, joins the place of its strongest earlier match, or opens a new place:

$$\ell(f_t) = \begin{cases} \ell(f_{t-1}) & \text{if } p_t \geq p_0, \\ \ell(f_{j^\star}) & \text{else if } s_{tj^\star} > \tau, \\ \text{a new place} & \text{otherwise,} \end{cases} \qquad (3)$$

where $j^\star = \arg\max_{j<t} s_{tj}$, $p_0 = 0.40$ and $\tau = 0.8637$. The care lies in four choices: a strict threshold, one strongest link, continuation first, and no merging of old places. Eq. 3 is exact for an objective of its own: once continuation is decided, each other fragment scores its best link into its own place, $s_{tj} - \tau$, or zero if it opens one, and the rule reaches this maximum for every fragment, online. It follows antecedent ranking in coreference [41] with FAB-MAP's new-place option [18]. The closest design, Ego-Topo's online zone graph [51], scores a zone by its mean match; we use the strongest link, put continuation first, and count by place.

Size under control. Each call to the reader sees at most $K = 128$ frames of one place, chosen uniformly at 1 fps, about 13 thousand tokens, whatever the video's length. Once the graph is built, answering needs only those frames, at most $|\mathcal{P}|\,K\,b$ for b bytes per frame, so K sets the memory, the input and, through it, the answer time (Sec. 5.5).

What is learned. The memory is question-blind because no part of Eq. 3 uses q. It is test-time because G_V is inferred for each video, and no weights are updated on the test video. What is trained is small: two logistic regressions with six features each and one linear projection, about

Benchmark	Videos	QA	Length	Returns	Blind
VSI-Bench [76]	288	>5,000	n/r	n/r	yes
VSI-SUPER-Count [77]	n/r	200	10–120 min	n/r	n/r
ReVSI [83]	381	n/r	n/r	n/r	partial
OST-Bench [42]	1,400	10,000	n/r	n/r	chance
EC-Bench [69]	152	1,699	57 min	n/r	n/r
TOC-Bench [14]	1,951	2,323	29 s	n/r	filter
LongSpace [40]	445	4,073	21 min	n/r	n/r
SVCBench [46]	406	1,000	to 20+ min	rotation	text-only
ARKit-Revisit test	103	460	11 min	recorded	yes
ARKit-Revisit held-out	87	403	12 min	recorded	yes
3RScan-Revisit test	134	362	n/a	recorded	yes
Long streams	38, 18	202, 68	30, 61 min	recorded	yes
MultiScan-Revisit	33	–	n/a	recorded	–

Table 2. Benchmarks we inspected and RevisitCount (below the line). Returns: how a room is seen again. Blind: blind baseline reported (partial: object-removed videos; filter: only to filter questions). n/r: not reported. Our videos: streams with questions (MultiScan: all streams).

295,000 weights, nearly all in the projection, fitted once with labels from recording and room IDs. The reader is frozen (Qwen3.5-9B [57]; 448-pixel frames, greedy decoding, prompts in Supp. B), and $\tau_g = 0.649$, like τ, was fixed before any test stream existed.

4. RevisitCount Benchmark

We further introduce RevisitCount, a benchmark for counting objects across real returns, which we will release with our code. The benchmarks we inspected report no separately recorded return to a room (Table 2), so a method can score on them without recognising a place: on VSI-SUPER-Count [77], a baseline that sees only the video's length and category rates fitted on its other durations reaches 80.8% at 60 minutes, against 32.8% for PlaceTally, while a length prior scores 17.5% on our 60-minute streams. Rendered walks do not fill this gap: on HM3D scans [58] and synthetic HSSD homes [39], a per-category prior fitted on other houses scores 58.4% and 67.1%, against 41.6% for the best blind baseline on ARKit-Revisit test (Supp. D.4, D.8).

ARKit-Revisit. Each stream joins real ARKitScenes recordings [8], one per visit; a return, a later visit to the same room, is a separate recording made the same day, and look-alike distractor rooms are included. We choose the order of rooms; each stream is a constructed walk. Truth comes from annotated 3D boxes: an object counts if it is visible in at least two frames of one visit. Objects are matched across recordings by aligning the scans (Supp. A). We keep a category only if at least one of its objects is seen on more than one visit (467 of 828 candidate questions on the test split): every question has an object seen on a return. The 14 categories are furniture, fixtures and appliances. Each stream has four rooms and two returns or six rooms and three, so about a third of visits are returns; we call this share the return rate, and Sec. 5.2 varies it. 193 training streams, from rooms in no test split, fit the continuation model and the blind baselines. The held-out split, built from 248 rooms used nowhere else, was evaluated once, after the method was fixed. A request error lost seven of the 467 test questions before scoring (Supp. A).

