Multiplicity is an Inevitable and Inherent Challenge in Multimodal Learning

Sanghyuk Chun
Works done at NAVER AI Lab

Abstract

Multimodal learning has seen remarkable progress, particularly with the emergence of large-scale pre-training across various modalities. However, most current approaches are built on the assumption of a deterministic, one-to-one alignment between modalities. This oversimplifies real-world multimodal relationships, where their nature is inherently many-to-many. This phenomenon, named multiplicity, is not a side-effect of noise or annotation error, but an inevitable outcome of semantic abstraction, representational asymmetry, and task-dependent ambiguity in multimodal tasks. This position paper argues that multiplicity is a fundamental bottleneck that manifests across all stages of the multimodal learning pipeline: from data construction to training and evaluation. This paper examines the causes and consequences of multiplicity, and highlights how multiplicity introduces training uncertainty, unreliable evaluation, and low dataset quality. This position calls for new research directions on multimodal learning: novel multiplicity-aware learning frameworks and dataset construction protocols considering multiplicity.

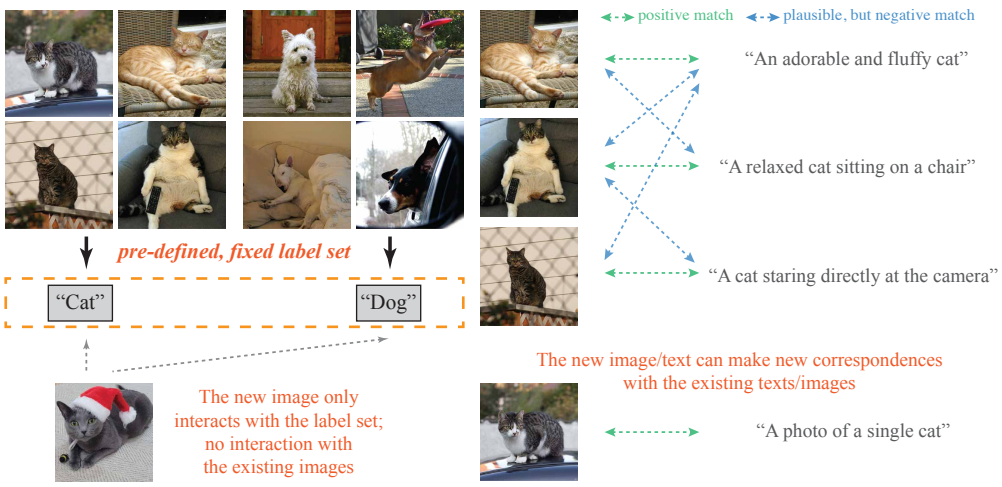
1 Introduction

Multimodal learning has emerged as a foundation in modern machine learning, showing recent breakthroughs in tasks involving vision, language, audio, action, and beyond [1, 2, 3, 4, 5, 6, 7, 8]. The advent of large-scale pre-training has significantly expanded the scope of what such systems can achieve. However, this success relies on a simplifying assumption: that mappings across modalities are *one-to-one*. Whether for contrastive pre-training or retrieval-based evaluation, each instance in one modality is assumed to correspond to exactly one correct counterpart in another, *e.g.*, one image to one caption. However, this one-to-one alignment assumption is fundamentally misaligned with the nature of real-world multimodal data. In practice, the relationship between modalities is inherently many-to-many, *e.g.*, an image can be described by multiple captions and vice versa, a property called **"multiplicity"**, the existence of multiple plausible correspondences between modalities.

This position paper argues that **multiplicity is an inevitable and inherent challenge in multimodal learning, and multimodal learning should be reframed around multiplicity.** Throughout the paper, it will be shown how multiplicity affects the entire multimodal learning pipeline, from data construction, training (*e.g.*, contrastive pre-training), to retrieval-based evaluation. Multiplicity is not a simple noise or side-effect, but a fundamental characteristic of multi-modal learning.

As shown in Fig. 1, a multimodal setting introduces a new challenge compared to unimodal settings. Supervised unimodal settings assume a pre-defined and fixed label set, where all the instances belongs to one class (sometimes multiple classes if considering multi-labeled classification [9]), namely, they assume *instance-wise annotations*. Therefore, even though we increase the dataset size, the newly added instances are irrelevant to the "ground truth" of the existing instances. On the other hand, multimodal settings usually assume *one-to-one* pairwise relationships (sometimes one-to-many if we collect multiple annotations for each instance, such as COCO Caption [10]). Unlike unimodal settings, multimodal tasks rely on *pairwise annotations*. Thus, adding a new multimodal pair increases the

Preprint. Under review.



