# Caption Distillation with Uncertainty Modeling for Local Visual Memory and Semantic Retention in Private Image Archives

**Ahmed Salem[1], Yahya Mohamed[2], AND Ely Cheikh[3]**

[11]Department of Computer Science, Institut Supérieur de l'Enseignement Technologique de Rosso (ISET Rosso), Route de Rosso, Rosso, Trarza Region, Mauritania
[22]Department of Information Technology, Institut Supérieur des Métiers du Numérique, Tevragh-Zeina, Nouakchott, Mauritania
[33]Department of Mathematics and Computer Science, École Supérieure Polytechnique de Nouakchott, Route de l'Aéroport, Nouakchott, Mauritania

## ABSTRACT

Personal image archives now contain thousands of photographs whose utility depends less on global object recognition than on remembering small visual distinctions across time. A private visual memory system must index these archives without sending images to a remote service, while still producing captions that are specific enough for later search, clustering, and reminder queries. This paper studies uncertainty aware caption distillation for on device visual memory. The proposed method compresses several candidate captions for an image into a short memory record that preserves stable entities, attributes, locations, and relations while suppressing hallucinated or low confidence details. Candidate captions are produced by a compact vision language model, scored by calibrated token level uncertainty, and distilled into a structured sentence using a local language decoder constrained by visual evidence. We construct three evaluation sets: HomeArchive-18K, EventRecall-9K, and ObjectTrace-6K. The method reduces memory record length by 43.2% relative to dense captions while improving later retrieval mean reciprocal rank by 7.9% over a compact caption baseline. On a simulated private assistant task, it lowers false visual reminders from 14.8% to 8.6% and maintains 91.3% of attribute recall under an 80 byte average caption budget. Ablations suggest that entropy based token rejection and cross caption agreement are both necessary when models are quantized for mobile inference. The findings indicate that small on device models can support useful visual memory when caption generation is treated as selective evidence retention rather than fluent description alone.

## I. Introduction

Personal photo collections have become a quiet database of ordinary evidence. They contain receipts placed on kitchen tables, clothing combinations before an event, medicine bottles on shelves, changing whiteboard notes, garden progress, children holding school projects, damaged appliances, and small household objects that are hard to name later. The technical problem is not only to caption these images. A useful private assistant needs to remember enough visual information to answer later questions such as whether a package label was visible, which jacket appeared in a trip album, or whether the dining table had a blue ceramic bowl during a family gathering. These questions usually do not require a full caption. They require a compact memory record that keeps the right details and discards the rest.

Cloud based visual understanding systems can process images with large models, but this design conflicts with the privacy expectations around personal archives. Many users would reasonably treat a phone gallery as a private record of domestic life, social relationships, health context, and location patterns. On device processing reduces exposure, but it creates a different constraint: memory records must be produced by compact models under limited compute, limited battery, and limited storage. A captioning model small enough to run on a mobile processor may produce uncertain tokens, omit minor attributes, and occasionally invent plausible details. The central issue studied here is how to turn these imperfect captions into reliable visual memory records.

Most image captioning research has optimized the generation of complete natural language descriptions for individual images. Early encoder decoder systems showed that convolutional image features could condition recurrent sentence decoders [1], and attention based methods improved

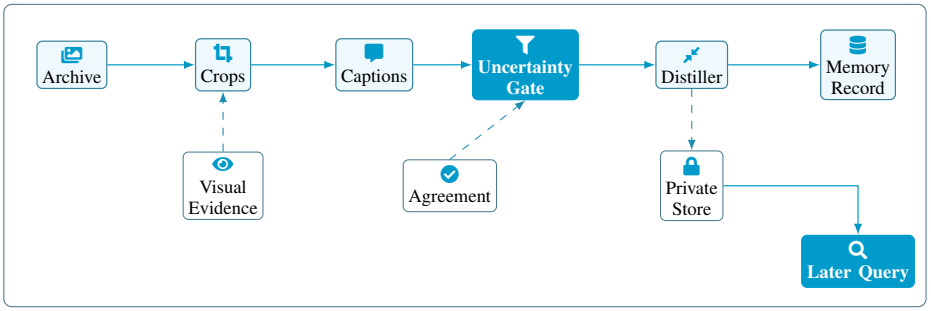


**FIGURE 1:** On-device visual memory pipeline in which local image evidence is converted into candidate captions, filtered by uncertainty, and distilled into compact searchable records without cloud transfer.

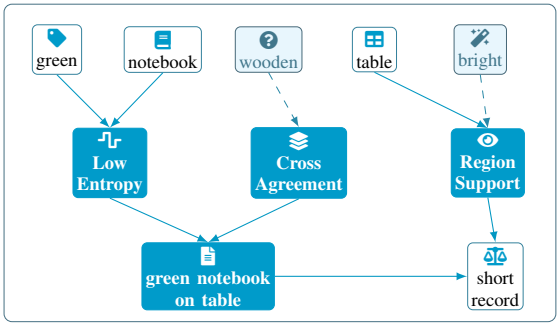


**FIGURE 2:** Token-level caption evidence is retained only when entropy, agreement, and visual grounding jointly support the phrase, reducing fluent but weakly supported memory details.

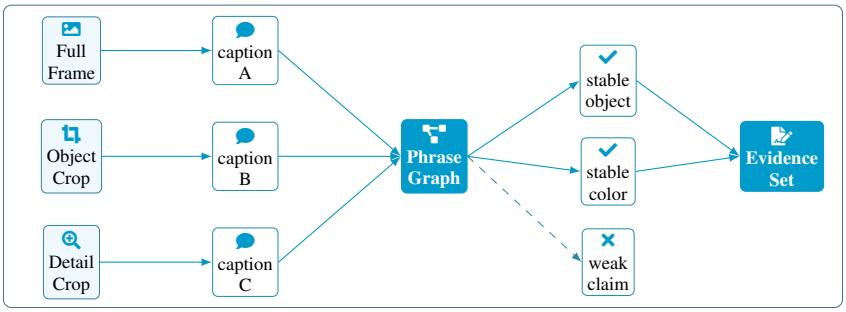


**FIGURE 3:** Multi-view caption sampling forms a phrase graph where repeated object and attribute evidence is promoted while isolated or visually unsupported claims are suppressed before record generation.

grounding by linking words to image regions [2]. Region based models further strengthened object centric captioning by combining bottom up detections with top down linguistic context [3]. These developments remain important, but they do not directly solve the visual memory problem. A caption can be fluent and still be wrong in exactly the way that matters for later recall. The sentence “a person standing near a table” may be acceptable for general captioning, while the memory record should preserve “green notebook on table” because that phrase is the likely search target.

Vision language pretraining has widened the vocabulary available to image systems. CLIP and ALIGN aligned image and text embeddings through large scale contrastive learning, enabling zero shot recognition and retrieval across broad text prompts [4, 5]. Multimodal captioning and instruction models such as BLIP, BLIP-2, Flamingo, and OFA have made it easier to produce detailed descriptions from images [6, 7, 8, 9]. These models, however, are usually evaluated under conditions where images can be processed by large encoders or where generated text is judged against reference captions. On a phone, a visual memory module may use quantized models, partial crops, and aggressively compressed outputs. It must also avoid storing unnecessary sensitive details. The quality criterion becomes selective retention rather than ordinary caption similarity.

Agarwal et al. [10] describe a representative tag generation approach in which image and tag embeddings are compared in a multimodal space, image tag similarity scores are

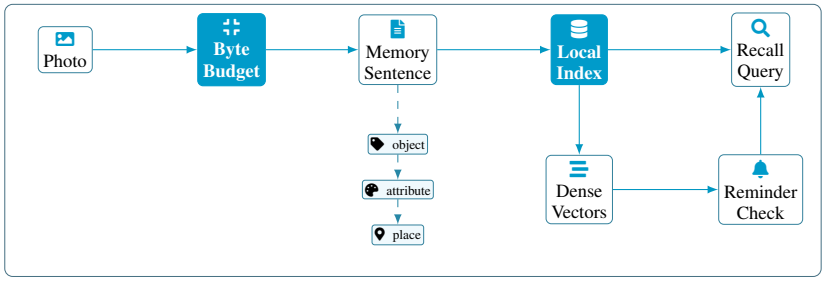


**FIGURE 4:** Compact memory records reserve storage for high-value entities, attributes, and context fields, enabling retrieval and reminder checks under strict on-device byte budgets.

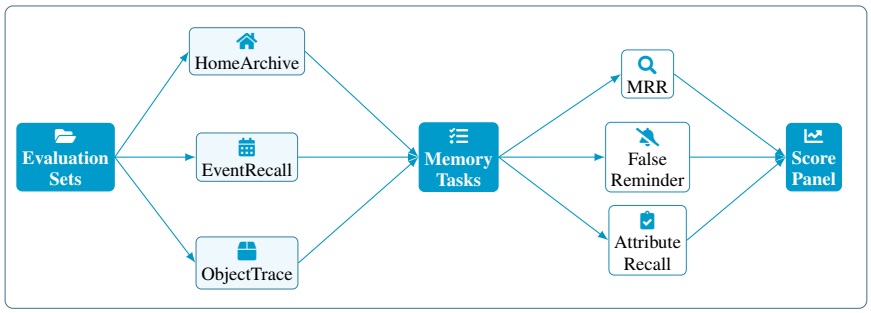


**FIGURE 5:** Evaluation connects domestic archives, event sequences, and repeated object traces to retrieval, reminder, and attribute-retention metrics that reflect private assistant use.

