*Eclipta Journals*

RESEARCH

# Variance Calibrated Concept Aggregation for Personalized Image Captioning Across Heterogeneous Small Visual Collections With Sparse Tags

Tuan Bui,[1] Khanh Vo,[2] and Viet Do[3]

[1]Department of Information Technology, Quang Binh University, 312 Ly Thuong Kiet Street, Dong Hoi City, Quang Binh Province 510000, Vietnam
[2]Department of Computer Science, Ha Tinh University, 447 26 Thang 3 Street, Dai Nai Ward, Ha Tinh City, Ha Tinh Province 450000, Vietnam
[3]Department of Information Systems, Hong Duc University, 565 Quang Trung Boulevard, Dong Ve Ward, Thanh Hoa City, Thanh Hoa Province 440000, Vietnam

**Abstract**

Personalized image captioning increasingly has to work with small groups of related images rather than with isolated photographs. A user may provide several images of one object, place, activity, or visual style, while the system has to infer a compact description that is stable across the set and still specific enough to guide retrieval or generation. This paper studies a variance calibrated concept aggregation method for set-conditioned captioning under sparse and noisy tags. The method combines multimodal image–text similarity, collection-level dispersion estimates, tag competition, and a controlled language generation stage that converts ranked concepts into captions. We evaluate the approach on a constructed benchmark of 18,240 small image collections derived from COCO, Flickr30k, and Visual Genome, with additional synthetic perturbations that simulate inconsistent user tags, repeated near-duplicates, and weak personalization signals. Across five collection sizes, the proposed method improves representative tag F1 from 0.612 to 0.684 over a cosine-mean baseline and improves caption specificity from 0.438 to 0.493 under human annotation. It also reduces hallucinated attributes by 14.7% relative to a direct multimodal captioning baseline. Ablations indicate that variance calibration is most useful when collections contain visually coherent objects with changing backgrounds, while a lexical diversity penalty matters more for mixed scene collections. The results suggest that set-level concept aggregation is a practical intermediate layer between image tagging and free-form caption generation, especially when personalization depends on a few repeated visual cues rather than a large user profile.

## 1. Introduction

Personalized image captioning is often framed as a direct mapping from one image to one sentence, but a large fraction of everyday personalization problems have a different shape. A user collects several images of the same bicycle, kitchen corner, clothing item, pet toy, school project, bedroom layout, or product prototype, and the desired description is not simply the best caption for any individual frame. The desired description has to preserve what repeats across the collection, ignore incidental changes in angle and lighting, and avoid averaging the images into a vague label such as object, room, or vehicle. This makes the task closer to set summarization than ordinary image captioning. The input is a small visual collection with partial redundancy, and the output is a tag or caption that should be representative without becoming generic. The difficulty is not that modern vision-language models lack vocabulary. The difficulty is that a shared embedding space can assign high scores to many plausible labels, and the labels that look locally plausible for one image may not be globally useful for the collection.

The present paper studies this problem through the lens of calibrated concept aggregation. The method takes a set of images and a candidate set of textual concepts, computes image–tag compatibility in a multimodal embedding space, estimates how stable each embedding dimension is across the collection, and then generates a caption from the best-scoring concepts. The technical aim is narrow. We do not try to build a full personal memory model, infer long-term preferences, or reconstruct the user's intent from conversation history. Instead, we ask whether a set-level scoring layer between visual recognition and language generation can improve caption specificity and reduce unstable attributes when only a small collection is available. The question is practical because many deployed editing, search, and generation workflows receive a few user-supplied images and a few weak tags, not a carefully annotated dataset.

The study is connected to several lines of work in multimodal representation learning. Contrastive image–text models made it possible to compare images and natural language labels in a shared space, with CLIP providing a widely used reference for zero-shot classification and retrieval [1].

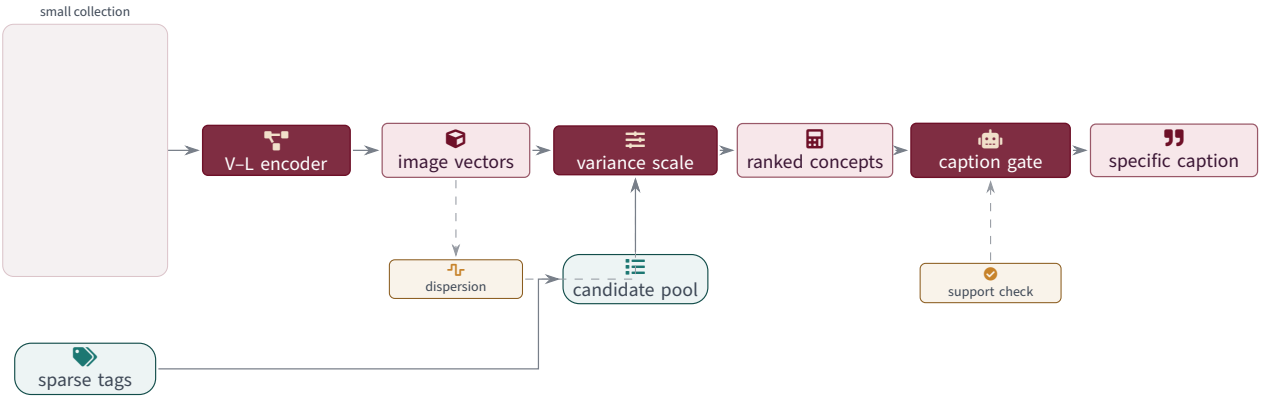



**Figure 1.** Variance calibrated concept aggregation converts heterogeneous image collections and sparse tags into stable concepts before constrained caption generation.

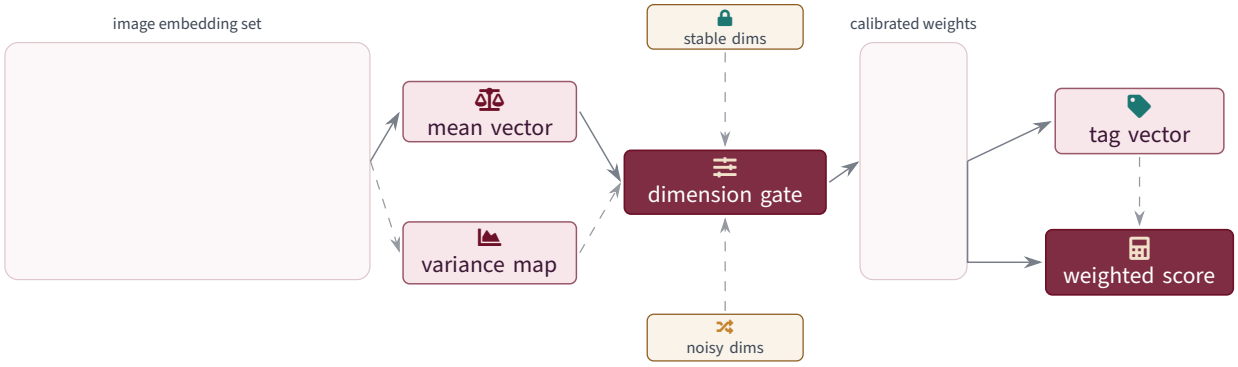



**Figure 2.** Collection dispersion rescales embedding dimensions so stable visual evidence contributes more than fluctuating background or viewpoint cues.

ALIGN showed that noisy image–text pairs at scale can still produce useful cross-modal alignment [2], while later systems such as CoCa and SigLIP refined the loss design and pretraining recipe for stronger matching and generation-adjacent performance [3; 4]. These models are attractive for set-level captioning because they allow tags, short phrases, and images to be compared without training a classifier for each possible concept. Yet this flexibility also creates a calibration problem. A tag can be semantically close to one image for reasons that do not hold across the collection, and a common tag can dominate simply because it appears in many captions, not because it captures the user's object of interest.

Agarwal et al. [5] provide an applied formulation that is significant for this study because it treats representative tagging as a collection-level decision rather than as a set of independent image labels. Their disclosed procedure obtains a plurality of images and tags, computes image–tag similarity in a multimodal embedding space, averages the relevant similarity scores, and selects the tag with the highest classification score. This is an important field-level reference point for our work because it makes explicit the operational link between caption-derived tags, cosine similarity, score aggregation, and representative tag selection. Our study uses that link as a starting point, but it focuses on the statistical behavior of aggregation when small collections contain repeated foreground objects, changing backgrounds, and sparse candidate labels. The emphasis is therefore not on claiming a new end-to-end captioner, but on testing a calibrated scoring stage that can be inserted before language generation.

The central concern is that a mean similarity score can be both useful and brittle. It is useful because it dampens single-image noise and rewards concepts that recur across images. It is brittle because it treats all embedding dimensions and all images as equally informative. If the foreground object is stable but the background shifts, dimensions associated with background context can pull the score toward broad scene labels. If two visually related concepts are close in the embedding space, a high mean score may not distinguish the user-specific phrase from a generic parent concept. If captions generated from individual images contain frequent but uninformative terms, lexical repetition can upweight common words that do not carry personalization value. These issues are small in large supervised settings where many examples can correct one another, but they are visible in small personal collections where five or eight images may be the entire evidence base.