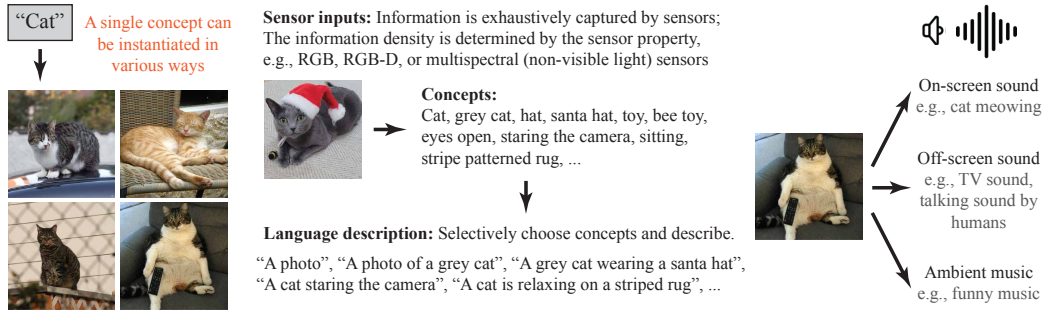
(a) Unimodal task (many-to-one) (b) Multimodal task (assuming one-to-one, but many-to-many)

Figure 1: **How unimodal task and multimodal task are different?** Unimodal tasks assume a fixed and pre-defined label set. Even though we add more instances in the dataset, the number of correspondences increases constantly, and the new instance does not affect to the existing instances. However, the correspondences in multimodal datasets, assuming one-to-one mapping, increase $O(N)$ by adding one multimodal pair.

number of annotations by $O(N)$, where N is the dataset size. Furthermore, while unimodal settings can carefully design the property of their label sets (*e.g.*, avoiding semantic overlapping between labels by considering the hierarchy from wordnet [11], or considering the popularity for balancing [9]), the nature of multimodal pairs is highly diverse, introducing multiple sources of multiplicity.

The roots of multiplicity are manifold and diverse. First, as illustrated in Fig. 1, multimodal datasets inherently introduce $\mathbf{O(N^2)}$ **pairwise annotations**, where the property of each modality instance is usually uncontrollable (*e.g.*, it is difficult to manually modify sensor inputs, such as images or sounds) – See Fig. 1 (b). Second, there exists **intra-modal variability problem**: multiple instances in one modality correspond to the same semantic concept. For example, a single concept (*e.g.*, cat) can be instantiated in diverse ways within an image modality. Many-to-many mappings naturally arise due to redundancy within each modality – See Fig. 1 and 2 (a). Third, there are **asymmetries in information density and representation mechanisms** (*e.g.*, dense image exhaustively captured by photographic sensors versus sparse linguistic descriptions with selectively chosen concepts by humans). The same modality item can be interpreted in multiple valid ways when expressed in the other modality, and it make complete and symmetric alignment infeasible. Ambiguity in what "counts" as a corresponding item leads to multiple valid alignments – See Fig. 2 (b). Finally, **the definition of correspondence depends on task objectives or context**. Different tasks demand different alignment notions, *e.g.*, for vision-language tasks, should an image be aligned to a caption describing its category, its background, its future implication, or its narrative framing? For audio-visual tasks, should a sound be aligned to on-screen actions, ambient context, or narrative tone? There is no single "true" counterpart. The set of valid correspondences varies by purpose, introducing conditional multiplicity – See Fig. 2 (c). Namely, there is no "truly corresponding unique pair" for a given instance; it depends on how we define the task. Section 2 will discuss more details of the source of multiplicity.

Multiplicity is unavoidable for multimodal tasks. Unfortunately, as the number of correspondences grows quadratically with the dataset size, making it infeasible to verify all possible matches. Namely, we should assume that we have sparse annotations for cross-modal matching; even though some correspondences are treated as positives, there might be additional plausible positives from "negative" relationships. This property introduces challenges throughout the multimodal learning pipeline. For example, in a standard training scenario with a standard multimodal dataset that assumes one-to-one correspondence, valid but unannotated positives are treated as negatives, leading to false negatives (FNs). These FNs can degrade evaluation reliability in retrieval-based benchmarks. FNs can also affect training by introducing ambiguity in instance or pairwise relationship. Considering these problems, multiplicity should be carefully considered during dataset construction, as design choices at this stage can either preserve or suppress the many-to-many nature of modality relationships.



(a) Intra-modal variability (b) Asymmetries in information (c) Task-dependent alignment

Figure 2: **How multiplicity occurs?** The source of multiplicity in multimodal datasets is diverse.

2 Multiplicity: An inevitable and inherent challenge

Definition. Let $\mathcal{R} \subseteq \mathcal{X} \times \mathcal{Y}$ denote valid cross-modal relations between two modalities $\mathcal{X}$ and $\mathcal{Y}$ (*e.g.*, vision-language [1], audio-visual [5]); Note that we assume two modalities for simplicity, but this definition can be easily extended to multiple modalities, such as vision-language-action [6, 7] and video-language-audio [12], $\mathcal{R} \subseteq \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n$. Standard practice presumes $|r : (x, y) \in \mathcal{R}| = 1$ for all x (and symmetrically for y). **Multiplicity** (or many-to-many correspondence) is a scenario that satisfies $|r : (x, y) \in \mathcal{R}| \geq 1$. Note that the number of relationships grows quadratically with the scale of the data set, $|\mathcal{R}|$ can also be quadratic with the size of the dataset in the worst case.