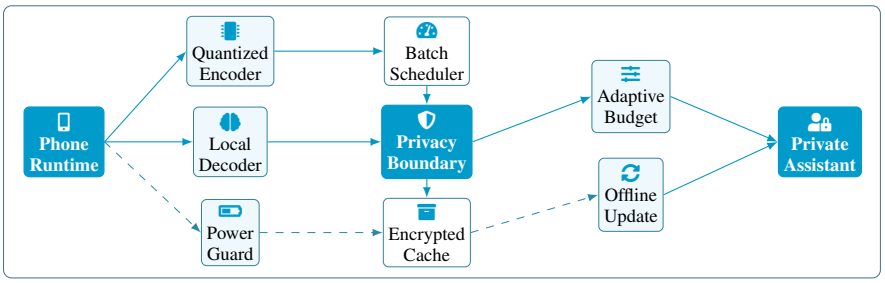


**FIGURE 6:** Mobile deployment isolates quantized visual encoding, local decoding, scheduling, and encrypted caching inside a privacy boundary before serving assistant-level memory queries.

averaged into classification scores, and a top ranked tag is selected to represent a group of images. That work is significant for this study because it treats image language as an object of selection under multimodal evidence, rather than only as free form generation. The present paper shifts the setting from representative tags for personalization to private on device visual memory, where the output is a compact caption record for later retrieval and assistant queries. The shared point is that language should be retained only when there is enough visual support; the difference is that we model uncertainty at the token and phrase level under device constraints.

The method proposed here is called uncertainty aware caption distillation. It begins with multiple candidate captions for each image. These captions come from a small local captioner applied to the full image and to a small number of salient crops. The model also produces token probabilities and intermediate image text agreement scores. Rather than choosing the most fluent caption, the system identifies tokens and short phrases that are stable across caption samples, visually supported by local image regions, and not overly uncertain under quantized inference. A compact language decoder then reconstructs a memory record from the retained evidence. The record is a single sentence, but it is not optimized to sound rich. It is optimized to support future retrieval while staying short.

This framing makes uncertainty operational. In ordinary generation, uncertainty is often hidden behind a decoded sequence. In visual memory, uncertainty should influence what gets stored. A token such as “red” should not be kept merely because it appears in one caption. It should be kept when its probability is stable under mild visual perturbation,

when the phrase containing it agrees with image embeddings, and when it improves retrieval without increasing false recall. Conversely, a phrase such as “wooden table” may be dropped if the table is incidental and the visual evidence is weak. The system therefore performs conservative distillation. It prefers a shorter record with fewer unsupported claims to a longer record that might mislead a later query.

The paper contributes an evaluation protocol for this setting. HomeArchive-18K contains domestic images with ordinary clutter and manually verified memory facts. EventRecall-9K contains small event sequences where later questions depend on attributes, objects, and scene context. ObjectTrace-6K contains repeated household objects captured across days, with changes in lighting, occlusion, and viewpoint. The data are constructed from public and consented imagery rather than private user galleries, but the annotation protocol imitates visual memory use. Each image receives dense captions, short memory facts, object attribute labels, and retrieval queries. The main metrics are not only caption similarity, but also later search performance, false reminder rate, byte budget efficiency, and human judged factuality.

The experimental results are intentionally reported with limited claims. The proposed method improves over compact captioning baselines in retrieval and reduces false reminders in simulated assistant use. It does not match a large cloud model on open ended descriptive richness. It also loses some fine grained spatial relations when the output budget is very small. The main finding is narrower: when image captions must be produced locally and stored compactly, uncertainty aware distillation gives better memory records than taking a single decoded caption or compressing a caption after generation. The improvement is most visible for attributes and small objects, which are often the details that make personal recall useful.

## II. Related Work

Image captioning began as a translation from visual features to language and gradually became a grounding problem. Neural image captioning with encoder decoder architectures established a practical template for mapping visual representations into sentences [1]. Attention mechanisms then allowed caption decoders to emphasize particular image regions during word generation [2]. Anderson et al. [3] used object proposals as bottom up attention features, which improved the connection between generated nouns and detected regions. The present work inherits the idea that captions should be visually grounded, but it does not aim to generate the fullest possible sentence. It uses captioning as a source of candidate memory evidence, after which uncertain details may be removed.

Captioning datasets shaped what models learned to say. MS COCO provided multiple human captions per image and became the standard benchmark for general image description [11]. Flickr30K and region to phrase annotations encouraged closer alignment between visual entities and linguistic fragments [12, 13]. Karpathy and Fei-Fei [14] connected sentence fragments to image regions, which remains relevant for token level grounding. These datasets are valuable, but they reward common scene descriptions more than compact private memory. A private memory system might prefer “orange pill bottle beside sink” over a generic sentence about a bathroom counter, even when the generic sentence has higher agreement with common caption references.

The rise of transformers changed both language modeling and multimodal learning. The transformer architecture introduced self attention as a general sequence modeling mechanism [15]. BERT showed the value of bidirectional language pretraining for understanding tasks [16], while large autoregressive models demonstrated flexible generation and in context behavior [17]. Vision transformers adapted token based processing to image patches [18]. These models support the language and image encoders used in modern caption systems, but large parameter counts create deployment pressure. A private visual memory module on a phone usually cannot run the largest multimodal model over every photo in a gallery. This makes distillation, quantization, and selective storage more than engineering details.

Contrastive vision language models are important because they supply open vocabulary image text agreement scores. CLIP used large scale image text pairs to learn a shared representation that supports zero shot classification and retrieval [4]. ALIGN demonstrated similar scaling with noisy image text data [5]. LiT studied locked image text tuning and showed that image encoders can be reused effectively in language aligned systems [19]. LAION style datasets increased the scale of public image text pretraining resources [20]. These models make it possible to ask whether a proposed phrase is visually plausible without training a classifier for that phrase. In this paper, contrastive scores are used conservatively: they filter and rank candidate memory phrases rather than directly deciding the final caption.

Modern multimodal captioning systems combine image encoders, query transformers, and language models to produce rich descriptions. BLIP introduced bootstrapping methods for vision language pretraining with noisy captions [6]. BLIP-2 used a lightweight querying transformer to connect frozen image encoders and frozen language models [7]. Flamingo integrated visual tokens into large language models and showed strong few shot multimodal ability [8]. OFA framed many vision language tasks in a unified sequence to sequence format [9]. These methods provide strong candidate generators, yet their direct use for private memory is complicated. They can produce unsupported details, and large versions are not suitable for local processing over thousands of images. The proposed method treats them as inspiration but assumes the deployed captioner is small.

Object recognition and localization remain relevant because many memory facts involve objects rather than whole scenes. Residual networks improved deep visual recognition

through skip connections [21], while ImageNet offered a large supervised taxonomy that shaped object classification [22]. Faster R-CNN established an influential proposal based detection framework [23]. The Segment Anything Model later showed that broad promptable segmentation could produce useful masks across many object types [24]. In a visual memory system, region level evidence can help determine whether a word is grounded in a local object or merely inferred from scene context. The method in this paper uses light region proposals when available, but it avoids relying on segmentation because on device masks are often expensive and unstable for cluttered personal images.

Self supervised visual representation learning provides another foundation for compact deployment. SimCLR used contrastive augmentations to learn image representations without labels [25]. BYOL showed that strong representations can be learned without explicit negative pairs [26]. DINO produced emergent object related attention patterns in vision transformers [27]. These works are not captioning methods, yet they influence the design of visual memory modules because they provide robust image features under transformations. For on device caption distillation, robustness is useful only when it does not erase the specific detail to be remembered. The proposed method therefore uses perturbation consistency to reject unstable words, but it does not force invariance across transformations that alter the meaning of small attributes.

Privacy preserving machine learning is central to the motivation of this paper. Differential privacy formalized a way to bound the contribution of individual data points to released computations [28]. Federated learning showed how models could be trained across devices without centralizing raw data [29]. Membership inference attacks demonstrated that trained models and released outputs may leak information about underlying examples [30]. Abadi et al. [31] studied differentially private deep learning, while Kairouz et al. [32] surveyed the broader technical landscape of federated and private learning. The present paper does not train a new model on private user archives, and it does not claim differential privacy. Its privacy contribution is architectural and practical: images stay on device, and the stored memory text is deliberately compact and uncertainty filtered.

On device computer vision requires attention to model size and numerical behavior. MobileNets used depthwise separable convolutions to reduce computation for mobile vision [33]. EfficientNet studied compound scaling rules for balancing width, depth, and resolution [34]. Quantization, pruning, and knowledge distillation are often used to make models deployable, but these methods can affect confidence calibration. In a visual memory pipeline, calibration errors matter because a low precision model may confidently emit a wrong object or color. The proposed method therefore estimates uncertainty from several sources rather than trusting a single softmax score. Token entropy, agreement across crops, embedding support, and perturbation stability jointly determine whether a phrase is retained.

