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

Supplementary Material

Version 1, 17 November 2026. The latest version is at https://alibayosmanli.github.io/placetally.

Sections A to H and the tables and figures numbered with S are in this supplement; all other section, table, figure, equation and reference numbers point to the main paper.

A. RevisitCount Benchmark Details

Tables S1 and S2 give statistics for every split. The release files, and scripts that rebuild every stream from the original downloads, are in the benchmark folder of the project's GitHub repository.

Rooms and recordings. An ARKitScenes [8] recording covers one room, and some rooms were recorded more than once on the same day. A return in RevisitCount is a later visit to a room through a different recording of it. A recording is usable if it has at least ten frames at 1 fps and at least one annotated object passes the visibility test below. All ARKit splits are built by one script, run with a different seed and pool of recordings for each split. For each stream it

- draws the returning rooms from rooms with at least two usable recordings, and uses one recording for the first visit and another for the return;
- draws the other rooms from the remaining rooms of the pool; in the first training set, only from rooms with a single usable recording;
- uses each room in at most two streams of the split.

Streams alternate between four rooms with two returns and six rooms with three. The first visits come in random order, then the returns in random order, each played backwards with probability 0.5. Every other pair of streams ends with a visit to one more new room after the returns. Each visit uses all 1-fps frames of its recording, at 640×480, and the stream's time runs on by 1 s per frame plus 2 s after each visit.

Look-alike rooms. In every other block of four streams, one first-visit room that is not revisited is replaced by a look-alike: the unused room whose recording has the highest SigLIP 2 [68] chamfer similarity to the first visit of a returning room. It is placed after that room's first visit and before any return. In the first training set the room it imitates is always the first returning room. In the other splits it is the returning room whose closest look-alike is the most similar.

Truth. Visibility in a frame comes from the annotated 3D boxes and the sensor depth. On each box face that faces the camera, 8×8 points are projected into the frame. A point passes if it lies in the image, has a depth reading, and is no more than $\max(5\text{ cm}, 3\%)$ of that depth behind the surface the sensor saw. An object is visible in the frame if at least 12 of its points pass, they make up at least 15% of its points, and the box around them covers at least 625 pixels. It is seen in a visit if that box covers at least 0.5% of the image in at least two frames of the visit.

Objects are matched between the two recordings of a returning room by label and position after aligning the two scans. A category is dropped from a stream if any of its objects in a returning room has an ambiguous match, or if both recordings have objects of that category left unmatched. The answer is the number of distinct objects seen over all visits. A question is kept if its answer is at least 1 and differs from the sum of the per-visit counts, so at least one object of the category is seen on more than one visit.

Categories and questions. ARKitScenes annotates 17 object classes, and RevisitCount uses 14 of them: bathtub, bed, chair, fireplace, oven, refrigerator, sink, sofa, stool, stove, table, toilet, TV or computer monitor, and washing machine. These are the names in the reader's prompts (Table S4). Cabinet, shelf and dishwasher are not used. Table S2 gives the questions per category and per true count.

Training, test and held-out splits. No training stream shares a recording or a room with a test or held-out stream (Sec. B). Of the 65 streams in the first training set, 17 lack 5-fps frames for one of their recordings. PlaceTally needs those frames to find cuts, so it could not run on these streams, and they were left out of every fit. They are kept in the release because the reader-headroom study (Sec. D.9) uses all 65 streams. The held-out split was built from 248 rooms used in no other split, with the same script and only the data folders and the seed changed. It was evaluated once, after the method was fixed. A request error lost seven of the 467 test questions before scoring; with them, PlaceTally and whole-video reading score 63.0% and 51.1%.

Long streams. The 30- and 60-minute streams use the same script on the 552 recordings of the test streams, with 14 rooms and 7 returns or 28 rooms and 14 returns, plus the extra new room in half of the streams. The test rooms had

Split	Streams	With questions	Questions	Rooms used	Length, min mean (range)	Rooms per stream	Visits per stream	Look-alike streams
ARKit-Revisit								
First training set, all	65	58	226	238	10.8 (5–24)	5.5	8.0	49%
of which fitted on	48	41	141	193	10.5 (6–18)	5.4	7.8	44%
Second training set	145	126	553	400	10.5 (5–21)	5.5	8.0	50%
Test	109	103	460	300	11.1 (4–19)	5.5	8.0	49%
Held-out	90	87	403	248	11.8 (5–23)	5.5	8.0	49%
Long, 30 minutes	39	38	202	292	30.1 (22–38)	14.5	21.5	49%
Long, 60 minutes	19	18	68	282	60.9 (52–73)	28.5	42.5	42%
3RScan-Revisit								
Training	22	20	42	60	n/a	5.5	8.0	45%
Test	150	134	362	415	n/a	5.5	8.0	49%
MultiScan-Revisit	33	–	–	92	37.5 (18–71)	5.5	7.9	48%

Table S1. RevisitCount splits. Questions are those scored; on the test split 467 questions on 105 streams were built and a request error lost seven before scoring. Training streams are not scored; their columns count the questions built. Rooms used: distinct rooms over the split's streams (on 3RScan test, 415 of the split's 418). Length includes the 2 s added after each visit; 3RScan keyframes have no fixed rate, so no length is given.

	First training	Second training	ARKit test	ARKit held-out	Long 30 min	Long 60 min	3RScan test
Questions per category							
bathtub	14	39	21	6	18	5	12
bed	20	40	43	35	20	2	47
chair	18	43	45	39	9	0	30
fireplace	13	31	33	32	19	10	7
oven	13	41	30	28	16	9	25
refrigerator	10	27	25	22	16	7	22
sink	30	58	46	38	22	3	74
sofa	19	42	35	36	10	1	14
stool	4	49	27	31	13	5	22
stove	15	27	18	19	10	6	30
table	20	48	41	41	2	0	5
toilet	19	41	28	21	21	10	38
TV or computer monitor	22	44	50	41	8	1	28
washing machine	9	23	25	14	18	9	8
Questions per true count							
1	56	128	103	93	20	1	144
2	57	160	118	104	26	2	98
3 or 4	62	152	142	89	69	19	77
5 to 8	36	83	65	70	56	30	21
9 or more	15	30	39	47	31	16	22
Built questions	226	553	467	403	202	68	362

Table S2. Built questions per category and per true count. Category names are those in the reader's prompts (Table S4). 3RScan counts are aggregates only; its answers are not redistributed.

been seen during the failure analysis, so the long streams are not a held-out test.

3RScan-Revisit and MultiScan-Revisit. 3RScan-Revisit is built in the same way from 3RScan [71], but a return is a rescan of the room made months later (Sec. 4). Its frames are the stored keyframes, which have no fixed rate, and its truth comes from the rendered labelled mesh with the same two-frame rule. The release holds only its recording IDs, frame indices and order; the build script derives the answers from the user's own copy of 3RScan. MultiScan-Revisit is built in the same way from MultiScan [49], at 1 fps from the video timestamps, and is used only to score place finding, so the release gives its room identities and no questions.

Known issues.

- Each recording has one orientation label. Frames recorded while the phone was turned stay sideways after the frames are made upright, and a visual check of sampled streams before the held-out evaluation found such frames.

- ARKit returns are recorded the same day. 3RScan-Revisit, whose rescans are months apart, covers rooms that changed between visits.
- Two MultiScan scenes, 48 and 71, show the same room (Sec. E).

B. Method Details

This section gives the settings of every part of PlaceTally in Sec. 3. All of them were fixed before any test stream was built, and none was changed afterwards.

Training data share no recording with a test stream. We compared, by ARKitScenes recording ID, every set used to fit a setting with the test and held-out streams. These are the training streams, the streams of our earlier studies on which τ and τ_g were set, and the recordings used to train the detector and the projection. None shares a recording with a test or held-out stream. The detector and the projection were trained on 22 October. At that time no test or held-out recording had features on disk, and their training code excludes the training streams and every recording selected for the test split by ID. The training streams also share no room with the test or held-out streams.

Detector. Frames are taken at 5 fps and made upright with the recording's orientation label. For each frame we keep a 28×28 grey thumbnail of the central square (zero mean, unit norm), a joint colour histogram with $8 \times 8 \times 8$ bins, and the SigLIP 2 code of the frame resized to 384×384. For each pair of neighbouring frames the detector computes the three similarities (thumbnail correlation, histogram intersection and SigLIP 2 cosine), each raw and minus its median over the five pairs on either side, and a logistic regression with balanced classes maps these six values to the log-odds of a cut. It was fitted on 120 synthetic streams, each joining six ARKitScenes recordings that appear in no RevisitCount stream. Each recording is reversed with probability 0.15, and every other join is a hard one, to the unused recording most similar to the previous one. The threshold is set low on purpose, because a false cut can be undone later and a missed one cannot. On 120 other synthetic streams from the same training recordings, 0.1% of the neighbouring frame pairs inside one recording score above 3.792, and 96% of the joins between recordings do. Peaks closer than 10 frame pairs (2 s) are suppressed.