3RScan-Revisit. This split uses rescans of rooms taken months apart, as irregular Tango keyframes [71], with 60 training and 418 test rooms. Truth is a rendered labelled mesh with the same visibility test. It tests our assumption: 41% of target objects in both scans were moved, yet only 1.7% of the 4,787 first-scan objects are listed as removed.

MultiScan-Revisit and long streams. MultiScan-Revisit uses iPad video [49], was used for no design choice, and has too few questions for counting, so we evaluate only its places. Long streams, which join ARKit test rooms into 30- and 60-minute streams, are not an independent test.

Metric. Strict mean relative accuracy (MRA) [76] for truth g and answer p is $\frac{1}{10}\sum_a \mathbf{1}[\,|p-g|/g < 1-a\,]$ over $a \in \{.50, .55, \ldots, .95\}$. An answer off by half or more scores zero. Place quality is F1 over same-place frame pairs. Both are in per cent; a difference of two scores is absolute; intervals are 95% bootstrap intervals over streams.

Baselines and controls. Many multimodal questions can be answered without the input [15, 30, 55, 67], so every split reports blind baselines. A blind baseline sees no video content that carries the answer: a constant, a per-category prior, and controls that scale a category rate by duration, recordings or places. All are fitted on training streams of the same dataset. Whole-video reading gives the same model all 1-fps frames of the stream in one pass, capped at 2,048. *Sum over visits* counts each true visit separately and adds the counts. A fixed correction multiplies this sum by one constant, 0.69, fitted on 48 training streams. An oracle uses the truth: PlaceTally given the true visits in place of its fragments. Cambrian-S-7B [77], CountVid [2], VideoAgent [24] and DirectMe [73] run with official code (settings and checks in Supp. C). CountVid runs on a fixed subset of 40 questions, as many as time allowed (cost in Table 5), DirectMe on the same 40, and timeouts count as zero. VideoAgent runs on all 460 test questions with a Qwen3.5-9B controller; we amended its parser after the first test reads and re-read every answer. StreamMem [79] has no released code, and the Spatial-TTT [44] checkpoint fails to run on our streams.

Method	ARKit-Revisit test	ARKit-Revisit held-out	3RScan-Revisit test
Blind baselines† [15, 30]			
Fitted on training streams	41.6 [37.7, 45.4]	39.2 [35.4, 43.3]	37.1 [34.1, 40.1]
Rate × PlaceTally's places*	39.8 [35.9, 43.4]	36.8 [32.9, 40.9]	32.0 [28.4, 35.8]
Adding up parts [56, 77]			
Sum over visits (privileged)	26.9 [23.5, 30.5]	30.8 [27.4, 34.3]	17.9 [14.6, 21.3]
PlaceTally without return links	29.6 [25.8, 33.5]	–	–
PlaceTally without return links × 0.69	57.7 [53.9, 61.5]	–	–
Fixed correction, sum over visits × 0.69 (privileged)	58.6 [54.8, 62.7]	–	69.1 [64.8, 73.3]
Whole-video reading			
Qwen3.5-9B [57] (same reader)	51.2 [46.9, 55.8]	44.3 [40.0, 48.9]	**55.1** [49.9, 60.1]
Gemma 4 12B [28] (other reader)	27.8 [23.8, 32.1]	–	–
Other video models			
Cambrian-S-7B [77], whole-video reading	34.1 [30.3, 38.2]	–	–
Cambrian-S-7B [77], surprise memory	2.1 [0.7, 3.9]	–	–
VideoAgent [24] (amended parser)	4.9 [3.3, 6.6]	–	–
CountVid [2] / DirectMe [73], 40 questions‡	0.0 / 0.0	–	–
PlaceTally (ours)	**62.5** [58.4, 66.5]	**61.2** [57.7, 65.0]	47.2 [41.7, 52.9]
with Gemma 4 12B as reader	51.6 [47.2, 55.6]	–	–
Oracle (privileged; not ranked)			
PlaceTally on true visits	65.1 [61.6, 68.6]	62.8 [59.2, 66.6]	65.4 [60.9, 70.0]

–: not run, for lack of time; the pre-registered held-out test has no other video models. † Highest of eight blind baselines in each column. * Category rate per place times PlaceTally's places. ‡ PlaceTally: 57.5 [46.7, 68.5] on these 40; timeouts count as zero.