In unimodal settings, $\mathcal{Y}$ becomes a fixed and pre-defined label set (*e.g.*, class labels or pixel-wise mask annotations), *i.e.*, $|\mathcal{Y}|$ is constant. Furthermore, label sets for unimodal settings are usually well-defined; there is little cases when $x \in \mathcal{X}$ belongs to multiple $y \in \mathcal{Y}$. Although some studies showed that popular classification tasks (*e.g.*, ImageNet [11]) are inherently multi-labeled tasks [13, 14, 15]), the scale of "multi-label" is relatively low compared to multimodal multiplicity. For example, Yun *et al.* [15] showed that while ImageNet images may correspond to multiple valid labels, approximately five labels per image can account for most of the semantic ambiguity. Hence, adding a new instance x in the dataset does not change "ground-truths" of existing data points – See Fig. 1 (a).

Multimodal tasks define ground truth not by fixed labels, but through pairwise relationships between instances from two modalities, $\mathcal{X}$ and $\mathcal{Y}$. Each data instance is typically represented as a positive pair (x, y), *e.g.*, an image and its corresponding text description. In contrast to unimodal settings where label sets are fixed and pre-defined, the space of possible instances in each modality is open-ended and inherently ambiguous. Moreover, adding a new multimodal pair (x, y) to the dataset can induce additional implicit relationships with existing instances. For example, adding a new caption "photo" to an image-caption dataset introduces plausible matches not just with one image, but potentially with all photographic images in the dataset – See Fig. 1 (b). This combinatorial nature of cross-modal alignment implies that the number of meaningful relationships can grow quadratically with the dataset size. In practice, each instance x is often associated with multiple valid counterparts in the other modality, *i.e.*, $|y : (x, y) \in \mathcal{R}| \geq 1$, due to structural properties of multimodal data.

Property 1. Intra-modal variability. Assume a data generation process (*e.g.*, structural causal models [16]) from the underlying "concepts" to the actual data. For example, assume visual and textual instances generated from concepts "grey cat", "santa hat", and "striped rug" as shown in Fig. 2 (b). This generation process is inherently stochastic, with no uniquely determined instance. Namely, each modality realizes the concepts in various shapes, *e.g.*, images with slightly different views or backgrounds, and diverse captions describing the same situation – See Fig. 2 (a). Namely, if there exist two semantically similar multimodal pairs with overlapping concepts (x_1, y_1) and (x_2, y_2), their cross-relationships (x_1, y_2) and (x_2, y_1) also should be positive even though they are treated as negative in the dataset. This problem becomes significant when we restrict the possible objects in the datasets and the data format (*e.g.*, COCO Caption [10] is built upon COCO [17] images of 80 common objects). Chun *et al.* [18] showed that COCO Caption contains many redundant captions, which results in false negatives (FNs) in the dataset; the average number of plausible human-verified positive images/captions for each caption/image is 8.5/17.9 (originally 1/5, respectively).

Property 2. Asymmetry between modalities. Modalities differ in how they encode and express information. For example, a photograph exhaustively records visual details, while a human-written caption selectively conveys only a few salient concepts. Although the same concept may appear in both modalities in varied forms, their information density differs significantly, especially in text, which reflects human cognition rather than sensor-based input. Theories from cognition, such as dual-coding theory [19] suggest that the mind processes information along verbal and nonverbal systems. When a person writes "a grey cat wearing a Santa hat" the verbal code is followed by a private visual image that may include additional details (background, action) never lexicalized. Different annotators, therefore, generate distinct but equally valid sentences for the same scene, and a single sentence can evoke multiple mental images, immediately yielding many-to-many alignments. Even sensor inputs have different information density by the choice of the sensor. For example, visual inputs captured by RGB, RGB-D, non-visible light, video camera, and motion sensors have different information from each other; the same scene will be expressed differently by the sensors.

Property 3. Task-dependent alignment. What counts as a correct alignment often depends on the task. For example, in vision-language tasks, should a caption describe only the main object in the image [10]? Should it exhaustively describe all the local visual information [20]? Infer what happened before and what happens next [21]? In audio-visual settings in Fig. 2 (c), the notion of alignment could range from on-screen sounds (*e.g.*, cat meowing sound) [22], off-screen sounds (*e.g.*, TV sound), talking speech following lip movement [23], or ambient sounds (*e.g.*, funny music or foley effects) [24]. Namely, the definition of a "positive" pair is ambiguous, context-sensitive, and task-dependent; if a pair is positive for a specific task, the pair could not be positive for another type of task (*e.g.*, ambient sounds could be negative if we only focus on on-screen sounds).