The proposed method, variance calibrated concept aggregation, addresses this brittleness by combining three

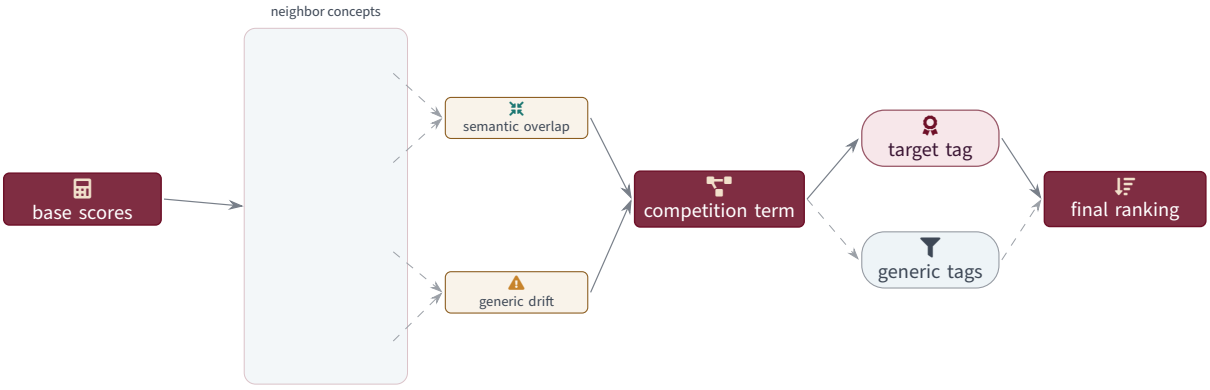



**Figure 3.** Candidate concepts compete locally, promoting specific repeated tags while suppressing visually plausible but weakly personalized generic labels.

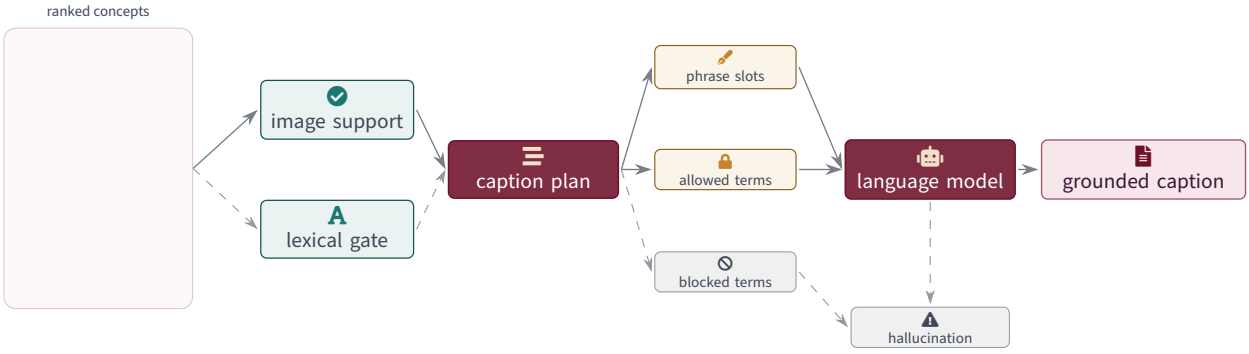



**Figure 4.** Constrained generation treats the language model as a fluency layer, allowing supported concepts while blocking weak attributes.

operations. First, it estimates the dispersion of image embeddings inside the collection and uses this dispersion to rescale the contribution of each dimension before computing image–tag similarity. Dimensions that fluctuate across the set are not discarded, but their contribution is reduced when the target is a collection-level concept. Second, it introduces a competition term among candidate tags so that a concept receives a stronger score when it is not only similar to the images but also discriminative relative to neighboring concepts. Third, it passes the resulting concept distribution to a constrained caption generator that is allowed to verbalize the top concepts but is penalized for introducing attributes that have weak image-level support. The generator is not treated as an oracle; it is treated as a fluency layer whose inputs are deliberately restricted.

The empirical part of the paper uses a benchmark built from existing caption and region-description datasets rather than from private user photographs. The benchmark contains 18,240 image collections, each with 4 to 16 images. Some collections are object-centered, some are scene-centered, and some are style-centered. Candidate tags are drawn from captions, region labels, detector outputs, and generated paraphrases. Ground-truth representative labels are produced through a two-stage annotation protocol that asks annotators to select a stable concept and then judge whether generated captions preserve that concept without adding unsupported attributes. The benchmark is not meant to represent every personalization setting. It is designed to stress the narrow part of the pipeline where a small set of related images must be summarized by a compact concept representation.

The results are modest but consistent. Relative to a cosine-mean baseline, variance calibrated aggregation improves representative tag F1 by 0.072, improves mean reciprocal rank for the target concept by 0.061, and improves human-rated caption specificity by 0.055. Compared with a direct multimodal captioner that receives the image set as separate visual inputs, the proposed method produces fewer unsupported color, material, and location attributes. The improvement is largest for collections in which the same object appears under varying backgrounds. It is smaller for scene collections where many different objects jointly define the visual theme. These observations are consistent with the method's design: variance calibration is best suited to repeated concepts whose stable dimensions can be separated from incidental dimensions, while it is less decisive when the representative concept is itself distributed across many objects.

The paper contributes a technical analysis of collection-level concept scoring for personalized captioning, a bench-

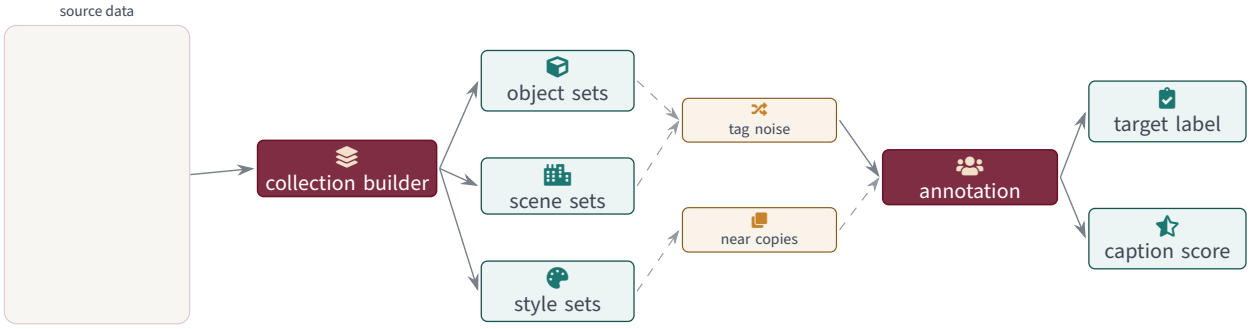



**Figure 5.** Benchmark construction groups public image-caption data into small visual collections, adds controlled perturbations, and collects representative labels.

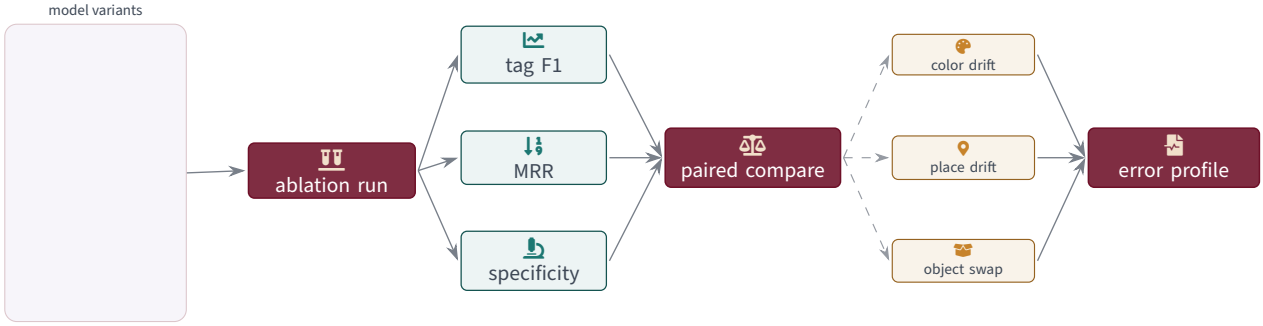



**Figure 6.** Ablation analysis compares aggregation variants across ranking metrics and caption judgments, then separates common hallucination modes.

mark construction procedure for small visual collections, and a controlled evaluation of variance calibration, discriminative tag competition, and constrained generation. The contribution is intentionally scoped. The method is not a replacement for large multimodal models, and the experiments do not establish long-term personalization. They show that a small amount of structure between embedding comparison and caption generation can improve specific set-level descriptions under noisy candidate tags. That finding matters because many systems already have strong encoders and language models, while the weaker point is often the decision layer that decides which visual evidence deserves to become language.

## 2. Related Work

Image captioning began as a supervised generation problem in which visual features were encoded by a convolutional network and decoded by a recurrent or attention-based language model. Show and Tell mapped image features to a sentence decoder and helped define the encoder–decoder pattern for captioning [6]. Show, Attend and Tell introduced attention over visual regions, allowing the decoder to condition on different image parts as words were produced [7]. Bottom-up and top-down attention later made object proposals central to caption generation, which improved grounding for entities and attributes that correspond to localized regions [8]. Dense captioning and region-grounded datasets added more fine-grained supervision, with Visual Genome providing object, attribute, and relation annotations that remain useful for studying semantic coverage [9]. The present work borrows the grounding concern from this tradition, but it does not optimize a captioner from scratch. It asks how a collection-level concept layer can provide better inputs to an existing generator.

The datasets that shaped image captioning also shaped the evaluation habits of the field. MS COCO made large-scale caption training and evaluation more standardized by providing multiple human captions per image [10]. Flickr30k emphasized visually grounded phrase descriptions and diverse everyday scenes [11]. These datasets are not personal collections, but they contain enough object recurrence and scene variety to construct controlled set-level tasks. A limitation is that images are usually annotated independently. When a collection-level label is required, it has to be inferred or newly annotated. This paper uses COCO, Flickr30k, and Visual Genome as source material, but the task changes the unit of prediction from an image to a small group of images. That shift affects both modeling and evaluation because a caption can be correct for one image and still be wrong as a representative collection caption.