Table 3. PlaceTally leads on real ARKit returns and falls behind whole-video reading under 3RScan's camera change. Strict MRA (%) with 95% bootstrap intervals over streams. Privileged rows use true visits; bold: best row without them.

Use of each split. Training data set every parameter. ARKit-Revisit test, 3RScan-Revisit and VSI-SUPER-Count guided our failure analysis; only the held-out split and MultiScan-Revisit stayed unseen until evaluated once. Post hoc analyses chose nothing.

5. Experiments

Key findings. Against existing methods, a memory model constructed per video at test time is (1) more accurate than whole-video reading with the same reader, most at large counts (Sec. 5.1); (2) level as rooms return more or less often, where a fixed correction holds at one return rate only, with frames kept from every stay in our streams (Sec. 5.2); (3) ahead of the other memories, which over-count here and cost more (Sec. 5.3); (4) built with one decision per fragment, enough because a strong match is rarely wrong (Sec. 5.4); and (5) controlled by one number, the frames kept per place, which sets its memory, input and answer time (Sec. 5.5). (6) Its limit is a change of camera, which makes places harder to find (Sec. 5.6).

5.1. Does Counting by Place Help?

PlaceTally outperforms whole-video reading with the same reader. On the test split, PlaceTally scores 62.5% MRA (Table 3), against 51.2% for whole-video reading with the same model (paired +11.3% [5.1%, 17.2%]) and 41.6% for the best blind baseline (+20.9% [15.7%, 26.3%]). On the held-out split, evaluated once after the method was fixed, it scores 61.2% against 44.3% (+16.9% [11.5%, 22.2%]) and 39.2% for the blind baseline chosen in advance (+22.0% [17.1%, 27.0%]); both gains meet criteria we pre-registered (Supp. F). With Gemma 4 12B [28] as the reader of both methods, it scores 51.6% against 27.8% (+23.7% [18.0%, 29.3%]), so the gain is not tied to one reader.

Whole-video reading gives a few small numbers. On the test split, 87% of whole-video answers are 1, 2 or 3, and their slope, the least-squares rise in the answer per extra true object (1 for a perfect counter), is 0.32 [0.27, 0.39], against 0.72 [0.65, 0.80] for PlaceTally (Fig. 1, right). Whole-video reading scores 84.6% at two objects, where its preferred answer is right [13], and 10.5% at nine or more, where PlaceTally scores 56.3% (+45.8% [37.1%, 53.9%]; Supp. D.1).

With and without the graph. To remove the graph and change nothing else, a same-frame control gives the reader exactly the frames that PlaceTally kept, with the same prompts, in one call. On training streams this accumulated memory scores 48.7% against 64.3% with the graph, at the same memory and the same 100 thousand input tokens (slope 0.27 against 0.77). Whole-video reading, with every frame and no graph, keeps more (2.9 against 2.3 MB per minute) and scores less. The gain comes first from counting in parts: 73% of visits with a category hold one such object and 89%

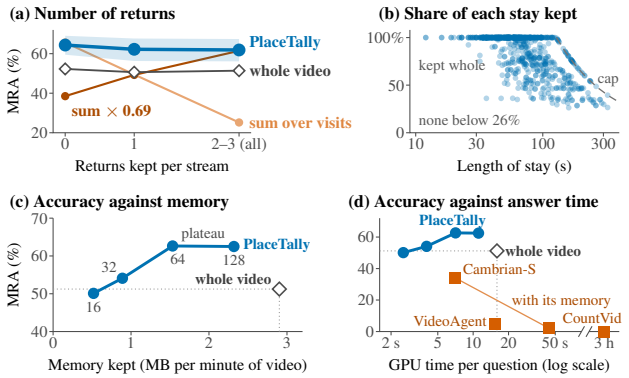


Figure 3. Robustness and cost on ARKit test rooms (post hoc, from actual answers). (a) Returns removed from the same 60 test streams; orange: the sum over true visits, as it is and times 0.69 (privileged); band: 95% interval. (b) The 823 test stays; dashed: the cap of 128 frames per place. (c, d) $K = 16$ to 128 frames per place at 11 minutes (gate refitted for 16 and 32); answer time excludes the one-off build (Table 5). Dotted lines pass through whole-video reading: above and left of them is better on both axes.