Continuation model. At each cut it reads six statistics: the detector's highest log-odds inside the cut, the learned and the raw similarity of the new fragment to the previous one, the highest learned similarity to any older fragment (zero for the second fragment), and the logarithms of the two fragments' lengths in 1-fps frames. A logistic regression on the standardised statistics (L2 penalty, $C = 1$) gives p_t, the probability that the cut is false, and the cut is undone when $p_t \geq p_0 = 0.40$. This level gave the most training streams with exactly the right number of places. The model was fitted on the 193 training streams, whose 1,791 cuts include 496 false ones. Table S3 lists its weights. The learned similarity to the previous fragment carries most of the weight, while a strong cut and long fragments count against continuation.

Statistic	Mean	SD	Weight
Detector log-odds inside the cut	15.28	7.60	−1.44
Learned similarity to the previous one	0.481	0.223	3.50
Raw similarity to the previous one	0.758	0.062	−0.15
Best learned similarity to an older fragment	0.573	0.284	−0.12
Log length of the new fragment	3.62	1.07	−1.48
Log length of the previous fragment	3.62	1.07	−1.60
Intercept			−4.38

Table S3. The continuation model. Weights apply to statistics standardised with the mean and standard deviation over the 1,791 cuts of the training streams; p_t is the logistic of the weighted sum.

Learned similarity. Eight consecutive 1-fps SigLIP 2 codes form a window, with a stride of 4 s. A linear map from 1,152 to 256 dimensions, initialised from the principal components of the training codes, projects each code, and the similarity of two windows or fragments is the chamfer similarity of their projected, normalised codes. Training used a logistic loss on pairs of windows: windows of one recording at least 24 s apart as positives, and windows of different recordings as negatives, with the 10% most similar negatives in each batch weighted five times. We used Adam with learning rate 10^{-3} for 2,000 steps of 512 pairs, with a fixed seed, and stopped early on 10% of the training recordings, kept out of the fitting. The training recordings are ARKitScenes recordings in no RevisitCount stream. The map holds 294,912 weights, nearly all of the method's trained weights.

Return evidence. s_{tj} is the symmetric chamfer similarity between the 1-fps SigLIP 2 codes of f_t and f_j: each frame's best cosine match in the other fragment, averaged over its fragment, and the two directions averaged. The threshold $\tau = 0.8637$ was set on 19 October on the exploration streams of an earlier study, before any test stream existed, and was never changed.

The best-link rule. Fragments are processed in stream order. For each fragment f_t after the first:

1. if $p_t \geq p_0$, f_t joins the place of f_{t-1};
2. otherwise, with $j^\star = \arg\max_{j<t} s_{tj}$, f_t joins the place of $f_{j^\star}$ if $s_{tj^\star} > \tau$;
3. otherwise, f_t opens a new place.

Places are never merged, and no step reads the question.

Eq. 3 exactly maximises a sum of per-fragment scores. Fix the continuation decisions, and give every other fragment f_t the score $\max_{j<t,\ f_j \in P(f_t)}(s_{tj} - \tau)$ if its place $P(f_t)$ holds an earlier fragment, and 0 if it opens a place. Each score is a maximum over a subset of the earlier fragments, so no fragment can score more than $\max(0, s_{tj^\star} - \tau)$ in any partition. Eq. 3 reaches this bound for every fragment at once. If $s_{tj^\star} > \tau$, f_t joins the place that holds $f_{j^\star}$ and scores exactly $s_{tj^\star} - \tau$. Otherwise it opens a place and scores 0. Later fragments do not change an earlier fragment's score, so the decisions can be taken online. We noticed this property after the rule was fixed. It does not make Eq. 3 an approximation of Eq. 2, which scores every pair.

Presence gate and reader. A place's frames are its 1-fps frames, of which at most $K = 128$ are chosen uniformly. Each is made upright with its recording's orientation label and placed on a 448×448 black square, and the reader receives them as one video with the stream's own timestamps. The reader is Qwen3.5-9B [57] (snapshot c2022362) served with vLLM, with greedy decoding. The count read allows eight new tokens and takes the first integer in the reply. The gate read takes one token: g_P is the probability of "yes" divided by the probability of "yes" or "no", each summed over four spellings (with and without a leading space and capital) among the 20 most likely first tokens. The threshold $\tau_g = 0.6487$ maximises Youden's J (0.874) over 1,183 present and 1,801 absent cases from the exploration segments of an earlier study. It was fixed on 20 October, before any test stream existed. Table S4 gives the prompts.

What is trained. The detector's logistic regression (six weights and a bias), the continuation model (six weights and a bias) and the projection (294,912 weights) are trained: about 295,000 weights, fitted with labels from recording and room IDs only. SigLIP 2 (google/siglip2-so400m-patch14-384, snapshot e8e48729) and the reader are frozen.

C. Baselines, Controls and Other Models

Blind baselines. There are eight variants, each fitted on training streams that share no room with the scored stream:

- a constant, the median answer;
- a per-category prior, the median answer for the category;
- six controls that multiply a count from the scored stream (its visits, its places or its length in seconds) by the mean or the median, over training questions of the same category, of the answer per unit.

The row "Rate × PlaceTally's places" in Table 3 uses the places PlaceTally finds, so it sees no truth.

For the held-out split one blind was fixed in advance. We scored all eight on the 553 questions of the second training set, each fitted without the scored stream's rooms. The per-category prior fitted on the 193 training streams scored highest, and it is the held-out row in Table 3.

Whole-video reading. The reader receives all 1-fps frames of the stream in one call. If a stream has more than 2,048, they are chosen uniformly, and the count is kept even. Frames are made upright and placed on 448×448 squares, the prompt is the one in Table S4, decoding is greedy, and the answer is the first integer.

Sum over visits and its corrections.

- *Sum over visits* reads each true visit with PlaceTally's count and gate reads, and adds the counts.
- *The fixed correction* multiplies this sum by one constant and rounds the result. The constant was chosen from 0.30 to 1.30 in steps of 0.01 to maximise strict MRA on the 48 streams of the first training set, which gives 0.69.
- *The constant for a mix of return rates* (0.81) was chosen the same way, on those streams at the three rates of Sec. 5.2.
- *The per-video rate* multiplies the sum by the number of PlaceTally's places among the visits, divided by the number of visits, and rounds the result (halves to even). Each visit counts in the place that holds most of its frames.

All four start from true visit boundaries, which makes them privileged.

Oracles. *PlaceTally on true visits* replaces its fragments with the true visits and leaves the rest of the method unchanged, so its places are the true visits joined by the return evidence. The second oracle uses the same places but does not read a place as one video. It reads each visit separately, keeps the largest gated count among the visits of each place, and sums over places. Both take the visit boundaries from the truth; neither uses the true room identities. On ARKit test the second oracle scores 63.6% [60.0%, 67.1%], against 65.1% [61.6%, 68.6%] for the first; on 3RScan test, 66.4% [61.8%, 70.9%] against 65.4% [60.9%, 70.0%].

Cambrian-S. We ran the released Cambrian-S-7B [77] with its official code, in its own environment and in both of its modes: reading the whole video, and the surprise-segmented memory it uses for counting. It processes all 1-fps frames through its own released pipeline, whereas our readers see 448-pixel letterboxed frames. On all 460 test questions it scores 34.1% and 2.1% (Table 3). A second run under a protocol fixed in advance, on 150 test questions, gives 31.7% [25.0%, 38.6%] and 1.7% [0.2%, 3.6%], against 59.4% [53.0%, 65.8%] for PlaceTally on the same questions.

CountVid. We ran the released CountVid [2] with its own settings on a fixed sample of 40 test questions, stratified

Read	Prompt
Count, per place	*This clip is one walk through a single room or space, filmed with a handheld camera.* [video] *How many different {plural} are in this space? Count each one once, even if it appears in several frames. Answer with a single number. If there are none, answer 0.*
Presence gate, per place	*This clip is one walk through a single room or space, filmed with a handheld camera.* [video] *Is there a {name} in this space? Answer yes or no.* ("an" before a vowel)
Whole-video reading	[video of all 1-fps frames, at most 2,048] *How many different {plural} have you seen in the whole video so far? Count each physical object only once, even if you saw it more than once. Answer with a single number.*

Table S4. Reader prompts, verbatim. {name} is the category's name, for example "table" or "TV or computer monitor" (the 14 categories are listed in Sec. A), and {plural} its plural, formed by adding "s". Whole-video reading uses an earlier list of plurals that differs in one category: "TV or monitors".

by category with a fixed seed. It reads complete streams at 3 fps, with the category given as a text prompt before the video. Each question may run for eight hours; a question that reaches this limit counts as zero, and four did. Each completed question took about 3.1 GPU-hours. Its published demonstration counts reproduce (3.67 against 4.0, and 69.14 against 69.1). To see whether returns or length explain its over-counting, we also ran it inside single true places on 24 training questions. It over-counts all of them, by a median factor of 15.4.