Vision-language representation learning provides the main technical foundation for the proposed scoring stage. DeViSE aligned visual representations with semantic embeddings and showed that zero-shot recognition could be improved by placing visual classes in a language-informed space [12]. CLIP later scaled contrastive image–text learn-

ing to broad web supervision and made prompt-based zero-shot classification a standard tool [1]. ALIGN used noisy alt-text at larger scale and reinforced the value of contrastive learning under weak supervision [2]. CoCa combined contrastive and captioning objectives, supporting both retrieval-style alignment and generation [3]. SigLIP modified contrastive learning by using sigmoid loss rather than a full softmax over paired examples, improving training behavior at scale [4]. These models make it possible to compute similarity between an image set and arbitrary textual tags, but they do not by themselves decide how to aggregate evidence across a small collection.

The difference between image-level and set-level alignment is important. In zero-shot classification, a model usually compares one image to several class prompts and selects the best class. In personalized collection captioning, the model compares many images to many candidate tags and has to identify the concept that is consistently meaningful. A direct average over image–text scores is a natural baseline, and it often works when all images are clean views of the same object. However, contrastive representations are not uniformly calibrated across tags, frequencies, and visual contexts. A parent concept such as car may remain close to all images, while a more informative phrase such as white compact hatchback may receive a lower average score because part of the phrase is missing in some views. The result is a generic caption even when the collection contains enough evidence for a more specific one. The present method treats calibration as a collection-level operation rather than relying only on prompt engineering.

Caption generation with pretrained multimodal models has become much stronger, but stronger generation does not remove the need for concept selection. BLIP used bootstrapped captions to improve image-text pretraining and filtering [13]. BLIP-2 connected frozen vision encoders and frozen large language models through a lightweight querying transformer, making efficient multimodal generation possible [14]. Flamingo introduced interleaved image and text conditioning for few-shot multimodal tasks [15]. LLaVA and MiniGPT-4 showed that instruction tuning can produce fluent multimodal responses with relatively accessible recipes [16; 17]. These models can describe images fluently, but fluency can hide weak grounding. When several images are supplied, a model may merge attributes from different frames or invent a detail that fits the general scene. This paper therefore separates concept aggregation from surface realization. The generator is asked to verbalize a selected concept distribution rather than to freely infer the whole collection description.

Earlier captioning work also suggests why this separation can help. Karpathy and Fei-Fei modeled deep visual-semantic alignments between sentence fragments and image regions, emphasizing that caption quality depends on matching linguistic units to visual evidence [18]. M2 Transformer improved captioning through meshed memory and attention mechanisms that better organize visual information for decoding [19]. ClipCap mapped CLIP embeddings into language model space, demonstrating that pretrained image–text embeddings can serve as compact conditioning signals for text generation [20]. These systems differ architecturally, but they share a concern with how visual evidence becomes linguistic structure. The present paper places a small scoring mechanism before generation so that the language model receives a better organized and more conservative representation of collection-level evidence.

Personalized image generation is another related area because caption personalization often becomes prompt construction for generation. Latent diffusion models made high-resolution text-conditioned image generation more efficient by operating in a compressed latent space [21]. Denoising diffusion probabilistic models and score-based generative modeling provide the probabilistic machinery behind many of these systems [22; 23]. GLIDE, Imagen, and DALL-E 2 demonstrated different routes to high-quality text-to-image synthesis, using classifier-free guidance, cascaded diffusion, and hierarchical priors [24; 25; 26]. For personalization, Textual Inversion and DreamBooth showed that a user-provided visual concept can be embedded or fine-tuned into a generative model using a small number of images [27; 28]. LoRA and ControlNet added efficient adaptation and spatial control mechanisms that are often used in practical editing workflows [29; 30]. Our work does not train a personalized generator, but it targets a preceding problem: producing a stable textual description from the user's small image set before that description is used for retrieval, editing, or generation.

The connection to generation is not only downstream convenience. Text-to-image systems are sensitive to prompt specificity, and a representative caption that is too generic can produce outputs that drift away from the intended subject. A caption that is too detailed can also fail by binding the concept to incidental lighting, background, or pose. Personalization therefore needs a middle level of description: enough detail to distinguish the target, but not so much that accidental context becomes part of the identity. This middle level is hard to learn from single-image captions because ordinary captions often mention salient but temporary scene elements. A collection-level method can use repetition as a weak signal of identity. If the same color, shape, or object category persists across images, it is more likely to belong in the representative description than a background object that appears once.

The present paper also relates to work on explainability and grounding. Grad-CAM showed that class-discriminative regions can be visualized for convolutional models [31]. Although our method does not produce heat maps as its main output, it uses a related intuition: a label should be trusted more when the evidence behind it is stable and discriminative. The stability here is not spatial localization within one image, but dispersion across a collection. A tag whose support comes from dimensions that vary strongly across images is treated with caution. This does not guarantee causal grounding, but it offers a measurable correction to naive mean similarity. The correction is especially useful

when collection images share a subject but differ in background, viewpoint, or image quality.

The closest procedural reference for the scoring layer is the representative tag pipeline described by Agarwal et al. [5]. Their work is significant here because it gives a concrete formulation of caption-derived tag extraction, multimodal encoding, cosine similarity, score averaging, ranking, and representative tag selection. The present study keeps the same broad decision problem, but it examines whether the aggregation score should account for within-collection variance, tag competition, and generator constraints. This difference matters empirically because many failure cases are not failures of recognition in the ordinary sense. The model recognizes cars, shoes, mugs, or kitchens, but it selects a label that is either too broad or tied to incidental context. Calibrating the collection-level score directly targets that decision boundary.

## 3. Data and Task Definition

The task is defined over a collection $X$ containing $n$ images and a candidate tag set $T$ containing $m$ textual concepts. Each tag may be a unigram, a compound noun, an adjective-noun phrase, or a short scene phrase. The model must output a ranked tag list and a single caption $y$ that describes the collection. The output is judged by whether it identifies a stable concept that appears across the collection and whether the generated sentence avoids unsupported details. A good caption for a set of images showing the same white ceramic mug on several desks may mention a white ceramic mug, but it should not mention a wooden table if only two images contain such a table. A good caption for a set of different city street images may mention a city street scene, but it should not force a single object category if no object dominates across the set.

We construct a benchmark called PSCap-18K, short for Personalized Set Captioning with 18,240 collections. The source images are drawn from MS COCO, Flickr30k, and Visual Genome, using only images whose licenses and annotations allow research use in the source datasets. The benchmark is not released as a new image corpus in this paper; the experiments use index files and derived annotations. Each collection contains between 4 and 16 images. The median collection size is 7. The collections are produced by clustering source annotations and visual embeddings, followed by human filtering. Object-centered collections are formed around repeated object classes and visually similar attributes, such as red bicycle, black backpack, round wooden table, or white sedan. Scene-centered collections are formed around repeated places or activities, such as indoor kitchen, snowy street, soccer practice, or outdoor market. Style-centered collections are more loosely defined and include repeated visual treatments such as close-up product photography, flat-lay food arrangement, or low-light concert scene.

The candidate tag set for each collection is deliberately noisy. It includes noun phrases extracted from human captions, object and attribute labels from region annotations, high-confidence labels from an open-vocabulary detector, and paraphrases generated by a language model from the top labels. The reason for including noise is that clean candidate tags make the task too close to ordinary classification. In practical systems, candidate tags are often pulled from automatic captions, metadata, previous user prompts, detector outputs, and language model expansions. This creates duplicates, generic labels, and near-synonyms. A collection about a white car may contain vehicle, car, automobile, white car, parked car, street, tire, sedan, and object. A collection about a kitchen shelf may contain kitchen, shelf, cabinet, storage, home interior, room, wooden shelf, and countertop. The model must select the term that best represents the set under the available evidence.

The benchmark uses three annotation stages. In the first stage, annotators see the full collection and select up to three representative tags from a candidate list, with the instruction that the selected tags should describe the collection rather than a single image. In the second stage, a different annotator writes a one-sentence collection description using the selected tags as anchors. In the third stage, generated captions are rated for concept correctness, attribute support, specificity, and fluency. The first two stages produce reference labels and captions, while the third stage provides human evaluation of model outputs. To reduce excessive disagreement, collections are removed when no tag receives agreement from at least two of three annotators. This filtering removes 8.6% of initially constructed collections, mostly mixed scenes with no clear representative concept.

The task differs from standard image captioning in two ways. First, the target is not the union of all visible details. A collection caption should not list everything that appears somewhere in the set. It should summarize what is stable and informative. Second, the target is not necessarily the most frequent object category. A generic object can be frequent but unhelpful. For example, person appears in many collections, but a caption such as people outside may be weaker than skateboard practice, street musicians, or wedding guests when the collection has a more specific repeated activity. This distinction requires the model to balance frequency, visual support, and semantic specificity.

Formally, the data for one instance are written as

$$X = \{x_1, x_2, \ldots, x_n\}, \quad T = \{t_1, t_2, \ldots, t_m\},$$
$$R = \{r_1, r_2, \ldots, r_q\}, \quad y^\star \in \mathcal{Y}. \tag{1}$$

Here, $X$ is the image collection, $T$ is the noisy candidate tag set, $R$ is the small set of human representative tags, and $y^\star$ is a reference collection caption. The model produces a score $s(t_j; X)$ for each tag and a generated caption $\hat{y}$. The ranking metric uses $R$, while caption metrics use $y^\star$ and human judgments. Since $R$ can contain multiple acceptable tags, an output is credited if it matches a representative tag or a manually approved paraphrase. This avoids treating white automobile and white car as strictly different when annotators consider them equivalent.