	ARKit test (598 places)			3RScan test (824)		
Variant	F1	Places	Exact	F1	Places	Exact
How the stream is cut (same linking)						
Fixed 8 s windows	32.0	4393	0	59.5	1699	8
TransNet V2 [65]	80.7	510	27	48.8	423	6
SigLIP [68] similarity drop	85.7	764	32	70.9	2261	2
Detector, no continuation	96.3	764	51	67.8	1737	3
How fragments are linked (shared log-odds weights)						
Eq. 3	96.2	595	83	74.9	781	39
Solver of Eq. 2 [38]	96.6	629	76	76.4	1148	25
Average link [36]	96.5	629	76	75.1	1134	26
Single link [36]	95.1	577	74	71.8	621	31
PlaceTally	96.9	594	85	75.3	752	**42**
with MegaLoc [9] returns	**97.6**	585	**92**	**76.5**	738	36

Table 4. Finding places, post hoc; F1 in %. Places: predicted (true in the header). Exact: streams with the right number of places, of all 109 and 150. Linking rows share log-odds weights, so Eq. 3 there differs from PlaceTally, which has its own thresholds. With MegaLoc, only the return evidence changes, its threshold fitted on training streams.

at most two, where the reader is best. Without return links every return is counted again (29.6% on test), and one graph shared across 40 training videos raises false merges from 4.0% to 11.7% and comparisons per fragment from 5 to 194.

Longer streams have larger counts. On 30- and 60-minute streams, whole-video reading falls to 41.3% and 26.0% and PlaceTally scores 68.2% and 65.6% (+26.8% [20.0%, 33.7%], +39.6% [29.3%, 48.9%]). Longer streams pass through more rooms and hold more objects (mean true count 3.6, 5.5 and 6.3), but whole-video reading's answer barely changes (3.2 to 3.5 at true counts of 5 to 8): one pass over many frames falls back on the few small answers the model prefers. Its fall comes mostly from larger true counts; at 60 minutes its 2,048-frame cap also drops frames.

5.2. Can a Fixed Correction Replace the Graph?

On the benchmark's own return rate, it nearly can. As every ARKit-Revisit question has an object seen on a return, the sum over visits scores only 26.9%. The fixed correction, multiplying this sum by 0.69, lifts it to 58.6%, which is 3.9% [−0.3%, 8.0%] below PlaceTally. The constant is about the share of first visits (5.7 rooms per 8.2 visits), which varies little, so at this rate a constant nearly matches counting by place, also at 30 and 60 minutes (Supp. D.3).

It holds only at that rate. A camera in use need not return at this rate, so we removed returns from 60 random test streams, keeping none, one, or all of them per stream, rebuilt each place graph and counted with the method unchanged (291 questions, Fig. 3a). Each return shows a room again, often from another side, yet from no returns to all of them PlaceTally stays level (64.4%, 62.2%, 61.9%; a change of −2.5% [−7.1%, 2.2%] on the same questions), while the fixed correction rises by 23.0% [15.9%, 29.8%] (38.5%, 49.2%, 61.4%). So the correction ties only at the rate it was fitted on, 0.4% [−4.6%, 5.6%] below PlaceTally on these streams, against 13.0% [6.7%, 18.7%] and 25.9% [19.1%, 32.4%] at one return and none. Whole-video reading is level but lower (52.3%, 50.6%, 51.4%), and PlaceTally's lead over it holds at every rate (10.5% to 12.1%, intervals above zero). A constant fitted on training streams at a mix of these rates (0.81) is still 9.2% [5.6%, 12.9%] below PlaceTally, and a per-video rate, PlaceTally's places over the visits, needs the place graph anyway (Supp. D.2). These results indicate that one number cannot replace a test-time memory model for each video: what can be added up depends on which rooms the camera returned to, and the question-blind place graph works that out before any question is asked.

Short stays are kept. Each place's frames are thinned evenly, so on test streams every stay keeps at least 12 frames and 95% of the shortest tenth are kept whole (Fig. 3b). Questions whose shortest stay with an asked object lasts 42 s or less score as well as those above 67 s (67.8% and 66.7%), also at equal true count (Supp. D.12).