DirectMe. We ran the released DirectMe [73] code (UCS-Bench repository, commit 13261b0) on the 40 questions used for CountVid. We followed the graph-building route its README gives for evaluation: Depth Anything 3 for depth and camera pose, YOLO-World with the first 300 Objects365 [62] names, SAM 2.1 masks, ten-frame chunks and the released thresholds. It reads each stream at its documented 1 fps and builds one map per stream before any question. Its own counting rule answers at the end of the stream: the answer is the number of map entries whose label matches the asked category. Its vision-language answer path chooses among multiple-choice options, so it cannot give the open counts our questions ask for. We added fireplace to its detector classes and our category words to its alias table through its documented function, kept its stock aliases, and asked about our TV-or-monitor category as "TV or computer monitor". A first pass at 3 fps was stopped before any answer was scored, after 11 of its first 14 streams reached the two-hour build limit; the change to 1 fps was registered that day (Table S21). One stream reached the four-hour limit and counts as zero. The run used 55.5 GPU-hours, the stopped pass included.

DirectMe scores 0.0% on all 40 questions, against 57.5% [46.7%, 68.5%] for PlaceTally. Of its 39 answers, 38 are too high, at a median of 7.7 times the truth, and none is within half of the truth. Counting the matching entries in its saved maps reproduces all 39 answers. Its maps hold far more entries than objects, for two reasons that the saved keyframes show: the same object is stored several times, and other things get the asked label. In one stream, the only bed has seven entries from a single two-minute recording: five show that bed from different sides, up to 4.5 m apart, and two sit on other things. In another stream, only one of the eight oven entries clearly shows the oven. This happens before any return: inside the first recording of each stream, 11 of the 15 questions with matching entries there already have more entries than the whole stream's true count, and only 24 of the 1,134 matching entries span two visits. Its stock aliases also count chairs as stools; with stools alone, one stool question earns partial credit and the score is 0.5%. Building a map takes 254 GPU-seconds per minute of video on one A100-40GB, about 60 times as long as PlaceTally (Table 5). The map file takes 1.0 MB per minute, but it points to stored frames for 58 of every 60 frames; with them, it keeps 3.9 MB per minute as JPEG and 24 MB as run. An answer takes 32 ms on a CPU. The authors' demonstration script uses another reconstruction model (Scal3R), 60-frame chunks and a larger detector; we did not run that setting.

VideoAgent. We used its official code [24], with two substitutions:

- Qwen3.5-9B (the same snapshot as our reader) replaces GPT-4 as the controller;
- a local sentence embedder (all-mpnet-base-v2) replaces the OpenAI text embeddings.

Its captioner, detector, tracker, re-identification, prompts and thresholds are unchanged. It reads our 5-fps frames, where its default is 15 fps.

After the first test reads we amended its answer parser to accept number words from zero to twenty and to retry on parser errors, and we re-read every answer. Then 452 of the 460 questions give a number, and the other eight score zero.

Its detector's vocabulary covers 9 of our categories directly; it has no label for bathtub, fireplace, stool, stove or washing machine. On first visits alone, it creates 7,442 identities for 3,055 true objects in those nine categories, 2.4 per object.

Other readers. Qwen3.5-35B-A3B, a mixture-of-experts model that activates about 3B parameters per token, reads the whole video with the same frames and prompt as Qwen3.5-9B; it scores 42.4% [38.1%, 46.9%] on the test split, against 51.2% for Qwen3.5-9B. Gemma 4 12B [28] replaces Qwen3.5-9B in both PlaceTally and whole-video reading with nothing refitted: the gate threshold stays at 0.649, and whole-video reading keeps its frame rule.

Methods we could not run. StreamMem [79] has no public code or weights; we checked its project pages and paper. Spatial-TTT [44] has released only its smallest checkpoint, Spatial-TTT-nano. We ran it with the official code, without changing its model or cache code, on one training stream, in four environments:

- with the package versions bundled with the code, the code could not load its Qwen3-VL backbone;
- with the versions its README names, it stopped with an error in its cache code;
- with two later Transformers releases, it stopped with an error in its video-feature code.

We therefore report no Spatial-TTT result.

Shot detector. The TransNet V2 [65] row of Table 4 uses the official weights and settings: 100-frame windows and a threshold of 0.5. It runs on our upright frames at 5 fps.

D. Additional Results

D.1. Answers by True Count

Table S5 splits the answers by true count. On ARKit test, PlaceTally's mean answer rises from 1.2 to 9.6 across the bins (slope 0.72), while whole-video reading's rises only from 1.5 to 5.4 (slope 0.32). Whole-video reading is best at a true count of 2, where a small answer is right (84.6%), and scores 10.5% at 9 or more. At 9 or more, PlaceTally leads it by 45.8% [37.1%, 53.9%] on the same 38 questions. The larger Qwen3.5-35B-A3B is flatter still (slope 0.26), and with Gemma 4 the same pattern holds (0.67 against 0.40). On 3RScan, whole-video reading leads overall because two thirds of the true counts are 1 or 2, but PlaceTally leads at 5 or more.

D.2. Return Rate in Full

Table S6 gives every line of Fig. 3a with its interval, and adds a fourth rate. From 60 test streams drawn at random we kept the first visit to each room (rate 0), these and half of the returns, drawn with a fixed seed (0.15), all visits (0.31, the benchmark's rate), or only the rooms that return (0.50). Fig. 3a counts returns instead: none, one (half of the returns is one return in each of the 60 streams) or all, two or three per stream (2.52 on average). The truth was recomputed with the same rule on the kept visits, and the same 291 questions remain at every rate. The place finder ran unchanged on each reduced stream, with place F1 97.3% [95.1%, 98.9%] on first visits alone and 97.6% [95.6%, 99.0%] with half the returns. PlaceTally's answers at 0 and 0.15 are reads of the reduced streams with the frozen method, and at 0.31 they are its stored test answers. At 0.50 it was not read, so Fig. 3a stops at all returns (0.31) and the table gives a stand-in, marked as such. The last two rows rescale the sum over visits (Sec. C): by 0.81, a constant fitted on training streams at the first three rates, and by each stream's own ratio of PlaceTally's places to kept visits. On the reduced streams, each kept visit counts in the place that holds most of its frames in PlaceTally's graph of the full stream. Pooled over the first three rates (873 question-rate pairs), the constant fitted on a mix is 9.2% [5.6%, 12.9%] below PlaceTally, and the per-video rate 3.2% [0.3%, 6.1%] below.

Table S7 shows how each method errs. The sum over visits is too high on 18% of questions without returns and on 83% at 0.50, while the fixed correction is too low on 70% without returns. PlaceTally is too high on about a fifth of questions at every rate, and whole-video reading is too low on 43 to 47%.

D.3. Fixed Correction on Long Streams

The long streams keep the benchmark's return rate, about a third (Table S1), and there the fixed correction ($\times 0.69$) stays level with PlaceTally: -4.2% [-8.5%, 0.1%] at 30 minutes and $+0.1\%$ [-6.1%, 6.3%] at 60 (Table S8), as Sec. 5.2 expects.

Whole-video reading's small answers can also be corrected after the fact. A map fitted on the 694 questions of the training streams sends each answer to the value with the highest strict MRA there: $1 \to 1$; $2 \to 2$; $3 \to 4$; $4 \to 8$; $5 \to 8$; 6 or $7 \to 10$; 8 to $10 \to 17$; other answers are left unchanged. It was fitted answer by answer, so it need not increase. It lifts whole-video reading from 51.2% to 54.8% on test, from 41.3% to 44.1% at 30 minutes and from 26.0% to 48.4% at 60 minutes, still 24.1% [17.2%, 30.7%] and 17.2% [8.4%, 26.3%] below PlaceTally at 30 and 60 minutes.

D.4. Long Streams in Full

Table S8 gives every method and blind baseline on the long streams. As the streams grow from 11 to 60 minutes, PlaceTally stays between 62.5% and 68.2% and place F1 between 96.8% and 98.3%. Whole-video reading falls from 51.2% to 26.0%. The blind baselines fall further: the constant and the per-category prior, fitted on streams of about 11 minutes, drop to 8.8% and 7.6% at 60 minutes, and the baselines that scale with the stream reach 19.3% at best. Because they reuse test rooms seen during the failure analysis, the long streams are not a held-out test.