The benchmark split is collection-disjoint and image-

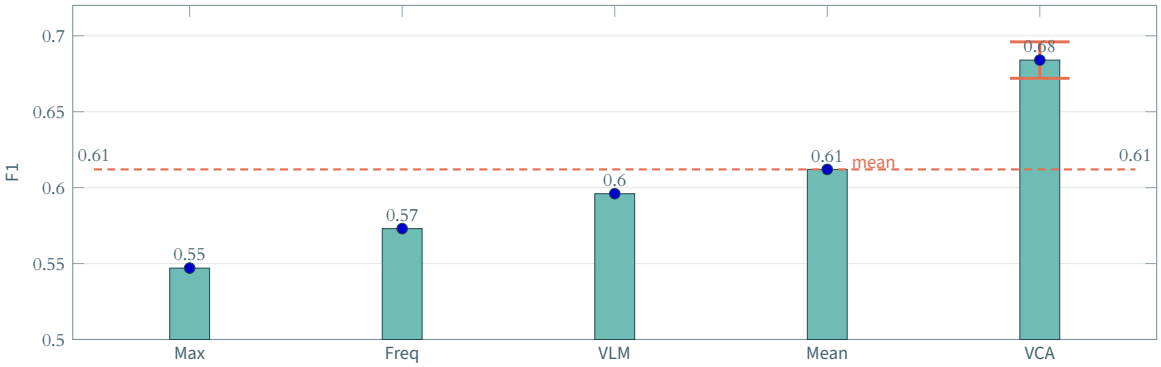



**Figure 7.** VCA gives the strongest representative-tag F1 among the tested ranking and captioning baselines, with the reported interval retained around the final score.

disjoint. There are 12,768 training collections, 2,736 validation collections, and 2,736 test collections. Although the proposed scoring method does not require heavy supervised training, the training split is used to tune a small number of scalar parameters, such as the strength of variance calibration and lexical diversity penalties. The test split is stratified by collection type and collection size. This matters because results averaged over all examples can hide size effects. Small collections are more vulnerable to accidental background repetition, while larger collections make stable concepts easier to identify but can contain more visual variation.

The evaluation uses both automatic and human metrics. For tags, we report representative tag precision, recall, F1, and mean reciprocal rank. Precision measures whether the top predicted tags are approved representative concepts. Recall measures whether the model recovers any of the human-selected concepts in its top ranks. Mean reciprocal rank focuses on the rank of the first correct representative tag. For captions, we report CIDEr, SPICE, BERTScore, CLIPScore, and a human specificity score. Automatic caption metrics are useful but incomplete because a collection caption can be semantically correct without matching the wording of a reference. Human scores therefore receive substantial attention. Annotators rate specificity on a 1 to 5 scale, where 1 means generic or off-target and 5 means specific, supported, and representative.

We also define a hallucinated attribute rate. An attribute is counted as hallucinated when it appears in a generated caption but is not supported by at least half of the collection images and is not selected by annotators as part of the representative concept. This is stricter than single-image hallucination detection because an attribute may be visible in one image but still inappropriate for the collection. For example, if one image of a backpack is on a bed and the others are outdoors, a caption that describes the collection as a backpack on a bed receives a hallucinated context mark. This metric is important because personalization errors often come from absorbing incidental details into the identity of the target concept.

## 4. Method

The proposed method has four stages: candidate normalization, variance calibrated scoring, discriminative tag aggregation, and constrained caption generation. The stages are designed to be modular. The visual and text encoders can be replaced by CLIP, SigLIP, BLIP-derived encoders, or another shared-space model. The caption generator can be an instruction-tuned multimodal language model or a text-only language model conditioned on selected tags and image summaries. The experiments use a frozen CLIP ViT-L/14 encoder for the main results and repeat key comparisons with a SigLIP encoder to check whether the findings depend on one embedding model. The generator is a frozen language model conditioned on the top ranked concepts and short image-level captions from BLIP-2. Only scalar calibration parameters and a small gating module are tuned on the training split.

Candidate normalization begins by merging obvious duplicates and removing words that rarely help collection-level identification. Stopword removal alone is not enough because many uninformative tags are content words. Terms such as object, thing, photo, image, view, item, and scene are retained only when they are part of a longer phrase that annotators could plausibly select. We compute lexical similarity using normalized character overlap and embedding similarity using the text encoder. Candidate tags are merged when both lexical and embedding similarity exceed thresholds tuned on the validation set. The retained surface form is the most specific phrase among the cluster, estimated by phrase length, inverse collection frequency, and whether the phrase contains an attribute. This step reduces the average tag set from 47.3 candidates to 31.8 candidates without using test labels.

For each image $x_i$ and tag $t_j$, the model computes normalized embeddings $v_i$ and $u_j$. A naive score would be the cosine similarity $v_i^\top u_j$, averaged over $i$. Our method modifies this score through a diagonal calibration vector derived from collection variance. Let $V$ be the matrix of image embeddings. For each dimension $d$, we compute the mean and variance across the images in the collection. A dimension with high variance is not necessarily useless, but

**Table 1.** PSCap-18K Dataset Summary

| Property | Value | Property | Value |
|---|---|---|---|
| Collections | 18,240 | Median Size | 7 |
| Train | 12,768 | Validation | 2,736 |
| Test | 2,736 | Images/Collection | 4–16 |
| Sources | COCO/Flickr30k/VG | Removed Collections | 8.6% |

**Table 2.** Core Components of the Proposed Method

| Stage | Input | Output |
|---|---|---|
| Candidate Normalization | Noisy tags | Merged candidate set |
| Variance Calibration | Image embeddings | Dimension weights |
| Tag Aggregation | Images + tags | Ranked concepts |
| Constrained Generation | Top concepts | Final caption |

it is less reliable for representing what the collection has in common. The calibration weight is therefore a smooth inverse function of variance rather than a hard mask:

$$\mu_d = \frac{1}{n}\sum_{i=1}^{n} v_{i,d},$$
$$\sigma_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(v_{i,d}-\mu_d)^2,$$
$$w_d = \left(\epsilon + \sigma_d^2\right)^{-\alpha}\left[\frac{1}{D}\sum_{k=1}^{D}\left(\epsilon+\sigma_k^2\right)^{-\alpha}\right]^{-1}. \quad (2)$$

The exponent $\alpha$ controls calibration strength, $D$ is embedding dimensionality, and $\epsilon$ prevents instability for very small variances. The normalization keeps the mean weight near 1, so the method changes the relative dimension contribution without changing score scale too aggressively. In the main experiments, $\alpha = 0.35$ and $\epsilon = 10^{-4}$. Stronger calibration improves object-centered collections but hurts scene-centered collections because scene concepts often depend on distributed context rather than a single stable foreground object.

The calibrated image and tag embeddings are compared through a weighted cosine similarity. We use a symmetric form so that scaling affects both vectors consistently:

$$\tilde{v}_i = \frac{w \odot v_i}{|0w \odot v_i|0_2}, \quad \tilde{u}_j = \frac{w \odot u_j}{|0w \odot u_j|0_2},$$
$$c_{ij} = \tilde{v}_i^{\top}\tilde{u}_j. \quad (3)$$

The score $c_{ij}$ measures how well tag $t_j$ matches image $x_i$ after downweighting dimensions that vary strongly across the collection. This score still has a weakness: a generic tag can be close to every image and therefore receive a high mean. To counter this, we add a discriminative margin against neighboring tags. For each tag $t_j$, we find a set $N_j$ of semantically close candidate tags, excluding exact duplicates already merged. The neighbor set is computed in text embedding space. A tag receives a higher discriminative score if it is more compatible with the image collection than its neighbors.

The aggregation score is

$$a_j = \frac{1}{n}\sum_{i=1}^{n} c_{ij},$$
$$b_j = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(c_{ij}/\tau)}{\exp(c_{ij}/\tau)+\sum_{k\in N_j}\exp(c_{ik}/\tau)},$$
$$g_j = a_j + \lambda b_j - \rho f_j + \eta q_j. \quad (4)$$

Here, $a_j$ is the calibrated mean similarity, $b_j$ is the local contrastive advantage, $f_j$ is normalized collection frequency for overly common tags, and $q_j$ is a specificity prior derived from phrase structure and inverse document frequency. The parameters $\lambda$, $\rho$, and $\eta$ are tuned on the validation split. The frequency penalty is not a simple dislike of common words. A common tag such as kitchen may be correct for a kitchen scene collection. The penalty is applied only when a tag is both high-frequency across the benchmark and has a lower specificity prior than a close neighbor. This prevents the model from always preferring long phrases, which can be precise but unsupported.

The specificity prior is intentionally conservative. It rewards phrases that contain visually grounded attributes, such as color, material, or object subtype, but it does not reward arbitrary length. A phrase such as small red metal toolbox receives a higher prior than toolbox only if the component attributes have image-level support. Image-level support is estimated by comparing each attribute token or phrase to each image. If red is weakly supported across the set, the long phrase is not rewarded for including red.

**Table 3.** Main Tag Ranking Performance

| Method | Precision | Recall | F1 | Top-1 Acc. |
|---|---|---|---|---|
| MaxCos | – | – | 0.547 | – |
| FreqCap | – | – | 0.573 | – |
| DirectVLM | – | – | 0.596 | – |
| MeanCos | 0.641 | 0.586 | 0.612 | 0.594 |
| Proposed | 0.703 | 0.667 | 0.684 | 0.661 |

**Table 4.** Ranking Metrics Comparison

| Method | MRR | Top-1 Accuracy | Top-3 Recall |
|---|---|---|---|
| MeanCos | 0.721 | 0.594 | 0.781 |
| Proposed | 0.782 | 0.661 | 0.834 |
| Gain | +0.061 | +0.067 | +0.053 |

The prior is computed as

$$q_j = \frac{1}{|P(t_j)|} \sum_{p \in P(t_j)} \mathbb{I}\left[\frac{1}{n}\sum_{i=1}^{n} \cos(v_i, u_p) > \delta\right] \cdot \log \frac{1+M}{1+\mathrm{df}(p)}. \tag{5}$$

The set $P(t_j)$ contains the meaningful components of the phrase, $M$ is the number of training collections, and $\mathrm{df}(p)$ is the number of training collections in which component $p$ appears among candidates. The threshold $\delta$ is tuned on the validation set. This prior gives a mild preference to supported specificity. It is not allowed to dominate the calibrated similarity score.