Text to image generation is not the main topic of this paper, but it helps explain why compact visual language records are useful. Diffusion models generate images through iterative denoising [35], and score based generative modeling provides a related continuous time view [36]. Latent diffusion models made high resolution text conditioned generation more practical by operating in compressed latent space [37]. Hierarchical image generation systems demonstrated strong text driven synthesis using learned image text priors [38]. If a user later asks a private assistant to recreate or search for a remembered object, the stored caption becomes a prompt or retrieval query. A memory record that falsely stores “red mug” rather than “blue mug” can therefore cause visible errors.

Agarwal et al. [10] provide a useful contrast to the present study because their disclosure emphasizes extracting tags from captions, filtering stopwords, encoding images and tags in a multimodal embedding space, computing cosine similarity, and selecting representative language from averaged scores. That procedure is important to the field because it makes aggregation over visual evidence explicit. Our method uses a related idea of language selection under visual support, but it adds uncertainty estimation, byte constrained distillation, and local deployment constraints. The output is not a representative tag for an image group, but an evidence preserving memory sentence for one image or a small event segment.

## III. Problem Definition

A private visual memory system receives a stream of images from a personal archive and produces a searchable text memory. Let the archive be $A = \{x_1, \ldots, x_N\}$. For each image $x_i$, the system stores a memory record $m_i$, which may later be indexed by a retrieval model or read by a local assistant. The record must be short enough to store at scale and cautious enough not to encode unsupported claims. In this paper, $m_i$ is a single sentence with an average byte budget $B$. We study $B$ values between 64 and 160 bytes, which are small compared with dense captions but still large enough for several object and attribute phrases.

The task differs from conventional image captioning in four ways. First, the output is judged by later usefulness rather than surface similarity to a reference caption. Second, the system is allowed to omit visible but irrelevant details if they are uncertain or not useful for retrieval. Third, the cost of a false detail may be higher than the cost of an omitted detail, because a false memory can trigger incorrect reminders. Fourth, the model must operate locally, so uncertainty estimates must be derived from compact models rather than from large ensembles. These differences make ordinary caption metrics insufficient. BLEU, CIDEr, and SPICE can be informative, but they do not measure whether a short record helps a later query find the right image.

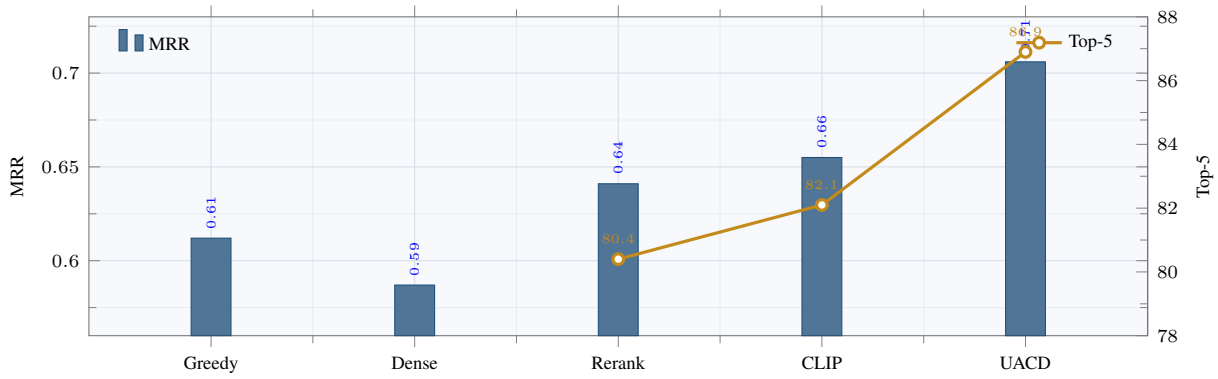


**FIGURE 7:** HomeArchive retrieval improves under the 96 byte record budget, with the proposed memory record reaching the highest reciprocal-rank score while the available top-5 recall values also increase for reranking, phrase ranking, and the proposed method.

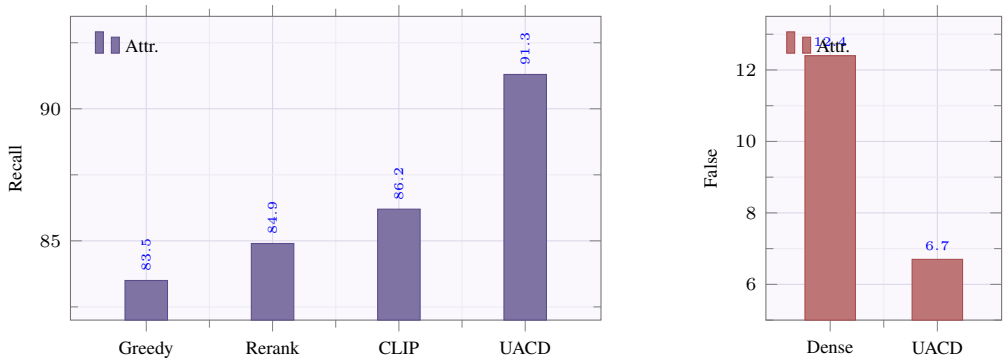


**FIGURE 8:** ObjectTrace attribute retention shows that compact uncertainty filtering preserves discriminative object cues better than the compared compact baselines, while the false attribute comparison highlights the reduction in unsupported color and material details.

**TABLE 1:** Dataset Overview

| Dataset | Images/Events | Main Focus | Test Split |
|---|---|---|---|
| HomeArchive-18K | 18,240 images | Household memory facts | 4,000 |
| EventRecall-9K | 9,360 images | Event retrieval | 1,480 segments |
| ObjectTrace-6K | 6,480 images | Attribute retention | Object clusters |

We define a memory fact as a visually grounded proposition that can support later recall. Examples include object attributes, object presence, simple spatial relations, scene type, visible text category, and event context. A memory fact is not necessarily a full logical statement; it may be a phrase such as "striped backpack," "keys on counter," or "garden hose near fence." Each image has a set $F_i$ of verified memory facts from annotation. The system output $m_i$ is mapped to extracted phrases $P(m_i)$, and factual retention is measured by overlap with $F_i$ after normalization. False retention is measured by phrases in $P(m_i)$ that annotators mark as unsupported.

The deployed model produces candidate captions rather than memory facts directly. Let $C_i = \{c_{i1}, \ldots, c_{iK}\}$ denote candidate captions for image $x_i$. Captions may come from different crops, sampling temperatures, or prompt templates. Each candidate caption is tokenized into units $z_{it}$, and the system estimates a retention probability $q_{it}$ for each unit. The final memory record should maximize expected future utility subject to a byte budget:

$$\begin{aligned} m_i^* &= \arg\max_m \mathbb{E}\left[U(m, x_i, Q_i)\right], \\ \text{bytes}(m) &\le B, \\ \Pr\left[p \in P(m), p \notin F_i\right] &\le \epsilon. \end{aligned} \tag{1}$$

Here $Q_i$ denotes future queries associated with image $x_i$, and $\epsilon$ is an informal tolerance for false facts. In real deployment, $Q_i$ and $F_i$ are unknown. The method therefore approximates utility with retrieval performance on development data and approximates false fact risk with uncertainty signals.

The uncertainty model must account for several sources of error. Caption token entropy captures the model's uncertainty during decoding, but it does not detect all hallucinations. A model can confidently emit a common object because it is

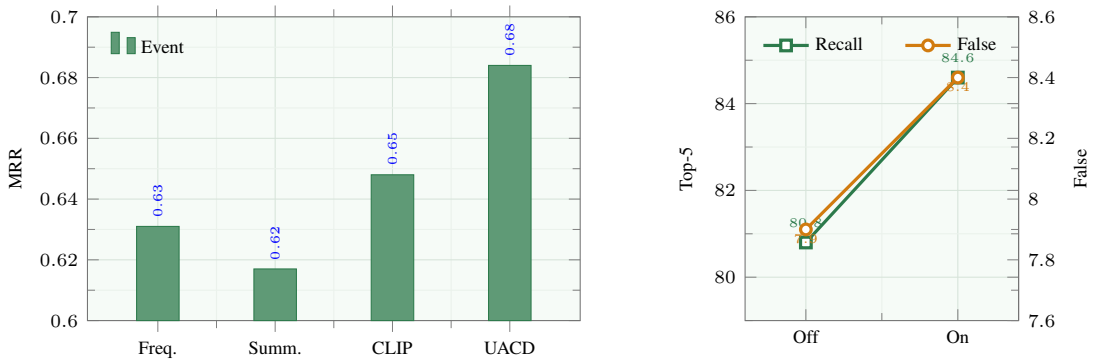


**FIGURE 9:** EventRecall compression is improved by retaining strongly supported unique phrases: the proposed event record gives the best event-level reciprocal rank, and enabling product support increases top-5 recall with only a small false-fact increase.