True count	1	2	3–4	5–8	9+	Slope	Overall
ARKit test	*n=102*	*n=117*	*n=139*	*n=64*	*n=38*		
PlaceTally, Qwen3.5-9B	69.6 (1.2)	64.1 (2.1)	61.0 (3.2)	55.2 (5.1)	56.3 (9.6)	0.72 [0.65, 0.80]	62.5 [58.4, 66.5]
Whole-video reading, Qwen3.5-9B	56.9 (1.5)	84.6 (2.1)	47.7 (2.5)	13.1 (3.2)	10.5 (5.4)	0.32 [0.27, 0.39]	51.2 [46.9, 55.8]
Whole-video reading, Qwen3.5-35B-A3B	17.6 (1.9)	73.5 (2.3)	56.2 (2.8)	18.3 (3.4)	3.4 (5.0)	0.26 [0.22, 0.29]	42.4 [38.1, 46.9]
PlaceTally, Gemma 4 12B	57.8 (1.5)	49.6 (2.2)	52.3 (3.7)	48.1 (5.8)	44.2 (9.2)	0.67 [0.59, 0.76]	51.6 [47.2, 55.6]
Whole-video reading, Gemma 4 12B	82.4 (1.1)	10.3 (1.4)	13.4 (2.2)	5.5 (3.1)	26.3 (5.9)	0.40 [0.31, 0.51]	27.8 [23.8, 32.1]
3RScan test	*n=144*	*n=98*	*n=77*	*n=21*	*n=22*		
PlaceTally, Qwen3.5-9B	50.0 (1.6)	38.8 (2.4)	53.4 (3.7)	43.8 (5.3)	47.3 (9.8)	0.72 [0.62, 0.80]	47.2 [41.7, 52.9]
Whole-video reading, Qwen3.5-9B	46.5 (1.6)	83.7 (2.1)	59.6 (2.7)	16.2 (3.3)	5.0 (5.6)	0.34 [0.30, 0.39]	55.1 [49.9, 60.1]

Table S5. Answers by true count: strict MRA (%), with the mean answer in brackets, for each bin of true counts; the slope of the answer regressed on the truth, and overall MRA, with 95% intervals over streams (5,000 draws). A slope of 1 means answers grow as fast as the truth.

Method	0	0.15	0.31	0.50
PlaceTally	64.4 [59.5, 69.1]	62.2 [56.3, 67.7]	61.9 [56.1, 67.4]	64.8 [59.8, 69.8]†
Whole-video reading	52.3 [46.5, 57.9]	50.6 [44.9, 56.4]	51.4 [45.8, 57.1]	–
Privileged: true visits				
Sum over visits	65.6 [60.7, 70.2]	49.5 [43.9, 54.9]	25.2 [21.2, 29.4]	13.8 [10.2, 17.4]
Fixed correction, $\times 0.69$	38.5 [33.3, 44.1]	49.2 [44.5, 54.0]	61.4 [56.2, 66.7]	56.5 [51.1, 62.0]
Constant fitted on a mix, $\times 0.81$	59.1 [54.1, 64.0]	55.5 [49.6, 61.2]	46.3 [40.6, 52.2]	–
Per-video rate	63.8 [57.6, 69.3]	54.6 [47.9, 61.2]	60.3 [54.9, 65.8]	59.9 [54.3, 65.1]

Table S6. Return rate in full: strict MRA (%) with 95% intervals on the same 60 test streams and 291 questions at every rate (column heads give the mean share of visits that are returns; 0.31 is the benchmark's rate). †At 0.50 PlaceTally was not read; the row is a stand-in that reruns the place finder on the reduced stream and sums the largest stored per-visit count in each place. At the other three rates the stand-in minus PlaceTally's actual answers is +0.2 [−2.3, 2.7], +1.2 [−1.0, 3.6], +0.4 [−2.3, 3.2]. –: not stored.

D.5. Stream Size and Number of Rooms

On ARKit test, PlaceTally's lead over whole-video reading grows with the stream: from 2.0% under 10 minutes to 23.9% at 15 minutes or more, and from 5.4% with four rooms to 15.8% with six (Table S9). The shortest streams and those with four rooms are the only groups whose intervals include zero. On the held-out split the lead is 14% to 19% in every group.

D.6. Accuracy against Input

Table S10 gives the accuracy against input at 11, 30 and 60 minutes (Fig. 3c shows 11 minutes), with a sweep on training streams in which the gate threshold is not refitted. There, PlaceTally rises from 47.4% to 64.2% as its frames per place grow from 16 to 128, and whole-video reading from 52.0% to 57.8% as its token cap grows eightfold. At its largest cap whole-video reading reads 2.4 times PlaceTally's input and still scores lower, and its GPU time per question is five times larger. Paired on the same questions, 128 frames per place score 3.6% [0.0%, 7.0%] and 3.4% [−1.3%, 7.6%] above 64 at 30 and 60 minutes, and 0.1% [−2.8%, 2.6%] below it at 11 minutes.

D.7. Finding Places

This part gives the studies behind Sec. 5.4 in full: the return evidence (Table S11), the pair statistics behind the best-link rule (Table S12) and the solver (Table S13). A control for the frames at recording ends follows (Table S14).

Return evidence. Table S11 changes only the return evidence. Refitting the SigLIP 2 threshold on training streams moves place F1 by at most 1.3%. MegaLoc's chamfer similarity [9] helps on ARKit test and is level on 3RScan. MegaLoc's single best pair of frames does not help on ARKit test and hurts on 3RScan, by 5.5%.

Why one best link. Table S12 gives the counts behind the percentages in Sec. 5.4. A pair above τ nearly always shows one place, while a pair below it says little: 42% of true-return pairs on ARKit test and 67% on 3RScan fall there. For a returning fragment, the best match to the earlier recording of its room clears τ far more often than the mean or the weakest match, most clearly on 3RScan.

The solver. The solver of Eq. 2 in Table 4 is greedy additive edge contraction with local moves [38]; it is optimal on 188 of 189 short training streams. Scored by Eq. 2, the places of Eq. 3 reach 72% of the solver's value on ARKit test and fall below zero on 3RScan, although their F1 is about the same (Table 4): the solver's F1 minus PlaceTally's is −0.3% [−1.9%, 1.2%] on ARKit test and +1.1% [−3.4%, 5.2%] on 3RScan. Table 4 scores places on all 109 and 150 test streams; counting uses the 103 and 134 that have questions. On both test splits the solver finds partitions that score higher

Low / exact / high	0	0.15	0.31	0.50
PlaceTally	30 / 51 / 20	32 / 48 / 20	33 / 47 / 20	21 / 59 / 20[†]
Whole-video reading	43 / 43 / 13	46 / 41 / 13	47 / 42 / 12	–
Privileged: true visits				
Sum over visits	30 / 52 / 18	20 / 36 / 44	11 / 13 / 76	7 / 10 / 83
Fixed correction, ×0.69	70 / 28 / 2	54 / 36 / 9	31 / 48 / 21	16 / 52 / 32
Constant fitted on a mix, ×0.81	47 / 45 / 8	38 / 41 / 21	22 / 30 / 48	–
Per-video rate	33 / 51 / 16	34 / 41 / 25	32 / 47 / 21	29 / 54 / 17

Table S7. Shares of answers below, equal to and above the truth (%), for the questions of Table S6; rows may not sum to 100 because of rounding. [†]Stand-in, as there.

	Test (11 min)	30 min	60 min
Streams	103	38	18
Place F1 (merges, splits)	96.8 [94.8, 98.3] (23, 21)	98.3 [97.2, 99.2] (16, 18)	97.7 [96.1, 98.9] (19, 18)
PlaceTally	62.5 [58.4, 66.5]	68.2 [63.2, 73.2]	65.6 [58.3, 72.6]
Whole-video reading	51.2 [46.9, 55.8]	41.3 [35.7, 47.0]	26.0 [20.5, 32.5]
Whole-video reading, calibrated	54.8 [50.9, 58.8]	44.1 [39.4, 48.9]	48.4 [41.2, 55.7]
Privileged: true visits			
PlaceTally on true visits	65.1 [61.6, 68.6]	67.6 [62.9, 72.3]	67.2 [60.1, 73.9]
Sum over visits	26.9 [23.5, 30.5]	36.0 [30.1, 41.7]	38.4 [32.5, 45.1]
Fixed correction, ×0.69	58.6 [54.8, 62.7]	64.0 [58.2, 69.7]	65.7 [59.9, 71.8]
Blind baselines, fitted on training streams			
Constant	32.7 [28.9, 36.6]	22.0 [17.5, 26.9]	8.8 [4.9, 12.8]
Per-category prior	41.6 [37.7, 45.4]	22.0 [17.1, 27.3]	7.6 [3.6, 11.9]
Rate × visits, mean	41.0 [37.1, 44.8]	30.8 [25.8, 35.7]	16.5 [9.0, 25.9]
Rate × visits, median	41.1 [37.2, 44.9]	30.6 [25.8, 35.5]	19.3 [11.6, 29.1]
Rate × seconds, mean	36.9 [33.2, 40.3]	25.2 [19.8, 30.5]	10.9 [3.3, 20.6]
Rate × seconds, median	37.3 [33.7, 40.6]	29.9 [24.3, 35.3]	17.5 [8.9, 27.7]
Rate × PlaceTally's places, mean	39.8 [35.9, 43.4]	29.7 [24.6, 34.9]	14.6 [7.6, 23.5]
Rate × PlaceTally's places, median	38.6 [34.9, 42.2]	30.2 [24.9, 35.4]	18.7 [11.2, 27.7]
Rate × true rooms, mean[§]	40.6 [36.7, 44.4]	29.8 [24.9, 34.8]	14.9 [7.9, 23.3]
Rate × true rooms, median[§]	39.7 [35.9, 43.4]	30.1 [25.1, 35.2]	18.8 [11.2, 28.2]

Table S8. Long streams in full: strict MRA (%) with 95% intervals over streams. The long streams reuse the test rooms (Sec. A), so length and the number of rooms grow together. Place F1 (%) is on the streams with questions. [§]Uses the true number of rooms.

on Eq. 2 than the true places (Table S13), so the true places do not maximise Eq. 2, and optimising it harder need not lead to them. Only 7 of 992 fragments on ARKit test, and 162 of 3,393 on 3RScan, have two or more earlier places above τ; only for these could a rule that weighs all links decide differently from the best link.