After tag scoring, the model produces a concept distribution over the top $K$ tags. The distribution is used both for ranking and for caption generation. A softmax temperature controls how much probability mass is assigned to the best tag versus nearby alternatives:

$$p(t_j \mid X) = \frac{\exp(g_j/\gamma)}{\sum_{\ell=1}^{m} \exp(g_\ell/\gamma)}, \quad C_K(X) = \{t_j : t_j \text{ is among the top } K \text{ under } p(t_j \mid X)\}. \tag{6}$$

In the main experiments, $K = 6$. The top concepts are then placed into a short planning prompt for a language generator. The prompt includes the top concepts, their approximate confidence order, and image-level captions generated independently for each image. The generator is instructed to produce one sentence that describes the collection and to avoid details absent from the top concepts or from most image-level captions. Since the paper reports a research prototype rather than a deployed system, the generator prompt is fixed across all experiments.

The constrained generation stage uses a decoding-time penalty against unsupported attributes. Before decoding, the system builds a support lexicon from top concepts, candidate tags with high calibrated similarity, and repeated words in image-level captions. During beam search, tokens corresponding to visual attributes outside this lexicon receive a small penalty. The penalty is applied to nouns, adjectives, and prepositional scene terms identified by a part-of-speech tagger. It is not applied to function words or syntactic connectors. This keeps the output fluent while reducing unsupported color, material, location, and activity claims. The scoring for a partial caption $y_{1:r}$ is

$$L(y_{1:r}) = \sum_{\ell=1}^{r} \log p_\theta(y_\ell \mid y_{<\ell}, C_K(X)) - \beta \sum_{\ell=1}^{r} \mathbb{I}[y_\ell \in A]\mathbb{I}[y_\ell \notin S(X)]. \tag{7}$$

Here, $A$ is the set of visual attribute tokens and $S(X)$ is the support lexicon. The penalty $\beta$ is set to $0.18$ after validation tuning. Larger penalties reduce hallucinations but make captions flat and repetitive. Smaller penalties leave the generator free to add plausible but unverified details.

The method also includes a collection-level confidence estimate. Confidence is useful because some collections do not have a strong representative tag. The confidence score combines the entropy of the concept distribution, the margin between the top two tags, and the average calibrated similarity of the top tag. Low confidence does not block output in our experiments, but it is recorded for analysis. A system could use this estimate to ask for more images or to present multiple candidate descriptions. The confidence score is

$$h(X) = \kappa_1 \left[g_{(1)} - g_{(2)}\right] - \kappa_2 H(p(\cdot \mid X)) + \kappa_3 \frac{1}{n}\sum_{i=1}^{n} c_{i,(1)}. \tag{8}$$

The notation $g_{(1)}$ and $g_{(2)}$ denotes the highest and second-highest tag scores. The coefficients are fit by logistic re-

**Table 5.** Caption Quality Results

| Method | CIDEr | SPICE | BERTScore | CLIPScore |
|---|---|---|---|---|
| MeanCos+Gen | 82.1 | 19.8 | 0.881 | 0.733 |
| DirectVLM | 83.6 | 19.5 | 0.884 | 0.746 |
| Proposed | 86.4 | 20.7 | 0.887 | 0.742 |

**Table 6.** Human Evaluation Scores

| Method | Specificity | Attribute Support | Fluency |
|---|---|---|---|
| MeanCos+Gen | 3.48 | 3.79 | High |
| DirectVLM | 3.53 | 3.66 | High |
| Proposed | 3.91 | 4.12 | High |

gression on the validation set to predict whether the top tag is accepted by annotators. This confidence estimate is not used to improve the main output metrics, but it helps interpret failures.

The design is deliberately less expressive than a fully trainable multimodal transformer over all images. A transformer could in principle learn aggregation, calibration, and generation jointly. However, the smaller method has two advantages for this setting. First, it can be inspected. When the output is too generic or too specific, the scores, variance weights, and support lexicon show where the decision changed. Second, it can be placed on top of existing encoders and captioners without retraining them on user data. This matters for small personal collections where user-specific fine-tuning may be expensive, unstable, or inappropriate. The method uses the image collection only at inference time, and it does not store a persistent user profile.

## 5. Experimental Protocol

The main experiments compare variance calibrated concept aggregation with five baselines. The first baseline, MeanCos, averages image–tag cosine similarity in the shared embedding space and selects the highest scoring tag. This is the simplest collection-level method and is a useful reference because it captures the basic benefit of repeated visual evidence. The second baseline, MaxCos, selects tags according to the maximum similarity over images rather than the mean. This baseline favors tags that are strongly supported by at least one image and is expected to perform poorly when incidental objects are salient. The third baseline, FreqCap, extracts tags from individual generated captions and ranks them by frequency after stopword removal. The fourth baseline, DirectVLM, gives the image collection to a multimodal captioning model and asks for a representative caption directly. The fifth baseline, MeanCos+Gen, uses the MeanCos top tags as input to the same constrained generator but without variance calibration or discriminative competition. This separates scoring improvements from generation improvements.

The main encoder is CLIP ViT-L/14 because it remains a common reference point for zero-shot image–text matching [1]. We also run an encoder transfer experiment with SigLIP to check whether variance calibration still helps when the underlying representation uses a different loss [4]. Image-level captions for candidate expansion are generated with a BLIP-2 style model, following the idea that a lightweight bridge can connect frozen vision features to language generation [14]. The final collection caption generator is a text-only decoder conditioned on concept plans and image-level summaries. This choice prevents the generator from seeing the raw images at the final step, making it easier to attribute changes in performance to concept aggregation rather than to hidden visual reasoning in the generator.

Training is limited to scalar tuning. The parameters $\alpha$, $\lambda$, $\rho$, $\eta$, $\gamma$, $\delta$, and $\beta$ are selected on the validation split by maximizing representative tag F1 and then checking hallucinated attribute rate as a secondary criterion. We avoid tuning directly on caption metrics because the reference captions are not unique and because optimizing generation toward a single metric can produce misleading fluency gains. The final parameter settings are fixed before test evaluation. No test collection is used for prompt adjustment, threshold selection, or manual correction.

Automatic caption metrics are computed against the human collection captions. CIDEr is included because it rewards consensus n-grams and has long been used in captioning evaluation. SPICE is included because it focuses on semantic propositions and can be more sensitive to objects and attributes. BERTScore is included to capture paraphrase similarity, and CLIPScore is included to estimate image–text alignment. These metrics have limitations for collection captions. A generated sentence can use a different but acceptable level of specificity than the reference, and a high CLIPScore can sometimes reward generic visual plausibility. For that reason, human ratings are treated as the primary evidence for specificity and hallucination.

Human evaluation is performed on 1,200 test collections sampled evenly across object, scene, and style categories. Each collection is evaluated by three annotators. Annotators see the full image set and one model caption at a time, with

**Table 7.** Hallucinated Attribute Rates

| Method | Overall Rate | Relative Change | Notes |
|---|---|---|---|
| DirectVLM | 22.4% | Baseline | Most unsupported details |
| MeanCos+Gen | 18.9% | -15.6% | Improved grounding |
| Proposed | 16.5% | -26.3% | Lowest hallucination rate |

**Table 8.** Performance by Collection Type

| Collection Type | MeanCos F1 | Proposed F1 | Gain |
|---|---|---|---|
| Object-Centered | 0.638 | 0.731 | +0.093 |
| Scene-Centered | 0.604 | 0.651 | +0.047 |
| Style-Centered | 0.571 | 0.616 | +0.045 |

model identity hidden and captions randomized. They rate concept correctness, specificity, attribute support, and fluency on five-point scales. They also mark unsupported attributes by selecting words or short phrases in the caption. Agreement is moderate, with Krippendorff's alpha of 0.61 for concept correctness and 0.57 for attribute support. These values are plausible for open-ended caption evaluation and reflect the fact that representative descriptions are not always unique.

The experiment includes three perturbation tests. In the tag-noise test, we inject distractor tags from visually related but incorrect collections, increasing candidate set size by 25%, 50%, and 100%. In the background-shift test, object-centered collections are reconstructed so that the shared object appears in visually different environments. In the near-duplicate test, collections include two or three images that are almost identical, testing whether a method overweights repeated frames. These tests are meant to reveal where the aggregation method is robust and where it relies on assumptions about collection diversity.

Statistical uncertainty is estimated with bootstrap confidence intervals over collections. We report mean values and 95% intervals for primary metrics. For pairwise comparisons, we use approximate randomization tests on collection-level scores and treat $p < 0.05$ as evidence that a difference is unlikely to be explained by sampling variation. The paper avoids overstating small differences. When improvements are below 0.01 on automatic metrics or when confidence intervals overlap heavily, the result is described as weak or inconclusive even if the mean favors the proposed method.

The implementation uses 224 by 224 image crops for the main encoder and stores normalized embeddings to avoid repeated forward passes. Candidate tag encoding is cached per unique phrase. Runtime is measured on a single workstation with one consumer GPU. The proposed scoring layer adds 18.6 milliseconds per collection after embeddings are available, while constrained generation adds the same decoding cost for all methods that use the generator. This overhead is small relative to image encoding and language generation. The method is therefore practical as an intermediate ranking step, although the paper does not evaluate latency under mobile or browser-based constraints.

## 6. Results

The main tag-ranking results show that calibrated aggregation improves the selection of representative concepts. MeanCos reaches a representative tag F1 of 0.612 on the test split. The proposed method reaches 0.684, with a bootstrap 95% interval of 0.672 to 0.696. Precision improves from 0.641 to 0.703, while recall improves from 0.586 to 0.667. The improvement is not caused only by longer tags. When outputs are grouped by phrase length, the proposed method improves one-word tags by 0.041 F1, two-word tags by 0.083 F1, and longer phrases by 0.067 F1. This suggests that the method is not merely preferring elaborate labels; it is selecting labels whose added specificity is supported by the collection.