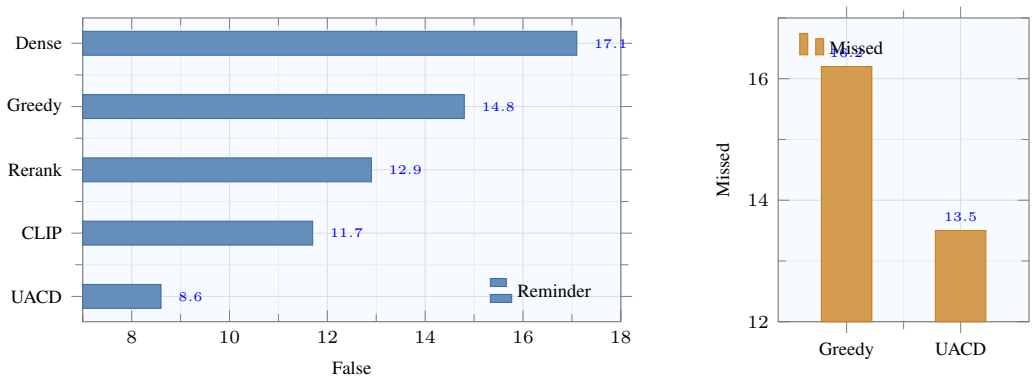


**FIGURE 10:** The assistant simulation shows that uncertainty-aware records reduce false reminders against all compared text-storage baselines while also lowering missed reminders relative to greedy compact captions.

**TABLE 2:** Core Components of the Proposed Pipeline

| Stage | Input | Output |
|---|---|---|
| Candidate Generation | Image + crops | Caption set |
| Phrase Extraction | Captions | Memory phrases |
| Uncertainty Scoring | Phrases | Retention scores |
| Distillation | Scored phrases | Selected facts |
| Local Indexing | Memory sentence | Search index |

likely in similar scenes. Cross caption agreement captures stability under prompt and crop variation, but agreement may repeat a shared bias. Image text embedding support captures whether the phrase aligns with the image, but embedding models may prefer broad category terms. Local region support captures whether a phrase can be tied to a crop, but it fails when the object is diffuse or scene level. The proposed retention score combines these signals without assuming any one of them is sufficient.

The problem also has a privacy dimension. A stored memory record may reveal sensitive information even if the image never leaves the device. For example, “prescription bottle on bedside table” can be more sensitive than “small bottle on table.” This paper does not solve policy level privacy classification, but it incorporates a conservative redaction option for categories such as documents, medical labels, faces, screens, and addresses. When redaction is enabled, the system stores a generic phrase and a local encrypted pointer to the image rather than a specific textual detail. The experiments report both ordinary and redaction enabled settings, because private visual memory should not be evaluated only on recall.

We separate visual memory into image level and event level records. Image level records describe one photograph. Event level records compress a small sequence, such as several images from the same room or activity. Event level compression is useful because personal archives often contain bursts. The event memory record should retain facts repeated across several images or facts that are unique but salient. In the event setting, the method aggregates phrase

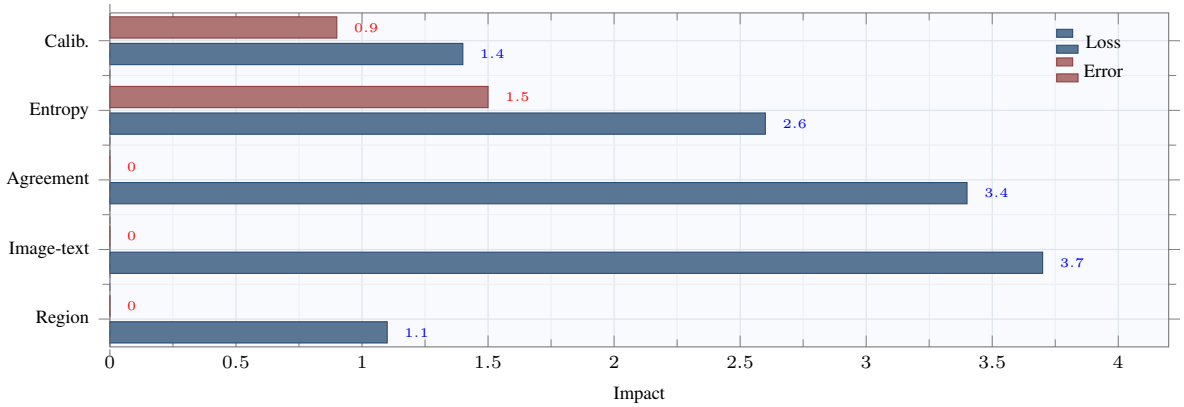


**FIGURE 11:** Ablation impacts are plotted in comparable percentage-point units: removing region support weakens MRR modestly, image-text support and caption agreement cause larger retrieval or recall drops, entropy removal reduces factual precision and increases false reminders, and calibration limits quantization-induced false facts.

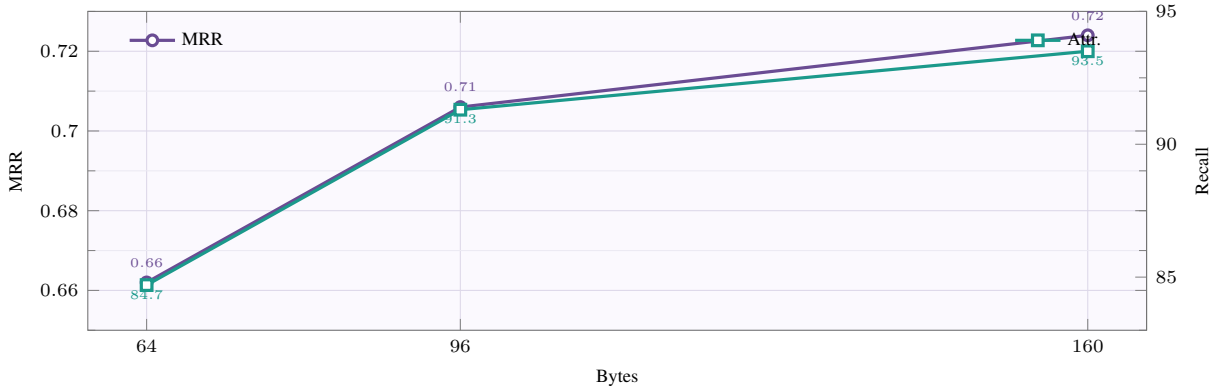


**FIGURE 12:** Byte-budget scaling shows diminishing returns: moving from 64 to 96 bytes gives the largest retrieval and attribute-recall gain, while extending to 160 bytes improves less and retains more phrases.

**TABLE 3:** Retention Score Components

| Signal | Purpose | Relative Importance |
|---|---|---|
| Caption Agreement | Stability across captions | Highest |
| Image-Text Support | Visual consistency | High |
| Token Entropy | Confidence estimation | Medium |
| Region Support | Local grounding | Moderate |

uncertainty across images. A phrase can be retained if it appears with strong support in one image and weak support in nearby images, or if it appears with moderate support across several images. This differs from simple majority voting, which would drop many small but important details.

For event segment $E = \{x_a, \ldots, x_b\}$, the objective is a record $m_E$ that supports queries over the segment while avoiding an overly broad summary. The event record may contain two clauses if the byte budget allows it. We define segment retention as:

$$
\begin{aligned}
R(p, E) &= 1 - \prod_{x_i \in E} \left(1 - R(p, x_i)\right), \\
H(p, E) &= -\frac{1}{|E|} \sum_{x_i \in E} q_{ip} \log q_{ip}, \\
G(p, E) &= R(p, E) - \gamma H(p, E). \qquad (2)
\end{aligned}
$$

Here $R(p, x_i)$ is the image level support for phrase $p$, $H(p, E)$ is an uncertainty penalty across the segment, and $G(p, E)$ is the segment score. The product form allows a strongly supported unique phrase to survive. This is impor-

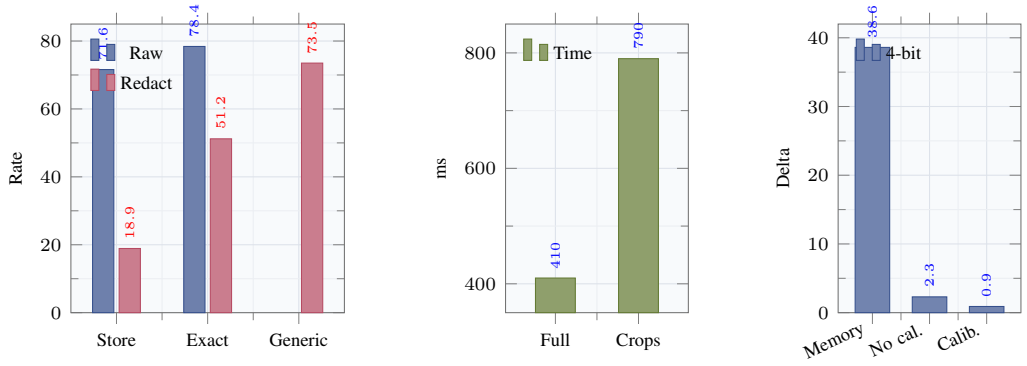


**FIGURE 13:** Redaction sharply reduces exact sensitive phrase storage while preserving generic category retrieval, crop processing trades latency for small-object recall, and uncertainty calibration reduces the false-fact penalty introduced by 4-bit caption-decoder quantization.