5.3. Do Other Video Models Close the Gap?

No: on these streams, the other memories give mostly too-high answers. VideoAgent scores 4.9% on all 460 questions, and 304 of its answers are too high. CountVid scores zero on its 40 questions, against 57.5% [46.7%, 68.5%] for PlaceTally: all 36 completed answers are too high, at a median

of 16 times the truth, also inside single true places (Supp. C); the cause remains open. DirectMe also scores zero: its map holds far more entries than objects before any return, so 38 of its 39 answers are too high. Cambrian-S reaches 34.1% reading the whole video and 2.1% with its surprise memory (Table 3). These memories also cost more: VideoAgent keeps 6.7 times the memory and takes 17 times as long to build it, CountVid takes about 3.1 GPU-hours per question, and Cambrian-S, building its memory in each query, 44.4 GPU-seconds against 11.1 plus our one-off build (Fig. 3d).

5.4. Is One Best Link Enough?

Cuts and false cuts. In Table 4, fixed windows and a shot detector miss boundaries that matter, and our detector reaches a place F1 of 96.3% on ARKit test. The continuation model then gives 85 of 109 ARKit streams the right place count, against 51 without it, with a small gain in counting (+2.0% [−0.6%, 4.6%]).

Why the best-link rule is good enough. A solver of Eq. 2 [38] finds no better places than PlaceTally (F1 −0.3% [−1.9%, 1.2%] on ARKit and +1.1% [−3.4%, 5.2%] on 3RScan) and it over-splits, with 1,148 places against 824 true on 3RScan (Supp. D.7). The reason is that the evidence is one-sided. On ARKit test and 3RScan test, 96% and 91% of pairs above τ show the same place, while 42% and 67% of true-return pairs fall below it. A returning fragment's best match in the earlier recording of its room clears τ in 85% and 71% of cases, the mean match in only 74% and 36%, so one strong link is the reliable test, and a mean over the place, as in a zone score [51], misses more returns.

Headroom. Given true visits, PlaceTally rises from 62.5% to 65.1% on test and from 61.2% to 62.8% on the held-out split, and only 1.9% [0.7%, 3.7%] of objects appear in two predicted places, so the parts are nearly object-disjoint. On training streams, most remaining error is the reader's: with true places it scores 66.5%, against 89.3% with exact counts in every gated place (Supp. D.9).

5.5. How Much Does the Memory Cost?

One number sets the cost. Input tokens are linear in stored bytes ($R^2 = 0.99$) and answer time is linear in the input ($R^2 = 0.93$), so K sets all three costs (Table 5; post hoc). With $K = 128$, memory grows linearly with the video, at 2.3 MB per minute, and the graph is built 15 times faster than real time, at 11, 30 and 60 minutes alike (Supp. D.11).

Less memory costs little accuracy. With $K = 64$, PlaceTally reads as much as whole-video reading, stores about half as much and answers 2.2 times faster; it scores 62.6% against 51.2%. With $K = 16$ it keeps 18% of whole-video reading's memory and nearly matches it (50.1% against 51.2%). More frames stop helping early: at 11 minutes, $K = 128$ takes 1.5 times the memory and 1.6 times the answer time of $K = 64$ for no gain (Fig. 3c), and on 30- and 60-minute streams it adds 3.6% and 3.4%, intervals reaching zero (Supp. D.6). As 63 to 65% of places are read in full at $K = 128$, a larger K changes only the rest. About half the input is the presence gate, which adds 2.1% [−0.9%, 5.3%].

Method	Build speed	Memory MB/min	Input k tokens	Query GPU-s	MRA (%)
PlaceTally, $K = 16$	15	0.52	19.0	2.5	50.1
PlaceTally, $K = 64$	15	1.53	68.3	7.1	62.6
PlaceTally, $K = 128$	15	2.32	105.8	11.1	62.5
Whole-video reading	–	2.90	69.1	15.9	51.2
VideoAgent [24]	0.88	15.5*	n/c	15.3	4.9
CountVid [2]	–	7.6*	–	11,084	0.0†
DirectMe [73]	0.24	3.9*	–	–	0.0†
Cambrian-S whole video [77]	–	n/m	n/c	7.0	34.1
Cambrian-S memory [77]	–	61.9‡	n/c	44.4	2.1

Table 5. Cost on ARKit test (11-minute videos), post hoc, on one A100-40GB GPU. K: frames kept per place ($K = 16$: refitted gate); PlaceTally and whole-video reading use Qwen3.5-9B. Build: × real time, once per video. Memory: kept after construction (PlaceTally: 3.0 MB/min while building). Input: all 460 questions; query: the same 50, nothing cached. –: none. n/c: counted differently. n/m: not measured. * JPEG estimate (as run: 87, 68, 24 MB/min). ‡ GPU state while answering. † 40 questions.