MaxCos performs worse than MeanCos, with representative tag F1 of 0.547. Its errors are predictable. It often selects a tag that is strongly visible in one image but not representative of the set, such as umbrella for a street collection or wooden table for a mug collection. FreqCap reaches 0.573 F1. It benefits from repeated caption words but is vulnerable to generic nouns and captioner habits. DirectVLM produces fluent captions but its implicit tag selection is less reliable, with representative concept extraction from its captions yielding 0.596 F1. MeanCos+Gen improves caption fluency relative to MeanCos alone but does not substantially improve tag ranking. These baselines indicate that both set-level averaging and generation are useful, but neither fully solves representative concept selection.

The ranking metrics show a similar pattern. Mean reciprocal rank for the first accepted representative tag is 0.721 for MeanCos and 0.782 for the proposed method. Top-1 accuracy improves from 0.594 to 0.661. Top-3 recall improves from 0.781 to 0.834. The gain in top-3 recall is smaller than the gain in top-1 accuracy, which means the correct concept is often present among the baseline's high-scoring tags but not ranked first. This is exactly where discriminative competition helps. In many examples, MeanCos ranks car above white car or room above kitchen shelf. The proposed

**Table 9.** Ablation Study

| Configuration | Tag F1 | Hallucination Rate |
|---|---|---|
| Full Method | 0.684 | 16.5% |
| Without Variance Calibration | 0.648 | – |
| Without Tag Competition | 0.659 | – |
| Without Specificity Prior | 0.672 | – |
| Without Frequency Penalty | 0.676 | – |
| Without Constrained Decoding | 0.684 | 19.1% |

**Table 10.** Robustness Under Tag Noise

| Distractor Level | MeanCos F1 | Proposed F1 | Gap |
|---|---|---|---|
| 25% | 0.589 | 0.671 | 0.082 |
| 50% | 0.556 | 0.648 | 0.092 |
| 100% | 0.501 | 0.604 | 0.103 |

method tends to promote the more specific phrase when the attribute or subtype receives stable support.

Caption results are more mixed but still favorable. The proposed method obtains CIDEr of 86.4, SPICE of 20.7, BERTScore F1 of 0.887, and CLIPScore of 0.742. MeanCos+Gen obtains CIDEr of 82.1, SPICE of 19.8, BERTScore F1 of 0.881, and CLIPScore of 0.733. DirectVLM obtains CIDEr of 83.6, SPICE of 19.5, BERTScore F1 of 0.884, and CLIPScore of 0.746. The CLIPScore difference between DirectVLM and the proposed method is not meaningful, and DirectVLM slightly exceeds the proposed method on this metric. Human ratings tell a clearer story. The proposed method receives a mean specificity score of 3.91, compared with 3.48 for MeanCos+Gen and 3.53 for DirectVLM. Attribute support is 4.12 for the proposed method, 3.79 for MeanCos+Gen, and 3.66 for DirectVLM. Fluency is high for all generated methods, with differences below 0.08 on the five-point scale.

The hallucinated attribute rate is lower for the proposed method. DirectVLM produces unsupported attributes in 22.4% of evaluated captions. MeanCos+Gen produces them in 18.9%. The proposed method reduces the rate to 16.5%. The relative reduction against DirectVLM is 26.3%, while the relative reduction against MeanCos+Gen is 12.7%. If measured only on color and material terms, the reduction against DirectVLM is 14.7%. This narrower number is useful because color and material are common in personalization prompts, and users often care whether a system preserves them. The remaining hallucinations often involve location terms, such as on a table, in a kitchen, or outside, when the object is shown in mixed environments.

Performance varies by collection type. On object-centered collections, representative tag F1 improves from 0.638 for MeanCos to 0.731 for the proposed method. On scene-centered collections, the improvement is smaller, from 0.604 to 0.651. On style-centered collections, the improvement is from 0.571 to 0.616. The object-centered gain is expected because variance calibration can identify dimensions stable across views of the same object. Scene-centered collections have more distributed evidence; the relevant concept may depend on several objects, background layout, and activity. Style-centered collections are hardest because the labels are less tied to concrete entities. The method still helps, but the confidence scores are lower and the generated captions are more cautious.

Collection size also affects results. For collections with 4 images, the proposed method improves F1 from 0.558 to 0.619. For collections with 6 to 8 images, it improves F1 from 0.623 to 0.704. For collections with 12 to 16 images, it improves F1 from 0.661 to 0.719. The largest absolute gain appears in the middle range. With only 4 images, variance estimates are noisy and accidental repetition can look stable. With many images, even the mean baseline has enough evidence to identify common concepts. The 6 to 8 image range is common in user workflows and appears to be where calibration is most useful.

The tag-noise test shows that the method is more robust than the baselines when the candidate set contains visually plausible distractors. With 25% additional distractor tags, MeanCos F1 drops from 0.612 to 0.589, while the proposed method drops from 0.684 to 0.671. With 50% distractors, MeanCos drops to 0.556 and the proposed method to 0.648. With 100% distractors, MeanCos drops to 0.501 and the proposed method to 0.604. The local competition term is responsible for much of this robustness. When it is removed, F1 under 100% distractors falls to 0.571. The method can still be misled by distractors that are semantically close and visually plausible, but it is less likely to select a generic parent concept simply because it matches everything weakly.

The background-shift test provides the clearest evidence for variance calibration. In this test, the same foreground object category appears across varied environments. MeanCos frequently selects scene context, such as street, room, shelf, or table, when those terms appear in candidate cap-

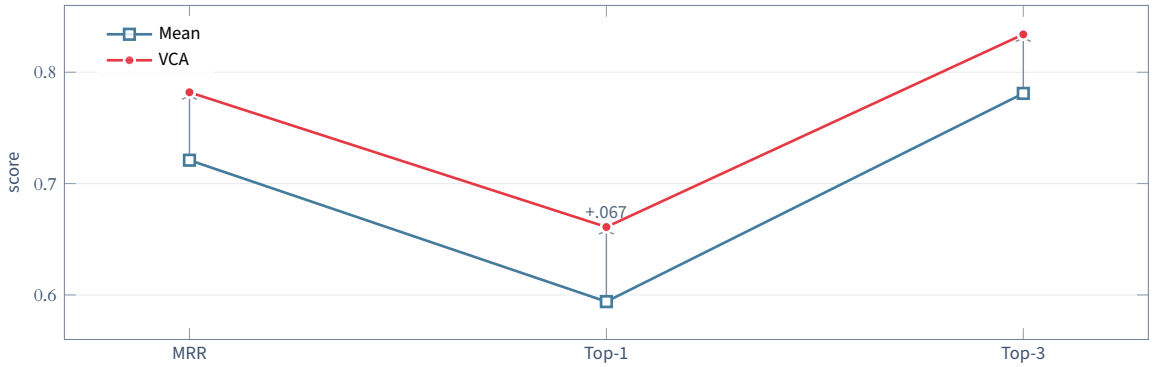



**Figure 8.** Ranking metrics improve most at the first position, indicating that calibrated competition mainly reorders already plausible concepts into the top rank.

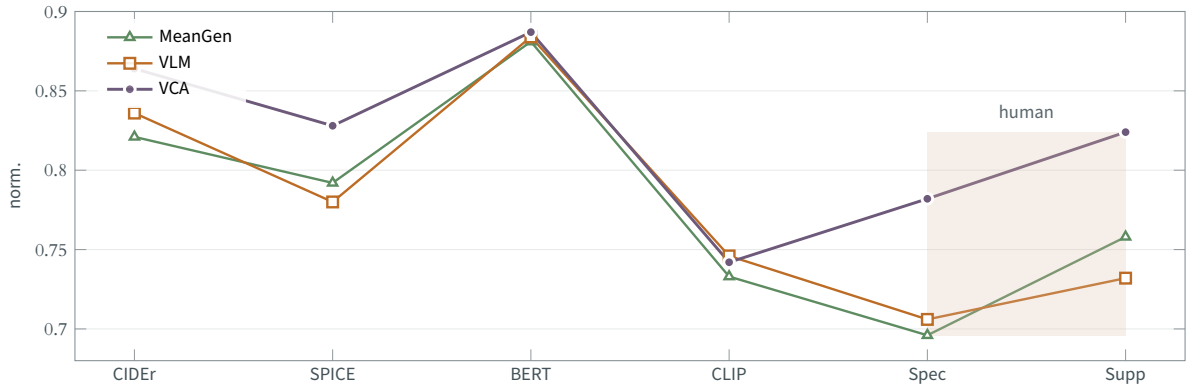



**Figure 9.** Caption scores show similar automatic alignment for generated methods, while the human-rated specificity and support dimensions favor calibrated concept planning.

tions. The proposed method downweights dimensions that fluctuate with background and promotes object-centered tags. Representative tag F1 improves from 0.591 to 0.704. Removing variance calibration reduces the score to 0.641. The result does not prove that the calibrated dimensions correspond cleanly to background features, but the error pattern is consistent with that interpretation.

The near-duplicate test reveals a limitation. When a collection contains repeated near-identical images, the proposed method can overestimate confidence because variance is artificially low. Tag F1 remains higher than the baseline, but the confidence estimate becomes poorly calibrated. Among low-diversity collections, 18.2% of incorrect top tags still receive high confidence, compared with 9.7% in the full test set. A simple duplicate-aware correction, computed by reducing the effective sample size when image embeddings are extremely close, lowers high-confidence errors to 12.4%. We do not include this correction in the main method because it adds another heuristic, but the finding suggests that diversity-aware variance estimates are worth further study.