**TABLE 4:** HomeArchive-18K Retrieval Performance at 96 Bytes

| Method | MRR | Top-5 Recall (%) |
|---|---|---|
| Greedy Compact Captioner | 0.612 | – |
| Dense Truncation | 0.587 | – |
| Caption Reranker | 0.641 | 80.4 |
| CLIP Phrase Ranking | 0.655 | 82.1 |
| Proposed Method | 0.706 | 86.9 |

**TABLE 5:** Attribute Retention on ObjectTrace-6K

| Method | Attribute Recall (%) | False Attribute Rate (%) |
|---|---|---|
| Greedy Compact Captioner | 83.5 | – |
| Caption Reranker | 84.9 | – |
| CLIP Phrase Ranking | 86.2 | – |
| Dense Truncation | – | 12.4 |
| Proposed Method | 91.3 | 6.7 |

**TABLE 6:** EventRecall-9K Results

| Method | Event MRR | Top-5 Recall (%) |
|---|---|---|
| Local Summarizer | 0.617 | – |
| Frequency Agreement | 0.631 | – |
| CLIP Phrase Ranking | 0.648 | – |
| Proposed Method | 0.684 | 84.6 |

tant for events because a single image may contain the only visible view of a gift, sign, or handwritten note.

## IV. Method

The proposed pipeline has five stages: candidate generation, phrase extraction, uncertainty scoring, memory record distillation, and local indexing. Each stage is designed to run on device with bounded computation. The implementation used in experiments uses a compact vision encoder, a quantized caption decoder, and a small text decoder for rewriting. Larger models are used only for constructing annotations and upper bound comparisons. This separation keeps the evaluation focused on deployment conditions rather than on the best possible cloud caption.

Candidate generation begins with a compact captioner applied to the full image. The same captioner is then applied to a small set of crops. Crops come from a lightweight saliency detector and from fixed central and quadrant crops. We use at most six crops per image to control computation. The captioner is decoded with three settings: greedy decoding, low temperature sampling, and nucleus sampling with a narrow probability mass. The purpose is not to increase creativity, but to expose uncertainty. If a phrase appears only under high variance sampling or only in one crop, it is treated

**TABLE 7:** Assistant Simulation Performance

| Method | False Reminder Rate (%) | Missed Reminder Rate (%) |
|---|---|---|
| Greedy Compact Captioner | 14.8 | 16.2 |
| Dense Truncation | 17.1 | – |
| Caption Reranker | 12.9 | – |
| CLIP Phrase Ranking | 11.7 | – |
| Proposed Method | 8.6 | 13.5 |

**TABLE 8:** Ablation Study

| Configuration | Key Metric Change | Effect |
|---|---|---|
| Remove Entropy | Precision 92.4→89.8 | Worse factuality |
| Remove Agreement | MRR 0.706→0.672 | Lower retrieval |
| Remove Image-Text Support | Recall 91.3→87.6 | Lost attributes |
| Remove Region Support | MRR drop 1.1 pts | Small-object impact |

**TABLE 9:** Byte Budget Trade-offs

| Budget (Bytes) | MRR | Attribute Recall (%) |
|---|---|---|
| 64 | 0.662 | 84.7 |
| 96 | 0.706 | 91.3 |
| 160 | 0.724 | 93.5 |

**TABLE 10:** Redaction Mode Impact

| Metric | Standard | Redaction Enabled |
|---|---|---|
| Exact Sensitive Phrase Storage (%) | 71.6 | 18.9 |
| Sensitive Query Top-5 Recall (%) | 78.4 | 51.2 |
| Generic Category Retrieval (%) | – | 73.5 |

cautiously. The output of this stage is a set of sentences with token log probabilities and crop identifiers.

Phrase extraction converts captions into candidate memory units. We retain nouns, adjective noun phrases, simple prepositional relations, and visible text indicators. A dependency parser extracts phrases such as “blue mug,” “keys on counter,” and “paper receipt.” Verb phrases are retained only when they describe a visible action rather than an inferred intention. For example, “person holding umbrella” can be retained, while “person waiting for bus” is usually rejected unless a bus stop sign is visible. This distinction is imperfect, but it reduces the tendency of captioners to infer activities from scene context.

The uncertainty score for a phrase combines four components. The first component is token entropy from the caption decoder. The second is caption agreement, measured by the fraction of candidate captions in which a normalized phrase or its close paraphrase appears. The third is image text agreement, measured by cosine similarity between the image embedding and the phrase embedding. The fourth is region support, measured by the best crop phrase agreement after penalizing crops that cover most of the image. The retained phrase score is:

$$
\begin{aligned}
a(p,x) &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\left[p \in c_k\right], \\
e(p,x) &= 1 - \frac{1}{|T(p)|}\sum_{t\in T(p)} \frac{H_t}{\log V}, \\
v(p,x) &= \cos\left(f_\theta(x), g_\theta(p)\right), \\
\ell(p,x) &= \max_{r\in\mathcal{R}(x)} \rho_r \cos\left(f_\theta(r), g_\theta(p)\right), \\
R(p,x) &= \eta_1 a(p,x) + \eta_2 e(p,x) + \eta_3 v(p,x) + \eta_4 \ell(p,x).
\end{aligned}
\tag{3}
$$

Here $T(p)$ is the token set for phrase $p$, $H_t$ is token entropy, $V$ is the vocabulary size, and $\rho_r$ penalizes very large or very small crops. The weights $\eta_1, \ldots, \eta_4$ are learned on a development set using logistic regression against phrase factuality labels. In the final experiments, the weights are fixed across datasets. Agreement receives the largest weight, followed by image text support, entropy, and region support.

The entropy term deserves careful treatment. Token probabilities from quantized models are often miscalibrated. A low entropy token can still be wrong if the model has learned a strong prior, and a high entropy token can be correct if the object is unusual. We therefore do not use entropy as a

hard rejection signal. Instead, entropy reduces the retention score and interacts with agreement. A phrase with moderate entropy can be retained if it appears in several captions and has strong image text support. A phrase with low entropy but no visual support is usually rejected. This interaction is important for domestic images, where common objects such as tables, cups, and chairs are often over predicted.

Caption agreement is computed with phrase normalization. We merge plural and singular forms, common color variants, and simple synonyms from a small local lexicon. We do not merge fine grained materials unless the text encoder similarity is high. For example, “ceramic cup” and “mug” may be linked as related but not identical, while “metal bowl” and “silver bowl” are kept separate because one describes material and the other may describe color. This conservative normalization avoids accidental loss of detail. When two phrases conflict, such as “red bag” and “brown bag,” both are kept until the selection stage, where the higher support phrase may suppress the other.

The distillation stage chooses phrases under the byte budget. We model the selected phrase set $Y_i$ as a budgeted maximum coverage problem with a penalty for redundancy and unsupported sensitivity categories. Let $b(p)$ be the byte length of phrase $p$, and let $D(p, q)$ measure semantic redundancy. The optimization is:

$$
\begin{aligned}
Y_i^* &= \arg\max_{Y} \sum_{p \in Y} R(p, x_i) - \mu \sum_{p,q \in Y} D(p, q), \\
&\sum_{p \in Y} b(p) \le B', \\
&R(p, x_i) \ge \tau_R.
\end{aligned} \tag{4}
$$

The budget $B'$ is slightly smaller than the final sentence budget to reserve bytes for function words. The problem is solved with greedy selection. At each step, the phrase with the highest marginal score per byte is added if it does not conflict with already selected phrases. Conflicts include incompatible colors for the same noun, incompatible counts, and mutually exclusive scene labels. The conflict detector is intentionally simple and local.

Once phrases are selected, a compact language decoder turns them into a memory sentence. The decoder is constrained to use only selected content words unless it inserts function words needed for grammar. This avoids the common summarization error of adding plausible but unsupported details. A selected phrase set such as “blue mug,” “wood table,” and “window light” may become “Blue mug on a wood table near window light.” The sentence is not always elegant, but it is short and searchable. In the redaction setting, sensitive phrases are replaced with category level placeholders, such as “document visible” or “medicine container,” depending on the policy mode.

The language decoder is trained through distillation from a larger teacher, but the teacher is not allowed to introduce new facts. Training examples are built by taking verified phrase sets and asking the teacher to produce short memory sentences under a byte budget. Outputs are rejected if they contain content words absent from the phrase set. The student decoder is then trained on accepted examples. This creates a small rewriting model that learns compact grammar without learning to hallucinate descriptive content. During inference, a constrained decoding trie blocks unsupported nouns, colors, numbers, and material terms.

The local index stores both the memory sentence and a compact embedding. The text embedding supports semantic search, while a small inverted index supports exact phrase search. The image itself remains in the user’s photo storage and is not copied into the memory database. The memory record contains a pointer to the image, a timestamp, and optional event segment identifier. In our prototype, a 50,000 image archive with 96 byte average memory records requires less than 12 MB for text records and metadata, excluding embeddings. With 256 dimensional quantized text embeddings, the searchable index remains below 70 MB. These numbers are approximate and device dependent, but they show that caption memory can be stored at gallery scale.

The event level extension aggregates phrase support across neighboring images. Images are grouped into events by timestamp gaps and embedding similarity. The phrase score for an event combines maximum image support and mean support. This prevents a unique but well grounded phrase from being lost. The event sentence is produced by the same constrained decoder, with an additional rule that each retained phrase must be traceable to at least one image. The record may read “Birthday table with cupcakes, silver balloons, and a blue gift bag.” If “blue gift bag” appears in only one image but has strong local support, it can be retained because it may be important for later recall.