Recording ends. Because each stream joins whole recordings, a return might be recognised from the frames where a recording starts or stops, which may look alike. The best-matching pair of frames lies within the first or last five frames of both recordings in 29.5% [24.0%, 34.9%] of true-return pairs above τ, against 3.8% [3.0%, 4.7%] if those pairs were spread evenly. Removing these frames from the return evidence lowers place F1 only from 96.9% to 96.5%, about as much as removing as many frames from the middle (Table S14). With 20 frames removed at each end, fewer streams get the right number of places than with 20 removed from the middle (52.3% against 63.3%), so these frames do help some streams.

Cost of an update. From cached features, adding a fragment to the graph takes on average 1.2 ms on CPU with fewer than 128 earlier frames and 3.6 ms with 512 to 1,024 on ARKit test, and 11.1 ms with 2,048 to 4,096 on 3RScan test. Decoding, features and the detector are not included.

D.8. VSI-SUPER-Count and Rendered Walks

VSI-SUPER-Count [77] joins clips from several datasets into videos of 10 to 120 minutes. It guided our failure analysis (Sec. 4) and is therefore not a held-out test. On it the frozen PlaceTally scores 21.0% to 37.0% (Table S15), below priors that multiply the video's length or its number of source clips by a rate, which reach 93.4% at 120 minutes, and below reading between the known clip seams; these priors were fitted on the other lengths of the same benchmark. Its place finder joins clips of different sources: against the source clips as nominal places, its precision is 42 to 71% while its

	Streams	Questions	PlaceTally	Whole-video	Difference
ARKit test					
Four rooms, two returns	52	202	61.1 [54.5, 67.2]	55.7 [49.1, 62.3]	+5.4 [−4.2, 14.5]
Six rooms, three returns	51	258	63.6 [58.3, 68.4]	47.8 [42.2, 53.6]	+15.8 [8.0, 23.2]
Under 10 minutes	42	161	59.0 [51.6, 66.5]	57.0 [49.1, 64.7]	+2.0 [−8.7, 12.2]
10 to 15 minutes	48	250	63.4 [58.3, 68.2]	48.7 [42.9, 54.5]	+14.8 [6.7, 22.6]
15 minutes or more	13	49	69.2 [55.8, 79.8]	45.3 [33.5, 56.4]	+23.9 [12.0, 34.9]
ARKit held-out					
Four rooms, two returns	42	186	63.0 [58.2, 68.1]	49.0 [42.1, 56.5]	+14.0 [5.3, 21.9]
Six rooms, three returns	45	217	59.7 [54.5, 65.1]	40.3 [35.5, 45.5]	+19.4 [12.5, 26.2]
Under 10 minutes	31	122	69.3 [63.4, 75.3]	52.5 [45.9, 60.2]	+16.7 [8.9, 24.1]
10 to 15 minutes	39	193	58.7 [53.6, 64.0]	40.8 [34.2, 48.0]	+17.9 [8.9, 26.0]
15 minutes or more	17	88	55.7 [49.1, 63.3]	40.8 [33.3, 49.0]	+14.9 [3.7, 26.9]

Table S9. Stream size and number of rooms: strict MRA (%) with paired 95% intervals over streams. Streams that end in a new room have one more room than named.

	MRA (%)	Input k tokens	GPU-s
Training streams, gate fixed			
PlaceTally, 16 frames per place	47.4 [43.2, 51.8]	18.5	0.55
PlaceTally, 32 frames per place	57.6 [53.8, 61.6]	36.1	0.94
PlaceTally, 64 frames per place	61.8 [57.8, 65.6]	65.6	1.66
PlaceTally, 128 frames per place	64.2 [60.0, 68.3]	100.6	2.59
Whole-video reading, 32k-token cap	52.0 [48.1, 55.8]	30.7	1.07
Whole-video reading, 64k-token cap	53.6 [49.7, 57.4]	62.0	2.37
Whole-video reading, 128k-token cap	54.9 [50.9, 58.8]	124.4	5.61
Whole-video reading, 256k-token cap	57.8 [53.6, 61.6]	237.1	13.63
Test and long streams, gate refitted for 16 and 32 frames			
PlaceTally, 16 frames per place, test	50.1 [45.8, 54.1]	19.0	–
PlaceTally, 32 frames per place, test	54.1 [49.9, 58.2]	37.1	–
PlaceTally, 64 frames per place, test	62.6 [58.6, 66.7]	68.3	–
PlaceTally, 16 frames per place, 30 min	52.4 [47.0, 57.6]	49.3	1.68
PlaceTally, 32 frames per place, 30 min	57.2 [52.3, 62.1]	96.5	1.97
PlaceTally, 64 frames per place, 30 min	64.6 [58.8, 70.0]	179.1	2.90
PlaceTally, 16 frames per place, 60 min	55.6 [43.8, 67.3]	96.4	4.22
PlaceTally, 32 frames per place, 60 min	58.2 [51.6, 65.6]	188.5	4.94
PlaceTally, 64 frames per place, 60 min	62.2 [53.4, 70.8]	347.5	7.47

Table S10. Accuracy against input per question at each length (Fig. 3c shows 11 minutes); its 128-frame points on test and long streams are in Tables S8 and S18. GPU-s: reader GPU time per question, batched requests sharing a batch's time equally. Tokens include the count and gate calls. –: not stored.

recall is 92 to 100%. Some of these joins pair clips whose physical identity could not be checked, such as ScanNet++ scenes or clips from different datasets, so the precision may be underestimated.

Three learned joining variants, built during the failure analysis to fix these joins, score 45.0% to 53.6% overall. The tree variant was then tested on 3RScan, where it made fewer false merges but more false splits and counted worse, so PlaceTally kept its frozen place finder.

Rendered walks. We also tried rendered video as a source of returns. On 12 houses of HM3D [58], which are scanned buildings, and 24 groups of synthetic HSSD homes [39], we rendered continuous 12-minute walks that return to rooms: 23 scored walks with 130 questions on HM3D and 48 walks with 250 questions on HSSD, over the same 14 categories. As in Sec. A, an object counts if it covers 0.5% of the image in two frames of a visit, and every question has an object seen on two visits. We had decided before scoring to drop a dataset if a blind baseline fitted on its walks reached 60%. On both, the per-category prior did, with 70.5% and 69.8%, and fitted on the other houses it still scores 58.4% and 67.1%, because the answer is 1 in 61 of the 130 and 133 of the 250 questions. We therefore did not run PlaceTally on these walks, opened none of the houses set aside for testing, and left both datasets out of RevisitCount. Rendered walks through static scenes would also not test captured camera motion.

D.9. Headroom in the Reader

On all 65 streams of the first training set, with true places (226 questions), the reader scores 66.5%, and exact counts in every place that passes the gate would give 89.3% (Sec. 5.4). Three checks place this loss in counting inside a place. First, truth recomputed from each place's own frames equals the place truth for 99.5% of places. Second, counts from the reader on the 1, 3 or 5 best photos of a place, chosen with the truth, give 64.1%, 66.4% and 67.6%, against 66.5% for the usual frames. Third, SAM 3 boxes nearly every true object: on sampled photos, perfect identity and rejection of its boxes would reach 79.9%, against 81.7% for all true objects on the same photos.

Table S16 lists the routes we tried to close this gap. None improved counting by the margin set for it in advance. The comparison with Molmo2 [17] had no margin set in advance. Molmo2 over-counts places with a single object (62 of 315, against 21 for our reader), and with two or more objects it errs both ways (57 too high and 55 too low of 144).

Localisation maps. This study followed a suggestion to give the reader maps of where the target objects are. SAM 3 was chosen over CountGD++ [3] in a first stage, because both found the targets (93.7% and 93.5% of them) but SAM 3

	ARKit test (598 true places)							3RScan test (824)						
Return evidence	τ	F1	Places	Exact	Merges	Splits	ΔF1	τ	F1	Places	Exact	Merges	Splits	ΔF1
SigLIP 2 (PlaceTally)	0.864	96.9 [95.1, 98.4]	594	85	24	21	–	0.864	75.3 [72.2, 78.4]	752	42	285	257	–
SigLIP 2, refitted	0.872	97.0 [95.2, 98.5]	598	83	23	24	+0.1 [−0.1, 0.3]	0.872	74.0 [70.8, 77.0]	797	38	290	282	−1.3 [−2.7, 0.1]
MegaLoc, chamfer	0.204	97.6 [95.9, 98.9]	585	92	21	9	+0.7 [0.1, 1.5]	0.204	76.5 [73.2, 79.9]	738	36	277	241	+1.2 [−2.0, 4.4]
MegaLoc, best pair	0.598	96.3 [94.1, 98.2]	600	84	22	24	−0.6 [−2.6, 0.8]	0.598	69.8 [66.3, 73.4]	936	40	305	361	−5.5 [−9.6, −1.8]

Table S11. Return evidence in full. Only the return evidence and its threshold change; cuts and the continuation model are PlaceTally's. Every threshold other than PlaceTally's τ was fitted on training streams and sealed before these test runs. F1 in %. Exact: streams with the right number of places. ΔF1: pooled place F1 minus PlaceTally's, with paired 95% intervals.