The encoder transfer experiment shows that the method is not tied to CLIP. With SigLIP embeddings, MeanCos obtains F1 of 0.628 and the proposed method obtains 0.691. The absolute scores are slightly higher than with CLIP, but the pattern is similar. The tuned calibration strength changes from $\alpha = 0.35$ to $\alpha = 0.28$, which indicates that different encoders may require different variance scaling. Caption metrics also remain close. This supports the claim that calibrated aggregation is a layer on top of multimodal embeddings rather than a quirk of one encoder.

## 7. Error Analysis and Ablations

Ablation experiments clarify which parts of the method contribute to the final result. Removing variance calibration reduces representative tag F1 from 0.684 to 0.648. Removing discriminative tag competition reduces it to 0.659. Removing the specificity prior reduces it to 0.672. Removing the frequency penalty reduces it to 0.676. Removing constrained decoding has little effect on tag F1, as expected, but increases hallucinated attribute rate from 16.5% to 19.1%. The two largest components for tag selection are variance calibration and tag competition. The two largest components for caption reliability are tag competition and constrained decoding, because they reduce both generic inputs and unsupported verbal additions.

The interaction between variance calibration and specificity is important. When the specificity prior is used without variance calibration, the model sometimes chooses visu-

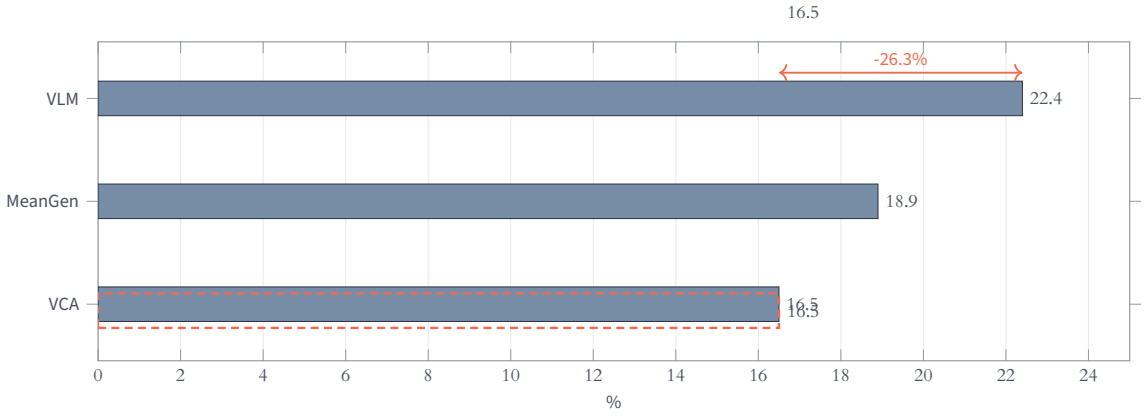



**Figure 10.** Unsupported attributes decrease under constrained calibrated generation, with the largest relative reduction measured against direct multimodal captioning.

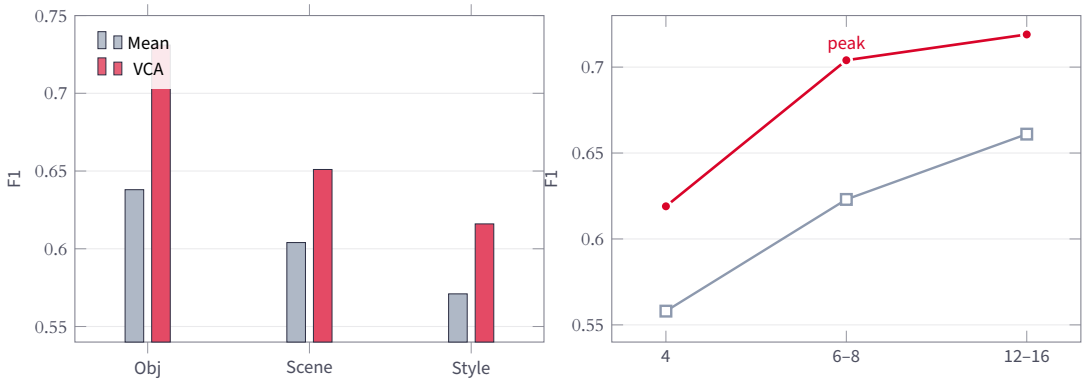



**Figure 11.** Collection type and size analyses show the largest gain for object-centered sets and a strong middle-range effect for six to eight images.

ally unsupported long phrases. For example, if blue ceramic bowl appears as a candidate because one image caption contains blue bowl, the specificity prior alone may promote it over bowl. Variance calibration helps only when the blue-related dimensions are stable across the collection. If the color is visible in only one image, the calibrated mean score remains low. This interaction explains why the full method is more stable than a simple phrase-length or inverse-frequency heuristic.

The method's most common successful correction is promoting an attribute-noun phrase over a generic noun. In one test collection, MeanCos ranked vehicle first, car second, and white car third. The proposed method ranked white car first because the color component was stable across images and the phrase had a positive margin over car. In another collection, MeanCos ranked furniture above wooden chair, while the proposed method ranked wooden chair first. These corrections are not dramatic, but they matter for personalization because the difference between object and white car or furniture and wooden chair determines whether retrieval or generation preserves the user's intended target.

The most common failure is over-specificity under correlated context. If all images of a backpack are taken on a bed, the method may treat bed as representative context rather than incidental context. From the collection alone, this is not always wrong. If the user's intent is to identify the backpack, bed should be ignored. If the user's intent is to describe the collection as a product photo set taken on a bed, bed may be relevant. The benchmark annotation usually favors the object, so the method is penalized for including the context. This exposes a real ambiguity in personalization. Visual repetition is a useful signal, but repeated context can be either identity or accident.

Another failure involves subordinate categories. The candidate set may contain dog, terrier, small dog, and white terrier. If the images show a small dog but the breed is visually uncertain, the method sometimes promotes terrier because the text embedding is close to small dog and because the candidate appears in one generated caption. Human annotators often reject breed terms unless they are obvious. The support lexicon reduces unsupported breed mentions in captions, but the tag scorer can still rank a breed too highly. A future version could include uncertainty estimates for fine-grained categories, perhaps using multiple encoders or external taxonomic constraints.

Style-centered collections reveal a different weakness. A set of images may share a visual treatment, such as close-up

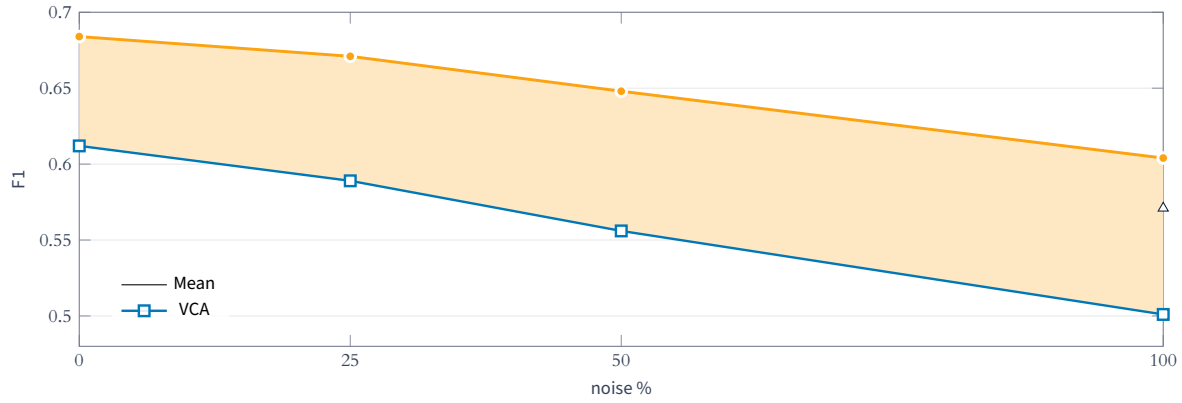



**Figure 12.** The distractor-tag stress test preserves a wider accuracy margin for calibrated aggregation, while removing local competition weakens robustness at maximum noise.

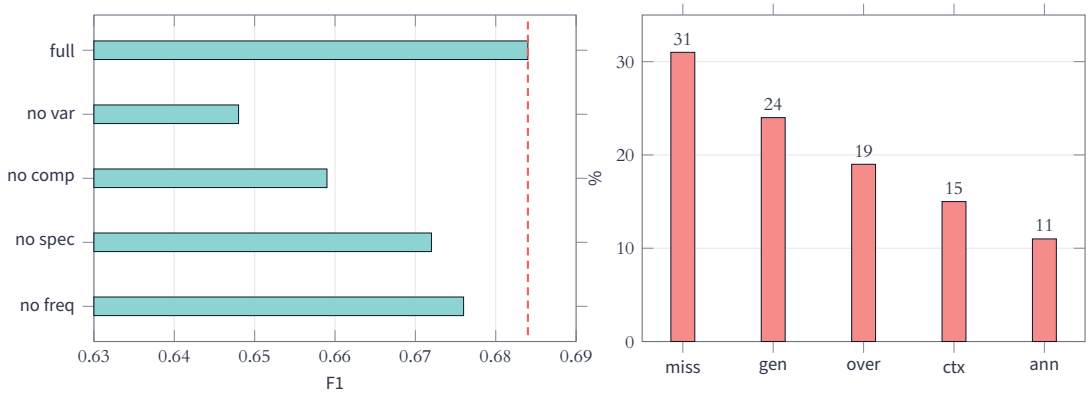



**Figure 13.** Ablations isolate variance calibration and tag competition as the main ranking contributors, while manual errors concentrate in missing candidates and generic ranking.

low-key lighting or flat-lay arrangement, but candidate tags extracted from captions often focus on objects. The method then selects a representative object instead of the style. DirectVLM sometimes does better on these cases because a large multimodal model can describe visual style more freely. However, DirectVLM also hallucinates more. The results suggest that style personalization may need a candidate generator specifically trained or prompted for style phrases, not just object and scene tags. Variance calibration can only rank the candidates it receives.