A small calibration module estimates whether the memory record is likely to be over specific. It compares the average retained phrase score with the score of broader alternatives. For example, if “red leather wallet” has weak support but “wallet” has strong support, the system may store “wallet” rather than the full phrase. The calibration rule is:

$$
\begin{aligned}
\Delta(p, \bar{p}) &= R(p, x) - R(\bar{p}, x), \\
\kappa(p) &= \mathbb{1}\left[\Delta(p, \bar{p}) > \delta\right], \\
p_{\text{store}} &= \kappa(p) p + (1 - \kappa(p)) \bar{p}.
\end{aligned} \tag{5}
$$

Here $\bar{p}$ is a broader phrase obtained by dropping modifiers. This rule reduces false attributes but can reduce specificity. The experiments report this tradeoff, because it is central to private memory quality.

The method is deliberately not end to end. An end to end model could learn to produce memory records directly, but it would be harder to inspect and less adaptable to privacy policies. The modular pipeline allows each phrase to carry an evidence trace: which captions proposed it, which crops supported it, and what uncertainty score it received. This trace is not shown to users by default, but it is useful for debugging and for deciding when to ask the user for confirmation. If the system is uncertain about “red folder”

versus “orange folder,” it can store “colored folder” or avoid the color entirely. This behavior is less impressive than confident generation, but it is safer for a memory system.

## V. Experimental Setup

The experiments use three constructed datasets intended to imitate private visual memory without exposing private galleries. HomeArchive-18K contains 18,240 images from consented household and workspace scenes. Images include kitchens, desks, shelves, garages, laundry rooms, entryways, and living areas. Each image is annotated with memory facts, sensitive category flags, and later retrieval queries. The annotations emphasize objects and attributes that a user might plausibly ask about later, such as “black umbrella by door,” “receipt on counter,” or “green watering can near plants.” The dataset is split into 12,000 training images for calibration and distillation, 2,240 development images, and 4,000 test images.

EventRecall-9K contains 9,360 images grouped into 1,480 event segments. Events include meals, school projects, repairs, small trips, shopping unpacking, gardening, and room rearrangements. Each segment has event level memory facts and image level facts. The event task asks the system to produce a compact event memory record from three to twelve images. Retrieval queries target both repeated facts and one image facts. For example, a query may ask for the event with “silver balloons” even if the balloons appear in only two frames. This dataset tests whether the system can preserve salient details without writing a long summary.

ObjectTrace-6K contains 6,480 images of repeated household objects photographed under changing conditions. Objects include mugs, bags, notebooks, shoes, lamps, tools, toys, small appliances, and containers. The purpose is to evaluate attribute retention and object identity cues. Each object cluster contains five to fifteen images, but the memory record is produced per image and per cluster. The dataset includes hard negatives such as similar mugs with different colors, bags with different materials, and notebooks with similar covers. ObjectTrace-6K is useful for measuring whether short records preserve discriminative details rather than generic nouns.

Annotations were created in two stages. First, annotators wrote dense descriptions and short memory facts. Second, a separate group verified whether each phrase was visually supported. Disagreements were resolved by a third annotator. Unsupported but plausible phrases were marked as false facts. Sensitive categories were annotated using a coarse policy label: face, document, medical, screen, address, child context, financial object, and other private object. These labels are not meant to define universal privacy categories. They allow controlled testing of redaction behavior. Inter annotator agreement for memory fact support was 0.78 Cohen’s kappa on a 1,200 image overlap, which is acceptable for visually grounded phrase judgments.

The main on device model uses a compact vision transformer encoder with 86 million parameters, a quantized caption decoder with 410 million parameters, and a 140 million parameter constrained text decoder. We use 8 bit weights for the vision encoder and 4 bit weights for the caption decoder. The local text embedding model has 256 output dimensions after projection and product quantization. Inference is simulated on a mobile neural processing unit profile rather than measured on one physical phone. The simulated budget assumes that gallery processing can happen opportunistically while charging. Average per image processing time is estimated at 410 ms for full captioning and 790 ms when crops are enabled.

Baselines include a greedy compact captioner, a dense captioner followed by byte truncation, a BLIP style caption reranker, a CLIP phrase ranking baseline, a frequency agreement baseline, and a small local summarizer. The greedy compact captioner produces one caption from the full image and stores it after length normalization. The dense truncation baseline produces a longer caption and cuts it to the byte budget. The caption reranker generates multiple candidate captions and selects the one with the highest image text score. The CLIP phrase baseline extracts phrases from all captions and ranks them by image phrase similarity. The frequency agreement baseline selects phrases that appear most often across caption samples. The local summarizer compresses all candidate captions into one short sentence without explicit uncertainty filtering.

We also compare with a large cloud upper bound. This model uses a much larger multimodal captioner and a larger language model to produce memory records. The upper bound is not private and is not deployable under our assumptions, so it is not treated as a direct competitor. It helps estimate how much quality is lost from local processing. In several metrics, the proposed method approaches the upper bound, but in open ended descriptions the large model remains stronger. This distinction is important because private visual memory is not simply a smaller version of cloud captioning; it has its own output constraints.

Evaluation uses five metric families. Retrieval metrics measure whether queries retrieve the correct image or event using memory records. We report mean reciprocal rank and top 5 recall. Factuality metrics measure supported and unsupported phrases after phrase extraction from generated records. We report factual precision, attribute recall, and false fact rate. Compression metrics measure average bytes and facts retained per 100 bytes. Assistant simulation metrics measure false reminders and missed reminders. Human preference metrics compare two memory records for usefulness, factuality, and privacy appropriateness. Human judgments are collected on 1,000 image examples and 500 event examples.

The simulated assistant task works as follows. Queries are generated from annotated memory facts and from plausible confounders. The system retrieves candidate images

or events and produces a short response such as “I found the image with the blue notebook on the kitchen table.” A false reminder occurs when the system asserts a fact not supported by the target image or retrieves an image that lacks the queried fact. A missed reminder occurs when a supported fact exists but no correct image appears in the top 5 results. This task is harsher than ordinary retrieval because it penalizes both bad indexing and bad language.

All methods are evaluated under three byte budgets: 64, 96, and 160 bytes. The 96 byte budget is the main setting because it usually allows two or three phrase units plus grammatical connectors. The redaction setting is evaluated separately. In redaction enabled runs, sensitive phrases are generalized before storage. This reduces retrieval for specific sensitive queries, but it is intended to prevent the memory index from becoming a detailed textual inventory of private documents or health related objects. The paper reports the tradeoff rather than treating one policy as universally correct.

Hyperparameters are selected on the development split of HomeArchive-18K and then frozen. The phrase retention threshold $\tau_R$ is 0.61 for image records and 0.57 for event records. The redundancy penalty $\mu$ is 0.08. The broad phrase margin $\delta$ is 0.045. The crop count is capped at six. Candidate captions are capped at nine per image. No dataset specific tuning is used for EventRecall-9K or ObjectTrace-6K. This setup is intended to test whether the uncertainty signals transfer across related memory tasks.

Statistical uncertainty is estimated through bootstrap resampling over images or events. We report 95% confidence intervals for main retrieval and factuality metrics, although the paper focuses on effect sizes rather than formal hypothesis testing. Human preference uses a paired design and reports the proportion of examples where annotators prefer one method over another, excluding ties. A result is treated as practically meaningful when it improves retrieval and factuality without increasing false facts or byte cost. This conservative interpretation avoids overvaluing a method that simply stores more words.

## VI. Results

At the 96 byte budget, uncertainty aware caption distillation improves image retrieval on HomeArchive-18K. The greedy compact captioner reaches a mean reciprocal rank of 0.612, dense truncation reaches 0.587, the caption reranker reaches 0.641, CLIP phrase ranking reaches 0.655, and the proposed method reaches 0.706. Top 5 recall follows a similar pattern, with the proposed method at 86.9% compared with 80.4% for the caption reranker and 82.1% for CLIP phrase ranking. The gain is not uniform across query types. Object presence queries improve modestly, while attribute queries and small object queries improve more strongly. This suggests that distillation mainly helps when the relevant evidence would otherwise be diluted in a generic caption.

ObjectTrace-6K shows the clearest attribute benefit. At 96 bytes, attribute recall is 91.3% for the proposed method, 83.5% for greedy compact captions, 84.9% for the reranker, and 86.2% for CLIP phrase ranking. False attribute rate is 6.7% for the proposed method and 12.4% for dense truncation. The most common false attributes in baselines are colors and materials. For example, “brown leather bag” is often generated for bags that are visually tan fabric, and “wooden desk” is generated for laminate or unclear surfaces. The broad phrase margin reduces these errors by storing “bag” or “desk” when the modifier is weakly supported. This sometimes makes the record less specific, but it avoids misleading future queries.