	ARKit test	3RScan test
Pairs above τ that show the same place	96.0 [93.6, 98.1] (315/328)	90.7 [87.5, 93.8] (2,734/3,013)
True-return pairs below τ	41.6 [29.5, 52.1] (215/517)	66.7 [63.5, 69.6] (3,075/4,609)
Pairs from one recording below τ	94.7 [90.8, 97.3] (232/245)	76.8 [73.8, 79.5] (3,974/5,174)
Pairs from different places above τ	0.3 [0.2, 0.5] (13/3,915)	0.8 [0.5, 1.1] (279/34,821)
Returns whose best match clears τ	85.1 [80.5, 89.7] (275/323)	71.4 [67.2, 75.4] (755/1,058)
Returns whose mean match clears τ	74.3 [67.7, 81.0] (240/323)	36.2 [32.2, 40.4] (383/1,058)
Returns whose weakest match clears τ	72.1 [65.1, 79.3] (233/323)	22.2 [18.8, 26.1] (235/1,058)

Table S12. The return evidence is one-sided (Sec. 5.4). Pairs are pairs of fragments that are not neighbours, scored with PlaceTally's SigLIP 2 chamfer similarity and $\tau = 0.8637$; the last three rows take, for each returning fragment, its best, mean or weakest match to the fragments of the earlier recording of its room. Percentages with 95% intervals over streams.

	ARKit test	3RScan test
Solver of Eq. 2	1,447	9,341
Greedy merging	1,041	1,131
True places	1,040	4,168
PlaceTally	1,038	−98
PlaceTally as a share of the solver	72%	−1%
Solver − PlaceTally, per stream	3.8 [1.9, 6.0]	62.9 [50.4, 77.0]
Fragments with $\geq$2 earlier places above τ	7 of 992	162 of 3,393

Table S13. The objective of Eq. 2 summed over streams, with every partition scored on the same signed log-odds weights. The solver is greedy edge contraction followed by local moves; it has no certificate of a global optimum. True places assign each fragment to the room that fills most of it. Only when two or more earlier places clear τ can the best link differ from another choice.

Frames removed	Masked	F1	Exact	Returns kept
None	0.0	96.9 [95.1, 98.4]	78.0 [69.7, 85.3]	58.4 [47.9, 70.5]
First and last 5	12.3	96.5 [94.7, 98.0]	74.3 [66.1, 82.6]	55.5 [44.9, 67.9]
First and last 10	24.5	96.4 [94.7, 97.9]	65.1 [56.0, 74.3]	51.3 [41.4, 63.2]
First and last 20	47.9	95.1 [93.4, 96.6]	52.3 [43.1, 61.5]	43.7 [35.5, 53.8]
5 from the middle	11.7	96.8 [95.0, 98.3]	71.6 [62.4, 79.8]	54.7 [44.9, 66.4]
10 from the middle	23.9	96.7 [94.9, 98.2]	69.7 [60.6, 78.0]	50.7 [41.4, 61.9]
20 from the middle	47.3	95.8 [94.1, 97.3]	63.3 [54.1, 72.5]	48.2 [39.7, 58.2]

Table S14. Do return links rely on the frames at the start and end of each recording? On ARKit test, the frames named in the first column (per recording, at 1 fps) are removed from the return evidence only; cuts and continuation are unchanged. F1 in %. Masked: share of frames removed (%). Exact: streams with the right number of places (%). Returns kept: true-return pairs still above τ (%).

put much less of its map on wrong objects (29% against 66% on sampled frames). Its maps then went to the reader in three ways: as a bias on attention, drawn over the frames, or as crops. The settings were fixed on 141 questions of the first training set and tested once on the 553 questions of the second. All three scored below PlaceTally without maps, with intervals that include zero, and none reached the +3% margin, so none was run on test streams. The study used about 10 GPU-hours.

D.10. Object Support

Counting by place adds counts correctly only if places share few objects. On ARKit test, 1.9% of the supported objects appear in two or more of PlaceTally's places, and 0.8 to 4.2% on the training streams (Table S17). On 3RScan it is 31 to 33%, where the places split many returning rooms (257 splits on test, Table S11).

The 3RScan rescans also change the rooms. Over the 903 pairs of a first scan and its rescan, 3,965 target objects of the first scans are also annotated in the rescan, 822 are not (82 of them listed as removed), and 217 appear only in the rescan; 40.6% of the shared objects were moved.

D.11. Cost Details

Table S18 extends Table 5 to the long streams. PlaceTally keeps about 2.3 MB per minute of video at every length, because its memory grows with the number of places. Whole-video reading keeps less per minute at 60 minutes, where its cap of 2,048 frames drops frames.

The construction cost in Table 5, 4.0 GPU-s per minute of video or 15 times faster than real time, was timed on one A100-40GB GPU on 5 test streams, and on 2 streams

	10 min	30 min	60 min	120 min	All
PlaceTally	32.2	21.0	32.8	37.0	30.8
benchmark's own metric	33.6	22.0	33.0	37.4	31.5
variant: linear joining	48.6	36.8	43.6	50.8	45.0
variant: tree joining	49.0	40.4	51.0	74.2	53.6
variant: hybrid joining	49.0	39.4	50.2	73.0	52.9
Reading between clip seams*	53.6	47.8	55.0	78.6	58.8
Reading by source recording*	49.8	–	59.8	79.8	63.1
Prior × length	42.2	51.6	80.8	93.4	67.0
Prior × source clips	45.4	60.6	84.2	90.6	70.2
Prior × PlaceTally's places	38.2	55.0	62.8	75.4	57.8
Place precision of PlaceTally	71 [61, 81]	42 [36, 50]	45 [40, 52]	46 [42, 51]	
Place recall of PlaceTally	100 [100, 100]	92 [89, 95]	100 [100, 100]	100 [100, 100]	

Table S15. VSI-SUPER-Count: strict MRA (%) on 50 questions per length. The row below PlaceTally uses the benchmark's own metric, which counts boundary cases as correct; under it the priors score the same, and no other entry changes by more than 1.4. Variants were built during the failure analysis and are not the frozen method. Priors multiply a count from the video by a rate fitted on the other lengths. *Privileged: uses the clip boundaries or source recordings known from how the videos were built; the prior on source clips uses their true number. Place precision and recall (%) are pair-based, against the source clips as nominal places.

Route	Best result against our reader
Open-vocabulary detectors as counters (SAM 3, CountGD [4], OWLv2, Grounding DINO, each with or without SAM 2.1)	none is better per photo; at two objects, SAM 3 is exact on 0.475 of photos against 0.495
Another reader (Gemma 4 12B, Gemma 4 26B-A4B)	−5.8 to −9.3 MRA inside places
Prompting (2× upscaling, thinking, listing before counting)	+4.9 exact at two objects per photo with upscaling; thinking doubles false objects
A detector proposes boxes and the reader judges them (crops, one box, numbered marks, hints)	the reader's judgements separate real from wrong boxes with AUROC 0.867, against 0.873 for SAM 3's own scores; marks split single objects in 18% of cases
Fine-tuning the reader for counting (LoRA)	+1.4 on the first training set, by moving its prior over answers
A probe on the reader's hidden state	−2.1 [−5.2, 0.8] on the second training set; without feature scaling (chosen afterwards) +2.8 [0.6, 5.2], still below the +3 margin
A learned judge of boxes on frozen features	keeps 32% of real boxes at 2% wrong ones, where 90% was needed
Molmo2-8B, video pointing with persistent IDs	54.2 [48.3, 59.7] against 66.5: −12.3 [−18.4, −6.7]
SAM 3 localisation maps given to the reader: attention bias, overlay, crops	−1.81 [−4.17, 0.52], −2.08 [−4.73, 0.55], −1.59 [−4.11, 0.98] on 553 questions of the second training set; margin +3

Table S16. Routes tried for counting inside a place, all on training streams, with true places unless stated. MRA and its differences are in %; differences are paired, with 95% intervals where computed. No route reached its margin, so none was adopted.

	Questions	Supported	Shared	Shared %
First training set (fitted on)	141	502	4	0.8 [0.0, 2.0]
Second training set	553	1,821	76	4.2 [2.1, 6.6]
3RScan training	42	85	28	32.9 [19.1, 51.2]
ARKit test	460	1,663	32	1.9 [0.7, 3.7]
3RScan test	362	1,009	314	31.1 [27.2, 35.3]

Table S17. Objects seen in more than one of PlaceTally's places. Supported: distinct objects that pass the two-frame rule in the questions' streams and categories; shared: those supported in two or more places (%, with 95% intervals over streams).

each of 30 and 60 minutes (3.9 to 4.0 GPU-s per minute at every length and every K). It includes decoding, features, grouping and writing the memory. Building an 11-minute video costs about 45 GPU-s, paid once. Against whole-video reading's 15.9 GPU-s per question, PlaceTally uses less GPU time in total from the sixth question on a video with 64 frames per place and from the tenth with 128; at the benchmark's 4.5 questions per video, whole-video reading uses less in total. Query times in Table 5 were measured with nothing cached, retries included. The JPEG sizes for VideoAgent and CountVid are estimates from sampled encodings, not full conversions; as run, both stored lossless frames (87 and 68 MB per minute).