5.6. When Places Are Hard to Find

A change of camera. Places stay the right unit under a change of camera, but finding them needs a continuation model fitted to that camera. On 3RScan-Revisit, PlaceTally on true visits reaches 65.4%, above whole-video reading at 55.1%, while with its own fragments it falls behind (47.2%, paired −7.9% [−14.8%, −1.0%]; Table 3). Place F1 is 75.3%, against 96.9% on ARKit, and 31.1% [27.2%, 35.3%] of objects appear in two predicted places. The continuation model, fitted on ARKit video, rejoins non-adjacent fragments of one recording 72 to 83% of the time (99% on ARKit) and merges adjacent fragments of different rooms 22 to 26% of the time (3%); one trained across cameras lost accuracy on ARKit, so we kept ours.

Look-alike rooms. On MultiScan-Revisit, place F1 is 90.2% [83.6%, 96.4%]: 97.8% on ordinary streams and 84.4% with look-alike rooms, which can be merged (Supp. E).

6. Conclusion

We present PlaceTally, a question-blind test-time memory model that constructs a place graph for each video before any question is asked. To count, a memory does not need to know what it is looking for, only where it has been. Experiments on

real ARKitScenes returns show that PlaceTally outperforms whole-video reading with the same reader on test and held-out rooms, and stays level as the return rate changes, while no single correction holds across rates. One number, the frames kept per place, sets its memory, input and answer time, and it is built 15 times faster than real time.

Limitations and future work. Places stand in for object identity while objects stay in their rooms, as our 14 categories do. Our streams join real recordings at a hard cut, and the other return rates come from removing visits from tested rooms; new recordings of continuous walks would test both. The next steps are a continuation model fitted to each camera and return evidence trained on look-alike rooms.

LLM Usage Statement

We used generative AI tools (Claude Code and Codex) for code implementation, debugging, for searching related work, and occasionally for suggestions on experimental variations. We also used them to improve writing clarity and grammar, refine figures, and check the manuscript for unclear presentation or misunderstandings. The authors decided the main hypotheses and the research methodology. All AI-assisted work was reviewed by the authors, and we take responsibility for the final content of the paper.

References

[1] Simone Alghisi, Massimo Rizzoli, Seyed Mahed Mousavi, and Giuseppe Riccardi. Getting to the Point: Pointing Improves LVLMs at Counting. *arXiv preprint arXiv:2603.21746*, 2026. 2

[2] Niki Amini-Naieni and Andrew Zisserman. Open-World Object Counting in Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 1, 2, 5, 6, 8

[3] Niki Amini-Naieni and Andrew Zisserman. CountGD++: Generalized Prompting for Open-World Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2

[4] Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. CountGD: Multi-Modal Open-World Counting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2

[5] Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 3

[6] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[7] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation Clustering. *Machine Learning*, 56:89–113, 2004. 3, 4

[8] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *NeurIPS Datasets and Benchmarks Track*, 2021. 3, 5

[9] Gabriele Berton and Carlo Masone. MegaLoc: One Retrieval to Place Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 3, 7

[10] Guillem Brasó and Laura Leal-Taixé. Learning a Neural Solver for Multiple Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[11] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying Short and Long-Term Tracking With Graph Hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 4

[12] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural Topological SLAM for Visual Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[13] Marc Chapus, Carlos Crispim-Junior, Véronique Eglin, and Atilla Baskurt. CountVideoBench: A Benchmark to Evaluate Prediction Biases on the Object Counting task by Video-Language Models. In *British Machine Vision Conference (BMVC)*, 2026. 1, 2, 6

[14] Junzhe Chen, Siyuan Meng, Yuxi Chen, Man Zhao, Wenyao Gui, and Xiaojie Guo. TOC-Bench: A Temporal Object Consistency Benchmark for Video Large Language Models. *arXiv preprint arXiv:2605.09904*, 2026. 2, 5

[15] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5, 6

[16] Zixu Cheng, Da Li, Jian Hu, Yuhang Zang, Ziquan Liu, Shaogang Gong, and Wei Li. GraphThinker: Reinforcing Temporally Grounded Video Reasoning with Event Graph Thinking. *arXiv preprint arXiv:2602.17555*, 2026. 3