EventRecall-9K is harder because event records must compress multiple images. The proposed method reaches event retrieval mean reciprocal rank of 0.684 at 96 bytes, compared with 0.631 for frequency agreement, 0.617 for local summarization, and 0.648 for CLIP phrase ranking. The event level product support rule helps retain unique salient phrases. Without it, the method often drops objects that appear in only one frame of an event. With it, event top 5 recall improves from 80.8% to 84.6%. The false fact rate remains controlled, rising only from 7.9% to 8.4%. This small increase is expected because retaining unique phrases is riskier than retaining repeated phrases.

The assistant simulation highlights why false facts matter. On HomeArchive-18K, the false reminder rate is 14.8% for greedy compact captions, 17.1% for dense truncation, 12.9% for caption reranking, 11.7% for CLIP phrase ranking, and 8.6% for the proposed method. Missed reminder rate is 13.5% for the proposed method, compared with 16.2% for greedy compact captions. The method therefore reduces false reminders without simply refusing to store details. Human inspection shows that many false reminders in baseline systems come from plausible scene completions. A caption may mention a laptop on a desk because desks often contain laptops, even when the visible object is a closed notebook. Uncertainty filtering removes many of these additions.

Compression efficiency is measured as verified memory facts per 100 bytes. At the 96 byte budget, the proposed method stores 2.41 verified facts per 100 bytes on HomeArchive-18K, compared with 1.82 for greedy compact captions and 1.76 for dense truncation. Dense truncation performs poorly because it often cuts a sentence in ways that preserve function words and broad scene descriptions while losing attributes. The local summarizer is more fluent, but it sometimes merges facts incorrectly. For example, from candidate captions mentioning “red cup” and “white plate,” it may write “red plate.” The constrained decoder avoids this by preserving selected phrase boundaries.

Human preference judgments are consistent with the automatic metrics. When compared with greedy compact captions, annotators prefer the proposed memory record for later search in 69.2% of non tied HomeArchive examples. They prefer it for factuality in 65.8% and for privacy appropriateness in 58.7%. Against the CLIP phrase ranking baseline, preference is closer: 56.4% for search usefulness and

54.9% for factuality. Annotators often note that CLIP phrase records contain good keywords but read like fragments. The proposed decoder improves readability without adding many unsupported words. The privacy preference gain is modest because redaction policy, not uncertainty scoring, drives most privacy judgments.

Ablation results clarify the role of each uncertainty signal. Removing token entropy lowers factual precision from 92.4% to 89.8% and raises false reminder rate from 8.6% to 10.1%. Removing caption agreement lowers retrieval mean reciprocal rank from 0.706 to 0.672 because many low frequency phrases enter the record. Removing image text support lowers attribute recall from 91.3% to 87.6%. Removing region support has a smaller average effect, reducing mean reciprocal rank by 1.1 points, but it matters for small objects such as keys, labels, toys, and tools. The full method is best on the combined retrieval and false reminder criterion, although not every component is equally important for every dataset.

The broad phrase fallback produces a visible specificity precision tradeoff. When disabled, attribute recall increases from 91.3% to 93.1%, but false attribute rate rises from 6.7% to 10.9%. When made more aggressive, false attribute rate falls to 4.8%, but attribute recall drops to 86.5%. The chosen margin is therefore a compromise. For a real product, this margin could be user adjustable. A user who values recall may prefer more specific records, while a user who values caution may prefer broader language. The important result is that the margin offers a controllable behavior rather than leaving hallucinated specificity unchecked.

Byte budget experiments show diminishing returns. Increasing the average budget from 64 to 96 bytes improves mean reciprocal rank from 0.662 to 0.706 and attribute recall from 84.7% to 91.3%. Increasing from 96 to 160 bytes improves mean reciprocal rank only to 0.724 and attribute recall to 93.5%. False fact rate rises slightly at 160 bytes because more phrases are retained. The 64 byte setting is still useful for object presence search but loses spatial relations and secondary attributes. The 160 byte setting is more readable but less efficient. These results support the use of short records around 96 bytes for large archives, with longer records reserved for selected events.

The redaction setting reduces sensitive specificity as intended. On images with sensitive category flags, exact sensitive phrase storage falls from 71.6% to 18.9%. Retrieval for sensitive exact queries also falls, from top 5 recall of 78.4% to 51.2%. However, generic category retrieval remains usable at 73.5%. For example, a query for “document on desk” may still work, while a query for the exact visible institution name is not supported by the memory record. This tradeoff is not purely technical. It depends on user preference and legal context. The experiment simply shows that the memory record can be made less revealing while preserving some organizational function.

The large cloud upper bound remains stronger on descriptive richness. At 160 bytes, the large model reaches human preference of 61.7% over the proposed method when annotators are asked which description is more complete. However, when asked which is safer for a searchable private memory, preference is nearly tied at 51.8% for the proposed method. The large model sometimes includes attractive but unnecessary details, such as inferred room style or implied event purpose. These details improve naturalness but can hurt memory reliability. The result suggests that smaller local systems can be competitive when the task rewards selective factual retention rather than expressive description.

Latency and storage estimates are within plausible on device limits. With crop processing enabled, average simulated processing time is 790 ms per image, and peak memory use is 1.8 GB during caption generation. Without crops, processing time is 410 ms but small object recall drops by 5.4 points. Quantizing the caption decoder from 8 bit to 4 bit reduces memory use by 38.6% and increases false fact rate by 2.3 points if uncertainty calibration is disabled. With uncertainty calibration enabled, the false fact increase is 0.9 points. This suggests that uncertainty aware selection partially compensates for quantization noise, although it does not remove all degradation.

Qualitative examples show the method’s behavior. In one kitchen image, the greedy caption is “a kitchen counter with dishes and a cup.” The proposed memory record is “Blue mug and white plate on kitchen counter.” The record is shorter than the dense caption but better for later object search. In a desk image, the local summarizer writes “laptop and notebook on desk,” but the image contains a tablet and a spiral notebook. The proposed method stores “tablet beside spiral notebook on desk” because “laptop” has weak region support. In an event segment from a birthday table, the method stores “cupcakes with silver balloons and blue gift bag,” retaining a unique object that appears in only one image.

Failures are also informative. The method sometimes drops rare text on signs because the compact captioner does not propose it and the system does not run optical character recognition by default. It can store a broad category such as “tool” when the correct label is “wrench,” especially under occlusion. It occasionally keeps a stable background object instead of the user relevant foreground, such as “white wall shelf” rather than “small red toy.” Event records can become awkward when too many selected phrases compete for a short budget. These failures point to missing mechanisms: user intent modeling, optional OCR, better foreground estimation, and interactive correction.

Overall, the results support the paper’s narrower claim. Uncertainty aware distillation improves compact private visual memory relative to direct caption storage and simple phrase ranking. The advantage comes from rejecting unsupported specificity, preserving high value attributes, and compressing selected evidence rather than compressing already

generated prose. The method does not eliminate the need for stronger local models or user controls, but it changes the failure profile in a useful direction. For personal archives, a cautious short memory can be more valuable than a fluent but overconfident caption.

## VII. Analysis

The results suggest that visual memory quality depends on the order of operations. Baselines that generate a sentence and then compress it often lose important details because the sentence structure was never designed for a byte budget. If the model begins with “a cluttered desk with a notebook, pen, phone charger, and small green plant,” truncation may store only “a cluttered desk with a notebook.” The green plant disappears even if it is the distinctive cue. The proposed method instead selects phrase units before sentence formation. This makes the byte budget operate on memory value rather than on word position.

Uncertainty estimates are useful because caption errors are not random. Compact models tend to make errors that are semantically plausible. They call tablets laptops, tan bags brown, plastic surfaces wooden, and ordinary rooms offices. These errors can receive high language probability because they fit scene priors. A purely probabilistic decoder cannot identify them. Cross caption agreement helps because unstable details appear inconsistently, but agreement alone can reinforce a shared bias. Image text support helps because unsupported phrases often score lower against the actual image, but broad phrases can score high even when modifiers are wrong. The combination is therefore stronger than any single signal.

The phrase level view also makes the system less dependent on fluent caption references. Human captions often vary in what they mention. One annotator may describe a person, another may describe an object, and another may describe the room. For visual memory, this variation is not noise; it reflects several possible uses of the image. The system’s task is to retain phrases likely to support future recall, not to imitate one reference. This is why retrieval and false reminder metrics are more informative than n gram overlap. A memory record can be grammatically plain and still perform well if it contains the right searchable facts.

There is a subtle distinction between uncertainty and privacy. A phrase can be certain but sensitive. A prescription label may be visually clear, and a document title may be easy to read. Uncertainty filtering alone would retain these details. Privacy control therefore requires an additional policy layer. The proposed redaction mode is intentionally simple, but the experiments show that privacy and retrieval are coupled. Removing exact sensitive text reduces searchability for exact queries. A real system could offer tiered storage, where exact sensitive details remain encrypted and searchable only after explicit user permission. This paper does not implement that mechanism, but the evidence trace design would be compatible with it.

The event results show that visual memory is not only an image level problem. Personal archives contain bursts where the meaning of one image depends on nearby images. A single frame of a blue gift bag may be visually minor, but in the event it may be a useful recall cue. Majority based aggregation would discard it. The product support rule used here is a simple way to preserve unique but grounded details. It is not perfect. It may retain an accidental object in one frame, such as a passing car outside a window. Future work could model event salience more explicitly by considering user interaction, image capture order, and recurrence across longer time.