Overall, the nature of multimodal correspondences is inherently many-to-many. In the following sections, we examine how this multiplicity impacts data collection, model training, and evaluation.

3 Multiplicity in training

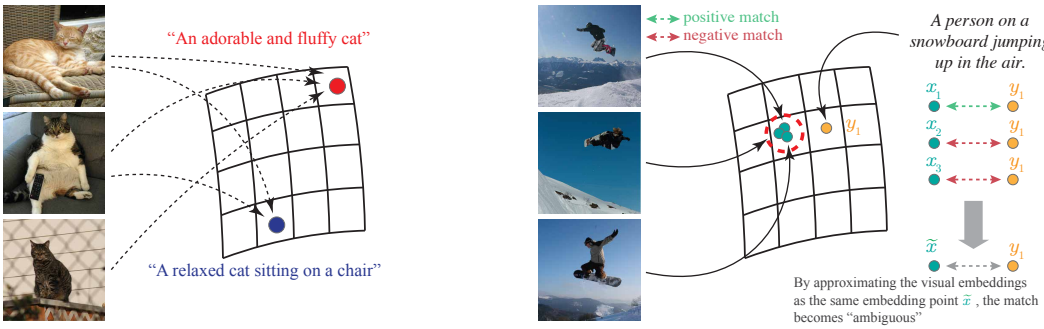
3.1 How multiplicity induces ambiguity in multimodal matching?

Mainstream multimodal architectures [25, 1, 26, 27] typically assume a one-to-one mapping, *i.e.*, each instance is encoded into a unique representation vector. However, multimodal inputs are inherently polysemous: a single instance can correspond to multiple valid interpretations or alignments, each deserving a distinct representation. If we assume an ideal dataset that annotates all plausible matches as positives, this multiplicity cannot be faithfully captured by one-to-one encodings. For example, as illustrated in Fig. 3 (a), a cat image should simultaneously match multiple captions with different meanings, which is fundamentally impossible by a one-to-one mapping. This introduces **input ambiguity**, or aleatoric uncertainty; an input, namely, an input, can be represented in various ways.

In practice, most multimodal datasets [20, 28, 29, 30, 31, 32] consist of one-to-one mapping, because a perfect dataset is infeasible due to annotation costs. However, considering that multimodal correspondences are inherently many-to-many, **false negatives** (plausible but unannotated matches) naturally emerges. While each input has only one "ground truth" (hence, the input-level ambiguity disappears), the ambiguity still exists at the level of pairwise relationships. In this case, models suffer from **matching ambiguity**: a given correspondence between an instance in one modality and another can be either positive or negative. This is another form of aleatoric uncertainty, not over the inputs themselves but over their cross-modal alignments. We examine how matching ambiguity arises.

As discussed in Section 2 *intra-modal variability*, multiple semantically similar items often exist within each modality. When we approximate such items (*e.g.*, images of the same object in different views, or captions describing the same scene with varying detail) into a single representation (by assuming that the encoder maps similar inputs into a very close and almost the same space), the resulting cross-modal matching becomes intrinsically ambiguous. Fig. 3 (b) illustrates the overview.

Formally, suppose $\{x_1, x_2, \ldots, x_K\} \subset \mathcal{X}$ are semantically equivalent inputs, approximated as a single representative $\tilde{x}$. Let $\{y_1, \ldots, y_K\} \subset \mathcal{Y}$ be their corresponding instances from another modality and the positive relationships r are $\{r : (x_i, y_i) \in \mathcal{R}\}$. Assume we randomly sample (x_i, y_j) from the mini-batch $\{(x_i, y_i) \mid i = 1 \ldots K\}$. Then, the matching label m between $\tilde{x}$ and y_j becomes a stochastic variable: $m(\tilde{x}, y_j) = m(x_i, y_j) = 1$ if $i = j$ and 0 otherwise. If we assume that y is approximated as $\tilde{y}$, the probability to have positive matching between $\tilde{x}$ and $\tilde{y}$ will be $1/K$.



(a) Multiplicity introduces "input ambiguity"

(b) Multiplicity induces "matching ambiguity"

Figure 3: **Multiplicity induces ambiguity.** (a) If we have an ideal dataset consists of the full pairwise annotations, an input should correspond to multiple instances from the other modality. The current one-to-one paradigm cannot handle this. (b) In practice, we have sparsely annotated pairwise annotations: each input only corresponds to one instance. In this case, multiplicity introduces a new uncertainty, named matching ambiguity.

Recap of current multimodal learning training algorithms. Modern multimodal learning heavily relies on training objectives that assume well-defined, one-to-one multimodal correspondences. Approaches such as triplet loss with hard negative mining [33, 34], contrastive learning [1, 27], pairwise matching [25, 26], and instruction tuning [4] all follow a similar principle: bring positive pairs closer while pushing negatives apart. They work under the assumption that each input has a single, correct counterpart in the other modality. When this assumption fails due to the input ambiguity or matching ambiguity, the model is penalized for preserving the correct semantic structure. This misalignment leads to undesirable outcomes: (1) distances between semantically compatible items become exaggerated, and (2) models may overfit to arbitrary choices among positive matches by disrupting the stability of gradient signals, especially when only one ground-truth is used in training.