D.12. Length of Stay

A stay is a visit: the time the camera spends in a room before it leaves. Each place keeps at most 128 frames, chosen evenly over all of its frames, so every stay keeps about the same share of its frames. Table S19 gives what the memory kept from each stay, from the logged reads (post hoc); Fig. 3b plots the test stays. Every stay keeps at least 10 frames, and the shortest tenth of stays are kept whole 92 to 96% of the time. No stay in our streams is shorter than 12 s; shorter stays are untested.

On the test split we grouped the questions by the shortest stay that shows an asked object, at its quartiles (42, 55 and 67 s). From the shortest group to the longest, Plac-

	Questions (streams)	Memory MB/min	Input k tokens	Reader GPU-s
Test streams, 11 minutes on average				
PlaceTally, 128 frames per place	460 (103)	2.32	105.8	–
PlaceTally, 64 frames per place	460 (103)	1.53	68.3	–
Whole-video reading	460 (103)	2.90	69.1	–
30-minute streams				
PlaceTally, 128 frames per place	202 (38)	2.34	283.1	4.8
PlaceTally, 64 frames per place	202 (38)	1.52	179.1	2.9
Whole-video reading	202 (38)	2.87	180.6	7.4
60-minute streams				
PlaceTally, 128 frames per place	68 (18)	2.26	548.8	12.2
PlaceTally, 64 frames per place	68 (18)	1.46	347.5	7.5
Whole-video reading	68 (18)	1.67	211.7	13.0

Table S18. Cost by stream length, ARKit test rooms. Memory: bytes kept per minute of video after construction (global features, the selected JPEG frames and the reader's index; no model weights, buffers or file-system overhead). Input and reader GPU time per question; batched GPU time is shared equally among the requests of a batch, so these times are lower than, and not comparable with, the unbatched query times of Table 5. –: not stored.

	Stays	Shortest / median (s)	Kept whole	Fewest frames kept	Shortest tenth kept whole
Test streams, 11 minutes	823	12 / 67	55.2	12	95
30-minute streams	817	12 / 68	53.7	10	96
60-minute streams	765	12 / 69	53.1	12	92

Table S19. What PlaceTally's memory keeps from each stay (visit) at 128 frames per place, from the logged reads (post hoc). Kept whole: share of stays (%) with every 1-fps frame kept. Shortest tenth: the same share among the shortest tenth of stays, up to 37, 38 and 37 s.

eTally scores 67.8% [59.6%, 75.8%], 54.3% [45.7%, 62.7%], 61.8% [55.5%, 67.7%] and 66.7% [57.7%, 75.2%], and whole-video reading 48.3% [39.9%, 57.8%], 50.9% [42.2%, 59.1%], 46.0% [35.5%, 56.5%] and 59.6% [51.2%, 68.1%]. The groups differ in true count (mean 4.3, 3.8, 3.8 and 2.6), as questions about more objects more often include a short stay. At equal true count, the shortest group minus the longest is +5.3% [−7.0%, 18.1%] for PlaceTally and −0.4% [−10.4%, 10.2%] for whole-video reading: short stays lower neither score, and whole-video reading's lower score on short stays comes from their larger counts.

E. Examples and Failure Cases

Figure S1 shows the two ways the place finder fails, on real streams. The last part of this section looks inside correct places, where the remaining errors come from the reader.

Look-alike rooms. On MultiScan-Revisit, place F1 is 97.8% on the 17 ordinary streams and 84.4% on the 16 streams with a look-alike room (Sec. 5.6), where 10 of PlaceTally's places mix two rooms. Look-alike pairs there score close to the return threshold: their median similarity is 0.860, against 0.831 to 0.848 on the ARKit splits.

We looked at three frames of each scene in the nine pairs of rooms that PlaceTally merged:

- One pair (scenes 48 and 71) is the same room, filed under two scene IDs, so that merge is correct and the truth counts it as two rooms.
- Two groups (47 and 59; 49, 57 and 70) look like parts of one open-plan floor.
- Five pairs look like different rooms in the same building. The example in Fig. S1a is one of them (scenes 107 and 111).

Part of the measured loss is therefore in the truth. Correcting it could only raise PlaceTally's place F1.

A return not recognised. 3RScan rescans were taken months apart, after furniture had moved and objects had changed, so one room can look different on its return. A return that is not recognised opens a new place, and its objects are counted again. Sec. 5.6 gives the rates on 3RScan-Revisit test.

A success and a miscount. Of the 460 ARKit test questions, 350 come from streams whose places match the true rooms exactly. On these, PlaceTally is exact on 183 questions, too high on 55 and too low on 112. In 163 of the 183 exact answers every place is counted right; in the other 20, errors cancel. Of the 167 wrong answers, 95 include a gate error, where a place without the object is counted or a place with it is counted as zero, and 72 come from counts alone.

Table S20 shows one question of each kind, chosen by fixed rules, leaving out the stream of Fig. 1. Example (a) has the largest true count among questions where every place is right and whole-video reading is wrong. Two of its rooms are seen twice, so adding up the visits gives 8, while PlaceTally reads each room once and answers 5. Example (b) has the largest true count among questions with exactly one wrong place: every returning room is read once and correctly, and one room seen once is read one short.

F. Pre-Registration Record

Table S21 lists every test that was registered in writing before its outcome was computed and that bears on a result in the paper or this supplement. Failed tests are included. Each registration is a dated file; we give the first eight characters of its SHA-256 hash, and the files can be provided on request. An amendment never replaced its registration: it sits beside it, with its own date and hash.

Several post hoc analyses were also specified in writing before their outcomes were computed: the return-rate study, the fixed-correction controls and the comparison of return evidence (Secs. D.2, D.3 and D.7). They set no bar and chose nothing, so they are descriptive and are not listed.

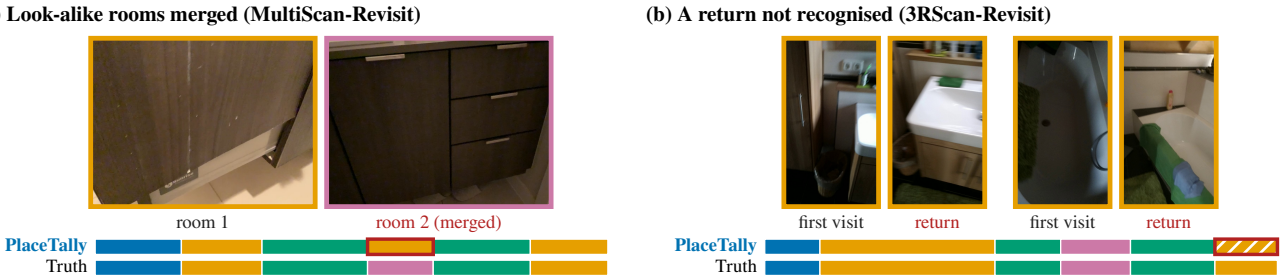


Figure S1. The two failure types of the place finder. Each bar runs along the stream, one colour per room; the upper bar shows PlaceTally's places and the lower bar the true rooms. Red boxes mark the wrong visits. (a) A MultiScan-Revisit stream with a look-alike room: the visit to the second room joins the first room's place, so their objects would be counted once. (b) A 3RScan-Revisit test stream: a bathroom rescanned months later is not recognised on its return and opens a new place (hatched), so its objects would be counted twice.

Place	Visits	True	Reader	P(yes)	Counted
(a) arkit-test-0035, TV or computer monitor					
Truth 5; PlaceTally 5; whole-video reading 3; sum over visits 8					
1	1, 5	3	3	0.96	3
2	2	0	0	0.06	0
3	3	1	1	0.88	1
4	4, 6	1	1	0.83	1
5	7	0	0	0.38	0
(b) arkit-test-0038, chair					
Truth 13; PlaceTally 12; whole-video reading 10; sum over visits 20					
1	1, 7	3	3	0.98	3
2	2, 8	4	4	0.98	4
3	3	0	0	0.17	0
4	4, 9	3	3	0.95	3
5	5	3	2	0.87	$2^{\dagger}$
6	6	0	0	0.10	0

Table S20. Two ARKit test questions, place by place, from the stored reads. Visits are numbered in stream order; a place with two visits is a room seen again. Counted is the reader's count if P(yes) $\geq \tau_g = 0.649$, and 0 otherwise. †Wrong place count. (a) Every place is right. (b) One place is one short.