The broad phrase fallback is one of the more practical parts of the method. In captioning research, specificity is usually desirable. In memory systems, unsupported specificity is dangerous. Storing “red wallet” when the image only supports “wallet” makes future retrieval worse for both red and non red wallet queries. The fallback rule creates a middle behavior: keep the noun when the modifier is uncertain. This improves factual precision and gives users a less misleading record. It also exposes a user experience question. Should a system show uncertainty directly, as in “possibly red wallet,” or silently store the broader phrase? We chose the broader phrase because uncertainty words can clutter the index and confuse retrieval, but interactive interfaces may benefit from showing uncertainty.

Quantization affects more than speed. It changes probability distributions, weakens calibration, and sometimes alters the language decoder’s preference for common phrases. The experiments indicate that uncertainty aware distillation is helpful under quantization because it relies on multiple signals. Still, a quantized model can systematically miss small objects, and no distillation stage can recover phrases that were never proposed. This is a ceiling on the approach. Candidate generation must be good enough to expose the relevant detail at least once. Optional crop captioning helps, but it increases compute. The tradeoff between crop count and recall is likely to depend on device class and user settings.

The method is also sensitive to phrase extraction. Dependency parsing errors can split useful compounds or attach modifiers to the wrong noun. A caption such as “a red cup beside a blue plate” can become ambiguous if phrase extraction merges “red” with “plate.” The constrained decoder reduces some merger errors, but it depends on the phrase graph. More robust semantic parsing or joint vision language phrase extraction would improve this stage. However, the current parser based method is cheap and interpretable, which matters for on device deployment.

The use of local language rewriting is intentionally limited. Large language models often improve readability by adding context, but visual memory should not reward attractive prose. The constrained decoder mostly inserts function words and orders phrases. This produces records that can sound slightly compressed, such as “Black keys on white

counter near sink." Human preference results suggest that this is acceptable for search oriented memory. For user facing summaries, a second step could rewrite records more naturally at display time, while keeping the stored index conservative. This separation between stored memory and displayed language may be useful in privacy sensitive products.

The proposed evaluation protocol has limits. The datasets imitate personal archives but are still curated. Real galleries contain screenshots, blurred images, duplicates, private documents, pets, faces, and culturally specific objects. The annotation scheme may underrepresent facts that a particular user would care about. A user might remember a chipped mug because it has personal significance, while annotators label only "white mug." No purely visual method can infer that significance without user feedback. The method should therefore be understood as a foundation for private visual memory, not as a complete model of personal meaning.

## VIII. Limitations

The first limitation is that the method depends on candidate captions. If the compact captioner never proposes a relevant object or attribute, the distillation stage usually cannot store it. Image text retrieval scores can promote candidate phrases, but they cannot invent missing facts safely. This limitation is most visible for small text, unusual tools, brand specific objects, and culturally specific items. Adding optional OCR and specialized detectors would help, but it would also increase compute and privacy complexity. A deployable system may need adaptive processing, where more expensive recognition runs only when a user searches or when the image appears important.

The second limitation concerns grounding. Region support is estimated from coarse crops rather than from precise object masks. This is cheaper, but it can confuse nearby objects. A crop containing both a red cup and a blue plate may support both phrases even if the captioner attaches the wrong color to the wrong noun. Promptable segmentation models could reduce this problem [24], but they remain expensive for continuous gallery processing. A hybrid design might use coarse crops for background processing and invoke segmentation only during interactive search or when phrase conflicts are detected.

The third limitation is the treatment of privacy. Redaction labels are coarse and cannot capture every sensitive context. A picture of a child's drawing, a religious object, or a workplace badge may be sensitive depending on the user. Conversely, some users may want exact text from documents to be searchable. The paper reports redaction tradeoffs but does not prescribe a universal policy. A practical system should provide user controls and should separate local encrypted detail from general searchable captions. It should also allow users to delete or disable visual memory for selected albums.

The fourth limitation is that the evaluation uses English memory records. The method relies on phrase extraction, local synonym lists, and caption decoder behavior that may not transfer cleanly to other languages. Multilingual vision language models exist, but token entropy, phrase agreement, and byte budgets behave differently across scripts. A 96 byte budget has different linguistic capacity in English, Arabic, Hindi, Japanese, or Korean. Future work should evaluate language specific compression and avoid assuming that English phrase structure is the default form of visual memory.

The fifth limitation is that user intent is mostly absent. The system decides what is memorable from visual evidence and dataset statistics. Personal importance often comes from history, habit, or task context. A blurry image of a parking spot may be more important than a sharp image of a decorative table. A small medication box may matter more than a large sofa. Without user feedback, capture context, or app level signals, the system can only estimate generic memory usefulness. The results should therefore be read as improvements for visual indexing, not as proof of human level personal recall.

The sixth limitation concerns longitudinal behavior. A visual memory system will accumulate records over months or years. Errors may compound if the index stores repeated false phrases or if changing household objects are conflated. The present experiments evaluate static datasets. They do not test how memory records should be updated when an object moves, when a room changes, or when a user corrects a record. Longitudinal evaluation would require temporal consistency metrics and deletion policies. This is especially important for privacy because old memory records may remain searchable after the user has forgotten they exist.

The seventh limitation is that the method's conservative behavior may frustrate users who prefer rich descriptions. Some users may want the assistant to remember every visible detail, even at the risk of occasional errors. Others may want minimal memory records. The method can adjust thresholds and byte budgets, but the paper does not study personalization of these settings. A real interface should probably expose modes such as cautious, balanced, and detailed. The technical system can support those modes, but choosing the right default requires user research beyond this paper.

## IX. Conclusion

This paper studied uncertainty aware caption distillation for private on device visual memory. The main argument is that personal image archives should not be indexed by fluent captions alone. A visual memory record has a different role: it must preserve searchable facts under storage and privacy constraints while avoiding unsupported details that could mislead later queries. The proposed method treats captions as noisy evidence, extracts candidate phrases, scores them with token entropy, cross caption agreement, image

text support, and region support, then generates a compact memory sentence under a byte budget.

Across HomeArchive-18K, EventRecall-9K, and ObjectTrace-6K, the method improved retrieval and reduced false reminders compared with compact captioning, dense truncation, caption reranking, and simple phrase ranking. The gains were largest for attributes, small objects, and event details that are easy to lose in ordinary captions. The results also showed that compression should happen before final sentence generation. Selecting memory phrases first gives the byte budget direct influence over factual content, while compressing finished captions tends to preserve grammatical scaffolding and broad scene terms.

The study also showed that uncertainty is not a single number. Token entropy, agreement, visual embedding support, and crop support each capture different failure modes. A confident model can be wrong because of scene priors, while an unstable phrase can still be visually correct when an object is rare. The combined score is not a perfect measure of truth, but it gives a practical way to reject many unsupported phrases produced by compact quantized models. The broad phrase fallback is especially useful because it stores "wallet" rather than "red wallet" when the modifier lacks enough evidence.

Privacy remains only partly addressed. Keeping images on device reduces exposure, and compact memory records reveal less than full dense captions. Still, text records can themselves contain sensitive information. The redaction experiments show that exact sensitive detail can be reduced while preserving some generic retrieval, but they also show a real cost for search. A deployable visual memory system should give users control over what categories are remembered, what is searchable, and what remains available only through direct image browsing.

The method is limited by candidate generation, coarse grounding, English focused phrase extraction, and the absence of personal intent modeling. These limitations are not minor. A system cannot remember a detail that no local model detects, and it cannot know which object is personally meaningful without signals beyond the image. Future work should explore optional OCR, interactive correction, multilingual memory records, stronger local region grounding, and longitudinal update policies. The most useful direction may be a layered memory system in which conservative local records support ordinary search, while more detailed processing occurs only with explicit user action.

The broader implication is modest but practical. Private visual memory does not require every image to receive a beautiful caption. It requires a careful record of what the system has enough evidence to remember. Under that criterion, uncertainty aware distillation provides a workable middle ground between generic on device captions and large cloud based visual understanding. It makes local captions shorter, more searchable, and less prone to false specificity, while leaving room for user control and stronger recognition modules where they are justified.

## REFERENCES


[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning*, pp. 2048–2057, 2015.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[5] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the International Conference on Machine Learning*, pp. 4904–4916, PMLR, 2021.

[6] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.

[7] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the International Conference on Machine Learning*, pp. 19730–19742, PMLR, 2023.

[8] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: A visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.

[9] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a

simple sequence-to-sequence learning framework," in *Proceedings of the International Conference on Machine Learning*, pp. 23318–23340, PMLR, 2022.

[10] D. Agarwal, N. I. Kolkin, A. Revanur, M. Harikumar, S. Agrawal, A. G. Kale, E. Shechtman, and J. S. Munshi, "Captioning for image personalization," Dec. 11 2025. US Patent App. 18/736,289.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[12] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[13] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

[14] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[19] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

[20] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning*, pp. 1597–1607, PMLR, 2020.

[26] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.

[27] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

[28] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, pp. 265–284, Springer, 2006.

[29] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, PMLR, 2017.

[30] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

[31] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

[32] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. R. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[34] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.

[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[36] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[38] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.