Settings. Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be items from two modalities. Each mini-batch contains N instances, with N annotated positive pair $\{(x_i, y_i) \mid i = 1 \dots N\}$. Without loss of generality, We suppose that x_1 actually has $K > 1$ equally valid matches in the mini-batch: $\mathbf{y}_+ = \{y \mid y_1, \dots, y_K\}$. In this case, the total number of true positive relations is $N - K + K^2$, whereas the dataset may only annotate N of them as positive, leaving $K^2 - K$ relations unobserved and thus treated as negatives, *i.e.*, false negatives (FNs). Let $f(x)$ and $g(y)$ denote the normalized embedding by encoders f and g.

Contrastive loss. Let $p_j := \frac{\exp(f(x_1)^\top g(y_j))}{\sum_{k=1}^{N} \exp(f(x_1)^\top g(y_k))}$, the softmax probability that x_1 and y_j are matched. For x_1, the original contrastive loss (*i.e.*, only considering N positives) is defined by $\mathcal{L}_{\text{sparse}} = -\log p_i$ and its gradient w.r.t. $f(x_1)$ is $\nabla_{f(x_1)} \mathcal{L}_{\text{sparse}} = \sum_{j=1}^{N} p_j g(y_j) - g(y_1)$; this gradient becomes 0 when $p_1 = 1$, namely, pushing $f(x_1)$ and $f(y_1)$ closer while pull $f(x_1)$ away from all other $f(y_j)$. However, if we suppose that x_1 has $K > 1$ actually valid positives but unannotated $\mathbf{y}_+$. Then, the gradient pulls $f(x_1)$ away from $g(y_+)$ despite their semantic similarity.

In contrast, an ideal loss considering all positives uniformly account for all K positives: $\mathcal{L}_{\text{ideal}} = \sum_{j=1}^{K} (-\frac{1}{K} \log p_j)$. Let $p_j^* = \frac{1}{K}$ for $j = 1 \dots K$ and 0 otherwise. Then, the gradient of $\mathcal{L}_{\text{ideal}}$ w.r.t. x_1 becomes: $\nabla_{f(x_1)} \mathcal{L}_{\text{ideal}} = \sum_{j=1}^{N} p_{ij} g(y_j) - \sum_{j=1}^{N} p_j^* g(y_j)$; this gradient becomes 0 when $p_j = \frac{1}{K}$ for all $j = 1 \dots K$. More specifically, the discrepancy between the actual and ideal gradients becomes $\sum_{j=1}^{N} (p_j - p_j^*) g(y_j)$. This mismatch makes the distance between x_1 and its plausible matching y_+ larger; despite x_1 and y_+ being actually positive, there exists a gap between the two modalities, which can lead to the modality gap problem [35]. As K increases, this mismatch amplifies, leading to slower convergence [36] and greater semantic fragmentation in the learned embeddings.

Hard negative mining (HNM). HNM is a widely used technique in multimodal metric learning to accelerate training by focusing on the most challenging negatives [33, 34]. However, it is particularly vulnerable to FNs when their similarity $f(x_i)^\top g(y_+)$ is comparable to that of the true positive $f(x_i)^\top g(y_1)$. In this case, HNM aggressively pushes $f(x_i)$ away from $g(y_+)$, often more strongly than contrastive learning, resulting in a distorted embedding space that violates semantic consistency.

3.2 Current attempts and future directions

Despite its significance, the impact of multiplicity during training remains underexplored, particularly in large-scale settings such as vision-language embeddings [1, 27] or multimodal LLMs [4]. Several attempts have been made using a smooth loss [37], pseudo-label [38, 39], or mixed label [39] using mixing augmentations [40, 41], but their impacts are yet limited. While smaller-scale datasets [10] have been used to study the issue, existing approaches show limited scalability and generalizability.

One line of work treats multimodal alignments as noisy correspondence (NC) [42] (*i.e.*, considering that a specific portion of "positive" and "negative" annotations are noisy), leveraging techniques from learning with noisy labels [43]. However, this approach has shown limited success in large-scale settings; for example, Chun [39] reported that this direction shows negligible benefits over standard contrastive learning. Moreover, architectures and training objectives for NC still assumes one-to-one correspondence, limiting in representing inherent input ambiguity. Nonethelese, rethinking a multimodal task with sparsely annotated many-to-many pairwise datasets as learning with noisy labels [43] or positive-unlabeled learning [44] will be an interesting future research direction.