[17] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28652–28668, 2026. 2

[18] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008. 3, 4

[19] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas

Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[20] Jinhong Deng, Limeng Qiao, and Guanglu Wan. HoloCount: A Holistic Visual Counting Benchmark for MLLMs. *arXiv preprint arXiv:2607.06420*, 2026. 2

[21] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming Video Question-Answering with In-context Video KV-Cache Retrieval. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[22] Nikola Djukic, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A Low-Shot Object Counting Network With Iterative Prototype Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[23] Ahmed Oumar El-Shangiti, Abzal Nurgazy, Hilal AlQuabeh, Nikolai Rozanov, and Kentaro Inui. The Count Is There, but Misaligned: Understanding and Correcting Counting Failures in VLMs. *arXiv preprint arXiv:2607.09544*, 2026. 2

[24] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A Memory-augmented Multimodal Agent for Video Understanding. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 5, 6, 8

[25] Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied VideoAgent: Persistent Memory from Egocentric Videos and Embodied Sensors Enables Dynamic Scene Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 3

[26] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3

[27] Dorian Gálvez-López and Juan D. Tardós. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012. 3

[28] Gemma Team. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026. 6

[29] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors. *Person Re-Identification*. Springer, 2014. 3

[30] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 6

[31] Bo Gu, Zhikang Zhang, Zizhuang Wei, Zhenyuan Chen, Lingyun Li, and Zhuoyi Song. SpaceMind++: Toward Allocentric Cognitive Maps for Spatially Grounded Video MLLMs. *arXiv preprint arXiv:2605.09449*, 2026. 3

[32] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 3

[33] Tao Han, Lei Bai, Junyu Gao, Qi Wang, and Wanli Ouyang. DR.VIC: Decomposition and Reasoning for Video Individual Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[34] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 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)*, 2024. 2

[35] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization. In *Robotics: Science and Systems (RSS)*, 2022. 2, 3

[36] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, 1999. 7

[37] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. AnyLoc: Towards Universal Visual Place Recognition. *IEEE Robotics and Automation Letters*, 9(2): 1286–1293, 2024. 3

[38] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjoern Andres. Efficient Decomposition of Image and Mesh Graphs by Lifted Multicuts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 7, 8

[39] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16384–16393, 2024. 5

[40] Shiqiang Lang, Jing Liu, Haoyang He, Peiwen Sun, Yuanteng Chen, Tao Liu, Lan Yang, Longteng Guo, and Honggang Zhang. LongSpace: Exploring Long-Horizon Spatial Memory from Perception to Recall in Video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5239–5256, 2026. 5

[41] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end Neural Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2017. 3, 4

[42] Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. OST-Bench: Evaluating the Capabilities of MLLMs in Online Spatio-temporal Scene Understanding. In *NeurIPS Datasets and Benchmarks Track*, 2025. 1, 5

[43] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. CounTR: Transformer-based Generalised Visual Counting. In *British Machine Vision Conference (BMVC)*, 2022. 2

[44] Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-TTT: Streaming Visual-based Spatial Intelligence with Test-Time Training. *arXiv preprint arXiv:2603.12255*, 2026. 3, 5

[45] Pengyiang Liu, Dongyue Lyu, Junbo Niu, Zhongyue Shi, Jiahao Xie, and Si Liu. When Should the Count Change? Learning State Maintenance for Causal Video Counting. *arXiv preprint arXiv:2609.35416*, 2026. 2

[46] Pengyiang Liu, Zhongyue Shi, Hongye Hao, Qi Fu, Xueting Bi, Siwei Zhang, Xiaoyang Hu, Zitian Wang, Linjiang Huang, and Si Liu. SVCBench: A Streaming Video Counting Benchmark for Spatial-Temporal State Maintenance. In *European Conference on Computer Vision (ECCV)*, 2026. 5

[47] Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. Vision-Language Memory for Spatial Reasoning. In *European Conference on Computer Vision (ECCV)*, 2026. 3

[48] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J. Leonard, David Cox, Peter Corke, and Michael J. Milford. Visual Place Recognition: A Survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016. 3

[49] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel X. Chang, and Manolis Savva. MultiScan: Scalable RGBD Scanning for 3D Environments with Articulated Objects. In *Advances in Neural Information Processing Systems*, 2022. 5