G. Compute and Reproducibility

Hardware and software. GPU work ran on single NVIDIA A100 GPUs with 40 GB. Whole-video reading, whose context holds 262,144 tokens, twice failed to start on them for lack of memory, so for the held-out and 3RScan splits it ran on 80 GB A100 GPUs; the Gemma 4 runs also used 80 GB GPUs. The Qwen3.5 readers were served with vLLM 0.19.1rc1 (a development build, as the run logs record), in an environment that now has PyTorch 2.11.0 and Transformers 4.57.6, with greedy decoding, seed 0 and at most eight new tokens; Sec. B gives the checkpoints and their revisions. The detector, the continuation model and the place graph run on CPU from cached features (Sec. D.7). Per-stage costs are in Table 5 and Table S18.

The held-out run. The reads of the single held-out evaluation used 23,323 GPU-seconds (6.5 GPU-hours) in all, failed and timed-out jobs included; feature extraction is not counted. PlaceTally's reads took 3,946 GPU-seconds for 403 questions. Whole-video reading took 3,396, plus 740 for a job that failed at engine start (the hardware amendment in Table S21). Two checks that the reader gives the same answers in separate jobs took the remaining 15,241, most of it one check job that reached its time limit.

Metric and intervals. For a question with true count g, the mean relative accuracy [76] of an answer p is

$$\mathrm{MRA}(p, g) = \frac{1}{10} \sum_{\theta \in \{0.50, 0.55, \ldots, 0.95\}} \mathbb{1}\left[\frac{|p - g|}{g} < 1 - \theta\right],$$

with a strict inequality, so a count of 1 or 2 must be exact. A method's score is the mean over questions, in per cent; a difference of two scores is absolute. An answer that cannot be parsed counts as 0, which scores zero. A 95% interval comes from 5,000 bootstrap draws (seed 0) that resample whole streams with replacement; each draw takes the summed scores over the summed questions of its streams, and the interval is the 2.5th to 97.5th percentile of the draws. A paired difference uses the same draws for both methods. VSI-SUPER-Count's own metric uses $\leq$; Sec. D.8 reports strict MRA, and the benchmark's metric for PlaceTally as well.

What the code reproduces. We checked the released code on CPU. From the stored answers it recomputes the ARKit numbers of Table 3, except DirectMe's, and those of Fig. 3a; from stored frame-pair counts, those of Table 4; and from the stored cost logs, every build, memory, input and query entry of Table 5, the memory and answer times in Fig. 3c and d, all of Table S18, and the input and reader times in Table S10. The answers behind DirectMe's score, and behind the scores with 16, 32 and 64 frames per place on test and long streams, are not in the release. From the stored cut statistics and return similarities it reruns the continuation model and the best-link

Test	Registered	Rule fixed in advance	Outcome
Claims of the paper			
Held-out counting test	29 Oct d7c5b2bb	(1) PlaceTally minus whole-video reading $\geq$ +5 MRA, paired lower bound $>$ 0. (2) PlaceTally minus the blind baseline fixed in advance, lower bound $>$ 0. (3) Report whether PlaceTally reaches 54.5, its test score minus 8; no decision.	(1) Met: +16.9 [11.5, 22.2]. (2) Met: +22.0 [17.1, 27.0]. (3) 61.2.
Amendment	30 Oct 8ef8e517	Hardware only: whole-video reading moves to one 80 GB GPU after an out-of-memory failure at engine start, before any request; no scientific setting changes.	Made before any whole-video answer existed.
MultiScan joining test	29 Oct dbae258d	Run a new joining method only if PlaceTally's place F1 on MultiScan is below 90, so that MultiScan shows the failure the method targets.	F1 90.2: not met, stopped before any reads. The F1 is reported in Sec. 5.6.
Cross-camera continuation	3 Nov 4013e352	A learned continuation model is eligible only if its place F1 on training streams is at least PlaceTally's. Then: 3RScan MRA $\geq$ +5, ARKit MRA ≥ -1.5, MultiScan F1 $\geq$ 88.2.	No candidate was eligible, so PlaceTally's continuation model was kept; the 3RScan bar is not met (0.0).
Second reader family	1 Nov 3796c4e7	Primary contrast: PlaceTally minus whole-video reading, both with Gemma 4 12B, on all 460 test questions.	+23.7 [18.0, 29.3].
Long streams	2 Nov a0903021	Descriptive. Expected: whole-video reading falls faster with length than PlaceTally.	As expected (Table S8).
Cost and input budgets	1 Nov 5b0450af	Descriptive accounting of cost and frame budgets. Amendments: an allocation repair (efb690d7) and finishing 188 missing whole-video requests (d3f93c52).	Table 5, Tables S18 and S10.
Routes for counting inside a place (Sec. D.9)			
Hidden-state probe	29 Oct 31b3bd24	$\geq$ +3 MRA over the reader and over a control probe, both lower bounds $>$ 0, on the second training set.	Not met: -2.1 [-5.2, 0.8].
Learned box judge	29 Oct 29282729	On training rooms kept out of its fitting, keep $\geq$ 90% of real boxes with $\leq$ 2% of wrong ones kept.	Not met: 32%.
Molmo2-8B in video mode	29 Oct c684d694	Run specification; no margin.	-12.3 [-18.4, -6.7].
SAM 3 localisation maps	1 Nov 3b967255	Maps must find $\geq$ 80% of targets; then $\geq$ +3 MRA over the reader.	Maps found 474 of 506; MRA not met: -1.8, -2.1, -1.6.
Audit of wrong boxes	29 Oct 4dada52c	A sample of wrong boxes, labelled by hand into error types.	Registered, not run: the labels were never made.
Other video models (Sec. C); written before these runs, after the test split had been used			
CountVid	29 Oct 643530e3	40 test questions, stratified by category with seed 0, native 3 fps; descriptive.	0.0 (Table 3).
Cambrian-S	29 Oct d7a429ae	A second run on 150 test questions; descriptive.	31.7 and 1.7 (Sec. C).
VideoAgent	2 Nov c788fa49	Run protocol, fixed before the first reads: official code with the substitutions of Sec. C; first integer of the answer.	4.9 (Table 3).
Amendment	3 Nov d50f061e	After the first test reads: accept number words and retry on parser errors.	4.9 with either parser on the re-read answers.
DirectMe	15 Nov 2f6d9040	The 40 CountVid questions; released code on its README route; its own counting rule; native 3 fps; descriptive.	0.0 (Table 3).
Amendment	16 Nov fd751ebb	Before any answer was scored: its documented 1 fps, after 11 of the first 14 streams reached the two-hour build limit.	0.0 at 1 fps.

Table S21. Registered tests. Dates are in 2026; hashes are the first eight characters of the registration file's SHA-256. MRA and F1 are in per cent; MRA differences are paired, with 95% intervals over streams.

rule, and rebuilds the place partitions of the ARKit test, held-out and MultiScan streams exactly; the same check passes on 3RScan, whose statistics are not released. Extracting features and reading with the reader need a GPU. On two training streams, extracting features from recordings rebuilt by the benchmark scripts and reading them gave the stored answers to all 9 questions. The 3RScan numbers need a licensed copy of 3RScan, from which the build scripts derive its answers (Sec. H).

H. Data, Licences and Privacy

Sources and licences. RevisitCount is built from three public datasets, each under its own terms:

- ARKitScenes [8]: Apple's licence, which allows non-commercial use and redistribution under its conditions;
- 3RScan [71]: terms of use for non-commercial research and education, with no clause for public redistribution;
- MultiScan [49]: CC BY-NC 4.0.

Our annotations are released under CC BY-NC 4.0, which fits these terms, and our code under the MIT licence. Our licence does not replace the terms of the source data.

What is released. The release holds no images, video, depth or meshes from the source datasets. For ARKit-Revisit, every split including the held-out one, it gives the stream definitions (recording IDs, frame ranges, order, direction and orientation), the questions, the answers and the per-visit counts. For MultiScan-Revisit it gives the stream definitions and room identities, and no questions. For 3RScan-Revisit it gives only recording IDs, frame indices and their order:

because 3RScan's terms do not allow redistribution, its answers are derived by the build scripts from the user's own licensed copy. The scripts extract each stream's frames at the paper's rates and orientation and apply the visibility rule of Sec. A.

Privacy. RevisitCount adds no new recordings. Its files hold recording IDs, frame indices, room identities and object counts, and no personal information. The scenes are interiors of real homes and other spaces from the source datasets, which stay under their own terms.

Datasheet.

- *Purpose:* to measure counting of distinct objects when the camera returns to rooms it has seen, with blind baselines on every split.
- *Composition:* 672 streams in seven files: 467 from ARKitScenes (training, test, held-out and long), 172 from 3RScan and 33 from MultiScan (Table S1); 14 object categories (Table S2).
- *Collection:* no new data; streams join existing recordings by the procedure of Sec. A, with a fixed seed per split.
- *Labels:* derived automatically from the source annotations by a fixed visibility rule; no manual labelling.
- *Intended use:* research on counting and memory in video. Not for commercial use, and not for identifying people or places.
- *Known limits:* each stream joins separate recordings, so it is not one continuous video; same-day returns on ARKit; one orientation label per recording (Sec. A).
- *Distribution and upkeep:* the project's GitHub repository, `https://github.com/AlibayOsmanli/placetally`, where corrections will be posted.