Another direction focuses on producing multiple embeddings, rather than single embedding for each instance [45, 46], where an instance is mapped to a set of representations to capture polysemous context, and similarity is defined via set-to-set relationships. This method assumes a fixed number of latent components per input (*e.g.*, two embeddings for each instance), each intended to capture a distinct concept. While this direction conceptually fits with both input uncertainty and matching uncertainty, it lacks flexibility when there exists more concepts than the pre-defined components and remains unproven at scale. Conceptually, mixture-of-experts (MoE) [47, 48] can be an alternative of this direction, but the link between MoE and multiplicity is still underexplored.

Probabilistic embeddings [18, 49, 50, 39, 51, 52] offer a more scalable alternative by modeling each instance as a probabilistic distribution, thereby naturally capturing uncertainty in both representation and alignment. This family of methods has been extended to large-scale VL models [52, 53], achieving performance competitive with CLIP [1]. Nonetheless, the empirical gains from probabilistic modeling remain modest in real-world applications, and their practical utility is still subject to debate.

Despite these directions, the field lacks a unified framework that systematically addresses multiplicity in multimodal training. This paper encourages rethinking multimodal training, including architecture, representation space, and training objectives, with the inherent input and matching uncertainties.

4 Multiplicity in evaluation

4.1 When and how Multiplicity makes multimodal benchmarks unreliable?

Multimodal models are often evaluated by one of the following approaches: (1) zero-shot evaluation by defining tasks via modality-specific information; (2) cross-modal retrieval, where the goal is to retrieve corresponding items across modalities (*e.g.*, image-to-text, text-to-audio); and (3) evaluation of generated outputs, such as captioning, audio synthesis, or robotic action plans. Multiplicity can arise unreliability to the benchmark under some scenarios. Specifically, cross-modal retrieval and generation evaluation are fundamentally vulnerable to multiplicity and its corresponding FN problem. Zero-shot tasks are relatively robust to this problem but we need a careful task definition.

Zero-shot evaluation defines tasks using modality-specific information (mostly based on textual description). For example, language-driven models perform zero-shot classification tasks by treating class labels as textual descriptions and performing classification via cross-modal similarity [1]. As another example, vision-language-action (VLA) models perform tasks based on text instruction sets, and evaluate the plan success rate [6]. This paradigm relaxes the pre-defined and fixed task condition by modality-specific information (mostly based on text descriptions, but not mandatory to be language – *e.g.*, task can be defined by audio, such as speech [54]). While zero-shot classification can sometimes avoid the pitfalls of multiplicity, this is largely contingent on how the label space is constructed. If class labels are distinct and mutually exclusive, the evaluation remains stable. However, in the case of taxonomic hierarchies (*e.g.*, "Cat" vs. "Russian Blue") or lexical ambiguity (*e.g.*, "laptop computer" vs. "notebook computer" in ImageNet classes [55]), the presence of multiple valid labels per instance challenges the assumption of single-label correctness [13, 14, 15]. To make zero-shot evaluation more reliable, the task should be carefully designed considering multiplicity.

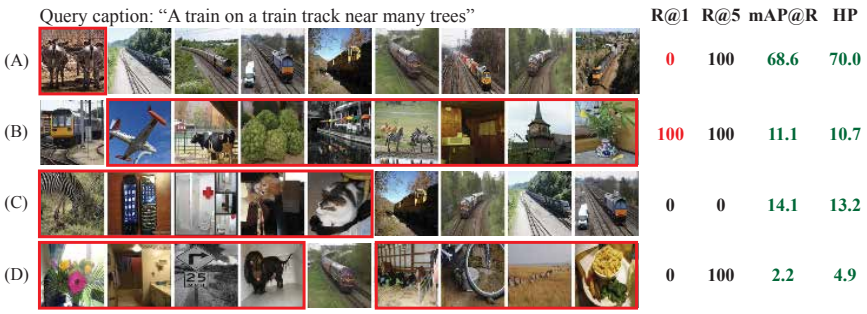


Figure 4: **Human preference vs. evaluation metrics under multiplicity.** Chun *et al.* [56] asked human annotators to compare four retrieval scenarios: (A) only top-1 is wrong, (B) only top-1 is correct, (C) top-1 to top-5 are wrong, and (D) only top-5 is correct. By comparing them in pairwise, the human preference (HP) score is computed by the BT model [64]. mAP@R [57] is highly correlated to HP, while R@Ks are often irrelevant.

In contrast, cross-modal retrieval is directly and severely impacted by multiplicity. Multiplicity inherently leads to false negatives (FNs), while most datasets assume a single correct target for each query. However, as the number of matches grows quadratically, it is infeasible to densely annotate all the possible matches between two modalities. Specifically, when a dataset is built upon limited objects (*e.g.*, 80 common objects) and a fixed format (*e.g.*, describing the main object), cross-modal retrieval results are often unreliable. For instance, the ECCV Caption benchmark [56] demonstrates that a significant portion of COCO Caption [10] treated as negatives are in fact semantically correct for human annotators ($\approx \times 4.4$ positive matches than the original dataset). Furthermore, if we consider multiple positives for each query, evaluation metric also matters in cross-modal retrieval benchmarks; the convention is Recall@K (R@K), but it is often misaligned to human perception.