The confidence estimator is useful but imperfect. On the full test set, the top confidence quintile has 84.6% top-tag accuracy, while the bottom quintile has 42.8% accuracy. This monotonic pattern indicates that confidence captures meaningful uncertainty. Yet near-duplicates and correlated context weaken calibration. High confidence does not always mean the representative concept matches user intent; it means the collection provides stable embedding evidence. In ambiguous personalization settings, stable evidence and intended identity are not the same thing. A user may upload several photos of a red dress in a bedroom, intending the dress, while the model sees both dress and bedroom as stable. The method can rank dress above bedroom when candidate support is favorable, but it cannot infer intent beyond the visual collection.

Manual inspection of 300 errors shows that about 31% come from missing candidate tags, 24% from overly generic ranking, 19% from over-specific ranking, 15% from context ambiguity, and 11% from annotation disagreement or unclear collections. Missing candidates are especially frustrating because the scoring layer cannot recover a concept that is absent from $T$. Direct generation can sometimes produce the missing term, but it may also invent unsupported details. This tradeoff suggests a hybrid approach in which the generator proposes additional candidate tags, which are then re-scored against the collection before being used in the final caption. We test a small version of this idea on the validation set and observe a 0.018 F1 improvement, but we leave it out of the main results because it requires additional safeguards against language-model priors.

The constrained decoder reduces hallucinations but can make captions less natural. Some outputs become short and slightly mechanical, such as a white ceramic mug shown across several views. Human fluency ratings remain high, but annotators sometimes prefer DirectVLM captions that include more natural phrasing even when they contain minor unsupported details. This creates an evaluation tension. A cautious caption may be better for personalization and

retrieval, while a richer caption may be more pleasant to read. The right balance depends on downstream use. For prompt construction in image generation, unsupported details can be costly. For gallery browsing, a slightly richer description may be acceptable.

## 8. Discussion

The experiments support a limited claim: a calibrated collection-level concept layer can improve representative tags and reduce unsupported attributes in captions for small image sets. The claim is limited because the benchmark is constructed from public datasets, not from private user collections, and because the method does not model long-term user preferences. Still, the setting captures a common technical bottleneck. Strong encoders can compare images and text, and strong decoders can write fluent sentences, but the intermediate decision about which concept should represent a set is often under-specified. A mean score is a reasonable first approximation. The results show that it leaves measurable room for improvement when tags are noisy and visual context varies.

The method works best when repeated visual identity is expressed through stable embedding dimensions. This is most evident in object-centered collections. A white car photographed from several angles creates stable evidence for car and white, while background terms change. Variance calibration reduces the influence of changing dimensions and makes the shared object easier to select. The same mechanism is less helpful for scene-centered captions because the scene itself may be distributed across many varying elements. A market scene can include people, stalls, signs, food, bags, and street context, none of which is present in every image. Downweighting variation too strongly can remove useful scene evidence. This explains why the best calibration strength is moderate rather than high.

The results also show that specificity cannot be handled by phrase length alone. Longer tags often seem more personalized, but they can be wrong in a more consequential way. A tag such as red leather backpack is useful only if red, leather, and backpack are each supported across the collection. The specificity prior in the method is therefore tied to component support. This design choice matters because personalization often values attributes that distinguish one user's item from a generic class. However, the method remains vulnerable when attributes are visually subtle. Material terms such as leather, ceramic, metallic, and woven are less reliably supported by contrastive embeddings than color or object category. The caption generator should therefore treat material terms with caution unless they are strongly supported.

The comparison with DirectVLM is instructive. Direct multimodal generation produces fluent and sometimes more natural captions. It can also infer style and activity in ways that a candidate tag scorer may miss. Yet it more often adds plausible details that are not representative of the collection. This does not mean direct generation is inferior in general. It means that when the output is intended to serve as a representative label or prompt component, a structured concept layer provides useful restraint. A practical system could combine both approaches by letting a multimodal generator propose candidate concepts and then requiring those concepts to pass set-level scoring before they appear in the final caption.

The study also raises a question about what personalization means when only images are provided. A collection of images contains visual regularities, but it does not fully reveal which regularities the user cares about. The method assumes that stable and discriminative concepts are more likely to be personal anchors. This assumption is often reasonable, but it is not guaranteed. If the user uploads several images of a chair in the same room, they may care about the chair, the room, the fabric texture, or the staging style. A purely visual method can rank likely concepts, but it cannot resolve intent without interaction. Confidence estimates and multiple candidate captions may be more honest than forcing one answer in ambiguous cases.

The benchmark design also affects interpretation. Because collections are derived from existing datasets, they contain visually coherent examples but not the full messiness of personal photo libraries. Personal images may include motion blur, screenshots, cropped objects, repeated social contexts, private symbols, and idiosyncratic naming conventions. The method should transfer to some of these cases because it uses general embeddings and does not require supervised labels for each concept. However, real personal collections may have weaker candidate tags and stronger intent ambiguity. The reported numbers should therefore be read as controlled evidence about a modeling component, not as a deployment guarantee.

A useful property of the method is inspectability. The score for a tag can be decomposed into mean similarity, local contrastive advantage, frequency penalty, specificity prior, and variance weights. This decomposition makes it possible to understand why the model selected white car instead of car or kitchen shelf instead of room. Inspectability does not automatically make the model fair, private, or correct, but it helps debugging. When a generated caption is wrong, developers can see whether the problem came from missing candidates, poor embedding alignment, overactive specificity, or language generation. Fully end-to-end multimodal models often make this diagnosis harder.

The privacy implications are modest but worth noting. The method does not require persistent user profiles or fine-tuning on personal images. It computes collection-level statistics at inference time. In a deployed system, embeddings and candidate tags could still leak information if stored or transmitted insecurely, so privacy is not solved by avoiding fine-tuning. However, a lightweight scoring layer can be run locally or with short-lived data more easily than a full personalization training procedure. This makes it potentially suitable for workflows where users want temporary personalization without committing images to a long-term model update.

The broader methodological lesson is that small set aggregation deserves more attention in vision-language systems. Many benchmarks reward single-image recognition or single-image captioning, while many applications involve small groups: a few product shots, a few reference images, a few frames from a scene, or a few examples of a desired style. These small groups are too small for training in the usual sense but large enough to provide useful evidence. Methods that use within-set statistics, uncertainty, and candidate competition can extract that evidence without pretending that the set is a full dataset. The proposed method is one instance of this pattern.

## 9. Conclusion

This paper examined personalized image captioning as a small-collection problem rather than as a single-image generation problem. The proposed variance calibrated concept aggregation method uses multimodal similarity, within-collection dispersion, local tag competition, and constrained caption generation to select representative concepts before producing a sentence. The method is intentionally modular and does not require fine-tuning a large multimodal model on user images. In a constructed benchmark of 18,240 collections, it improved representative tag selection and reduced unsupported attributes relative to mean cosine aggregation and direct multimodal captioning baselines.

The strongest result is that collection-level variance contains useful information. When an object remains stable but context changes, a simple mean similarity score can still be pulled toward generic or incidental tags. Downweighting high-variance embedding dimensions helps the model preserve repeated foreground concepts without fully discarding context. The benefit is not uniform. Scene-centered and style-centered collections are harder because their representative meaning is distributed across objects, layout, activity, or visual treatment. This unevenness is important because it shows that calibrated aggregation is a useful component, not a complete theory of personalization.

The experiments also show that caption fluency and representative accuracy are separable. Direct multimodal captioning often writes natural sentences, but it can add plausible details that are not supported by the collection. A constrained generator conditioned on ranked concepts is sometimes less expressive, yet it produces fewer unsupported color, material, and location claims. For applications such as visual search, collection organization, or prompt construction for image generation, this caution can be preferable to richer but less grounded language. For more conversational interfaces, the balance may need to shift toward naturalness and user clarification.

The benchmark is derived from public captioning datasets and cannot fully represent private image collections. Candidate generation is a bottleneck; when the right phrase is missing, the scoring layer cannot select it. Fine-grained categories and material attributes remain difficult, especially when language priors supply plausible but visually uncertain terms. Confidence estimates help identify ambiguous cases, but they can be miscalibrated when near-duplicate images make the collection look more stable than it really is. These limitations suggest that future systems should combine calibrated scoring with better candidate proposal, diversity-aware statistics, and interactive clarification.

A small amount of structure between multimodal encoding and language generation can improve set-level captioning when personalization depends on repeated visual evidence. The method does not remove the need for stronger encoders, better multimodal generators, or user feedback. It does show that representative tag selection is a distinct technical step with measurable consequences for caption specificity and attribute support. Treating that step explicitly makes personalized image captioning easier to analyze, easier to debug, and somewhat more reliable under sparse tags and heterogeneous small image collections.

## References

[1] 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 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 2021.

[2] 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 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916, PMLR, 2021.

[3] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," in *Transactions on Machine Learning Research*, 2022.

[4] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

[5] 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.

[6] 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.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, PMLR, 2015.

[8] 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.

[9] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," in *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2017.

[10] 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 *European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[11] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for seman-

tic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[12] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[13] 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 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, PMLR, 2022.

[14] 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 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, PMLR, 2023.

[15] 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. Binkowski, 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.

[16] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[17] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[18] 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.

[19] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10578–10587, 2020.

[20] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[21] 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.

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

[23] 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.

[24] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804, PMLR, 2022.

[25] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 36479–36494, 2022.

[26] 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.

[27] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *International Conference on Learning Representations*, 2023.

[28] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

[29] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

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