[50] Claire Mathieu, Ocan Sankur, and Warren Schudy. Online Correlation Clustering. In *Proceedings of the Symposium on Theoretical Aspects of Computer Science*, 2010. 3

[51] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. EGO-TOPO: Environment Affordances from Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4, 8

[52] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to Count to Ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[53] Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. DAVE – A Detect-and-Verify Paradigm for Low-Shot Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[54] Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A Novel Unified Architecture for Low-Shot Counting by Detection and Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[55] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. 5

[56] Muhammad Fetrat Qharabagh, Mohammadreza Ghofrani, and Kimon Fountoulakis. LVLM-Count: Enhancing the Counting Ability of Large Vision-Language Models. *Transactions on Machine Learning Research*, 2026. 1, 2, 6

[57] Qwen Team. Qwen3.5-9B model card. Official model card, 2026. Accessed 29 October 2026. 5, 6

[58] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In *NeurIPS Datasets and Benchmarks Track*, 2021. 5

[59] Ananth Ranganathan. PLISS: Labeling Places Using Online Changepoint Detection. *Autonomous Robots*, 32:351–368, 2012. 3

[60] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning To Count Everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[61] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric Topological Memory for Navigation. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[62] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[63] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe De Vleeschouwer. SoccerNet Game State Reconstruction: End-to-End Athlete Tracking and Identification on a Minimap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3293–3305, 2024. 3

[64] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[65] Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the ACM International Conference on Multimedia*, 2024. 7

[66] Lauri Suomela, Jussi Kalliola, Harry Edelman, and Joni-Kristian Kämäräinen. PlaceNav: Topological Navigation through Place Recognition. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 3

[67] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5

[68] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025. 3, 7

[69] Fumihiko Tsuchiya, Taiki Miyanishi, Shunsuke Yasuki, Mahiro Ukai, Nakamasa Inoue, Shuhei Kurita, Yusuke Iwasawa, and Yutaka Matsuo. Diagnosing Long-Video Quantitative Reasoning in Multimodal LLMs via Enumeration and Counting. *arXiv preprint arXiv:2603.29943*, 2026. 2, 5

[70] Vishaal Udandarao, Shyamgopal Karthik, Surabhi S. Nath, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Solving Spatial Supersensing Without Spatial Supersensing. *arXiv preprint arXiv:2511.16655*, 2025. 2

[71] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D Object Instance Re-Localization in Changing Indoor Environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 5

[72] Jiaming Wang, Jizhuo Chen, Diwen Liu, Atharva Ghotavadekar, Jiaxuan Da, Linh Kästner, and Harold Soh. Change-Robust Online Topological Memory for Long-Term Relocalization and Semantic Navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. 3

[73] Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep It in Mind: User Centric Continual Spatial Intelligence Reasoning in Egocentric Video Streams. In *International Conference on Machine Learning (ICML)*, 2026. 2, 3, 5, 6, 8

[74] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. In *Robotics: Science and Systems (RSS)*, 2024. 3

[75] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM Capabilities in Visual-based Spatial Intelligence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3

[76] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 5

[77] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards Spatial Supersensing in Video. In *International Conference on Learning Representations (ICLR)*, 2026. 1, 5, 6, 8

[78] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3D-Mem: 3D Scene Memory for Embodied Exploration and Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[79] Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. StreamMem: Query-Agnostic KV Cache Memory for Streaming Video Understanding. *arXiv preprint arXiv:2508.15717*, 2025. 2, 5

[80] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. 3

[81] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient Real-Time Understanding for Long Video Streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[82] Juexiao Zhang, Gao Zhu, Sihang Li, Xinhao Liu, Haorui Song, Xinran Tang, and Chen Feng. Multiview Scene Graph. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[83] Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. ReVSI: Rebuilding Visual Spatial Intelligence Evaluation for Accurate Assessment of VLM 3D Reasoning. In *International Conference on Machine Learning (ICML)*, 2026. 5

[84] Xinyi Zheng, Yunze Liu, Chi-Hao Wu, Fan Zhang, Hao Zheng, Wenqi Zhou, Walterio W. Mayol-Cuevas, and Junxiao Shen. SpatialMem: Metric-Aligned Long-Horizon Video Memory for Language Grounding and QA. *arXiv preprint arXiv:2601.14895*, 2026. 3

[85] Zoran Zivkovic, Olaf Booij, and Ben Kröse. From Images to Rooms. *Robotics and Autonomous Systems*, 55(5):411–418, 2007. 3