Most cross-modal retrieval benchmarks assume that each query corresponds to exactly one positive target. This leads to the widespread use of R@K, which simply check whether the positive appears within the top-K retrieved items. However, previous studies [57, 56] have shown that R@K is not only less informative than ranking-based metrics such as mAP@R (where R denotes the number of positives), but can also be misleading. In particular, R@K ignores the overall ranking quality and fails to reward models that retrieve multiple semantically appropriate items, making it insensitive to models that produce coherent and diverse outputs – it makes a case when R@K is 100% but mAP@R is not 100% [57] – See Fig. 4 (B). Furthermore, as shown in Fig. 4, ranking-based metrics offer a more nuanced and human-aligned perspective [56] – mAP@R and human preference (HP) are highly correlated than R@K. Unfortunately, enlarging K cannot be a solution; Chun *et al.* [56] showed that the rankings by R@K with different Ks are highly correlated each other, while the ranking by mAP@R is less correlated to them. This indicates that the need of carefully annotated cross-modal retrieval benchmarks and more reliable evaluation metrics for retrieval benchmarks under multiplicity.

Finally, evaluating generated outputs under multiplicity introduces a different set of challenges. Generative tasks are inherently open-ended, and the space of plausible outputs is vast and diverse [58]. Traditional automatic metrics evaluate generated outputs by comparing them to a limited set of reference outputs [59], typically using surface-level measures like n-gram overlap [60, 61, 62] or latent-level comparison [63], However, this approach fails to account for the fact that many semantically appropriate generations may differ from the reference. For example, "a grey cat in the house" and "a Russian Blue playing inside" are different phrasing but equally valid; automatic metrics cannot distinguish them. In this setting, multiplicity leads to systematic underestimation of model quality, as diverse but valid outputs are treated as incorrect. This highlights a fundamental limitation of current generation-based evaluation protocols in the presence of multimodal ambiguity.

4.2 Current attempts and future directions

The most direct way to address multiplicity in evaluation is to exhaustively annotate all plausible cross-modal pairs. However, this is infeasible in practice due to the quadratic growth in the number of possible correspondences. Instead, existing work has explored two main directions.

The first is to automatically identify additional positives using side information such as attributes or semantic similarity. For instance, PCME [18] introduced densely annotated retrieval benchmarks

on CUB [65] and COCO [17] datasets with fine-grained attributes and object labels. This approach helps mitigate FNs and enables to use precision metrics, thanks to the multiple positives per query. However, it may suffer from false positives, especially when captions refer to scene elements not captured by the predefined object labels. As another example, Wray *et al.*. [66] considered semantic similarity proxies computed on captions (*e.g.*, bag-of-words or part-of-speech overlap) for a more reliable video retrieval evaluation. This highly relies on the quality of the similarity proxies.

The second direction is to manually annotate a reduced set of candidate pairs, selected via automatic methods [67, 56]. For example, ECCV Caption [56] used five different retrieval models to select up to 25 candidate matches per query. Human annotators then verified whether each candidate was a true match. This is significantly cheaper than full annotation, but still has a risk of FNs if valid matches are omitted during candidate selection. Also, the scalability of this approach is not promising.

While multiplicity has been relatively actively discussed in retrieval evaluation, its implications are even less explored in other settings. In generation-based evaluation, human judgment remains the de facto standard to handle semantic diversity, as automatic metrics often unreliable. Although human evaluation better reflects real-world diversity, the lack of scalable and reliable automatic metrics continues to slow progress. In zero-shot tasks, multiplicity can be partially addressed with ideas from unimodal tasks. Previous works [13, 14, 15] have proposed rethinking single-label benchmarks as multi-label tasks or refining label sets to reduce ambiguity. Similar strategies could be applied to zero-shot multimodal evaluation, such as revisiting prompts or category definitions in benchmarks.

Ultimately, a more faithful evaluation framework must explicitly account for the many-to-many nature of multimodal relationships, both in how relevance is defined and how performance is measured.

5 Multiplicity in dataset construction

5.1 The relationship between the degree of the multiplicity and multimodal dataset quality

Recent studies have shown that multimodal model performance is closely tied to both model and dataset scale [68]. As traditional dataset construction is labor-intensive (*e.g.*, manual captions written by human annotators [10]), recent approaches focus on collecting large-scale but noisy multimodal pairs (typically crawled from the web) and filtering them to remove low-quality examples [28, 31]. Specifically, the existing dataset construction process concentrates on "alignment", measured by a lareg-scale pre-trained model [32, 69, 70]. For example, large-scale image-text datasets, such as LAION-5B [31], discards image-text pairs whose CLIP similarity is smaller than a pre-defined threshold. This heuristic has become a rule-of-thumb for scalable multimodal dataset construction.

However, as dataset size increases, the strategy that discards or keeps pairs with CLIP similarity may not be enough. As shown in Fig. 1 (b), adding even a single multimodal pair can influence the multiplicity structure of the entire dataset. For example, underspecified instances (*e.g.*, "photo" or "a person is standing") tend to align with a large number of items (*e.g.*, all general photos or human figures), amplifying multiplicity – leading to input- and matching-ambiguity as discussed in Section 3. Several studies [71, 72, 73, 74] attempted to avoid this challenge by training multimodal models solely with unimodal datasets (*e.g.*, text-only training), but this cannot be a fundamental solution.

Whether a dataset preserves or suppresses this multiplicity depends on design choices of multimodal pair collection and task definition: retaining only specific, narrowly defined examples may reduce multiplicity, but this is often infeasible since alignment depends on task-specific semantics. Bringing VL tasks as an example, we can reduce the potential matches of the given image by describing all the localized details in the image [20]; this may reduce the multiplicity, but this process is expensive and sometimes not helpful to general-purpose tasks, such as zero-shot classification. On the other hand, if we focus on the salient objects in the image [10], the captioning process becomes cheaper, but the possible matching images per each caption will dramatically increase [56].

5.2 Current attempts and future directions

Despite its importance, multiplicity has received limited attention in the context of dataset construction. While multiplicity-aware modeling and architecture design may eventually need to account multiplicity, minimizing unnecessary multiplicity at the dataset level remains a critical and cost-effective strategy, especially in the current paradigm where scaling-law still holds [1, 2, 68].

Multiplicity should be considered even before data collection, *i.e.*, starting from task definition. Cross-modal alignment is inherently task-dependent. Previous works [75, 76] showed that collecting task-relevant instances improves multimodal training. Without clear criteria for valid matches, datasets may introduce unintended multiplicity, causing downstream instability.

In addition, a careful multimodal pair collection process will be helpful to reduce the level of multiplicity. For example, filtering strategies should go beyond coarse alignment scores (*e.g.*, CLIP similarity) and explicitly target instances that amplify multiplicity (*e.g.*, underspecified inputs). One possible direction is a filtering based on specificity, such as HYPE [77]. By selecting more specific instances (defined by the embedding property), HYPE leads to higher-quality datasets and improved downstream performance. This supports the broader hypothesis that reducing multiplicity at the data level yields tangible benefits throughout the multimodal pipeline.

6 Discussions

Alternative views. The existence of multiplicity in multimodal learning is no doubt. However, as an alternative viewpoint, practitioners can argue that multiplicity is a neglectable issue and simply scaling up the model and dataset under the one-to-one assumption can achieve high-performing multimodal models. This might be true in the current status; as shown by recent studies [1, 2, 68, 32], scaling up the models with noisy multimodal pairs looks promising in terms of building strong multimodal models. However, as observed in the ImageNet classification task, the fundamental flaw in dataset or task becomes significant when models become very strong. For example, Beyer *et al.* [13] showed that ImageNet accuracy can be flawed when it goes beyond 90% due to the wrong labels. We now have strong multimodal models, but they are not yet sufficiently strong enough as much well-defined as supervised classifiers. It may not be the correct timing to consider multiplicity for practitioners. However, this paper argues that despite the multiplicity not being a critical factor of the performance as of now, it eventually should become a fundamental bottleneck to achieve a superhuman-level multimodal model because multiplicity is inevitable, and the impact of the multiplicity also exists in the entire multimodal learning pipeline. Furthermore, the current one-to-one paradigm fundamentally cannot handle multiplicity; we need a new paradigm for representing multiple different ideas of the given instance, although it would not be sufficiently effective in benchmark evaluation.

What will be the future direction? This paper has shown that multiplicity is a fundamental and unavoidable challenge across the entire multimodal learning pipeline. This paper argues that future multimodal learning research should be reframed around multiplicity. It calls for **a novel modeling beyond one-to-one mapping** (*e.g.*, one-to-many mapping [45] or stochastic embeddings [18]). As of now, stochastic modeling has shown a promising scalability while achieving comparable performance with large-scale foundation models [52]. Another possible direction is a conditional modeling that takes additional conditions to specify the given instance, *e.g.*, transforming an embedding with the given text condition [78], selecting a specific local area in the image [79] or lexically specifying the characteristic of the corresponding audio from the video [12]. These approaches reduce the ambiguity of the input by specifying what instance should be matched to the given instance by additional contexts (*e.g.*, pixel-level mask or text condition). Similarly, compositionality-aware modeling [80] will be an interesting direction, which models an input as a composition of its parts (or underlying concepts). By tackling compositionally, we can handle one of the major sources of multiplicity.

Multiplicity also introduces another important challenge, **how to construct datasets with multiplicity in mind**. It involves both evaluation and training datasets; while controlling multiplicity in datasets is the biggest task for both, evaluation datasets are more focused on the hidden pairwise annotations due to the multiplicity, and training datasets aims to reduce the multiplicity in the dataset. One promising direction is to design scalable dataset construction protocols that explicitly minimize structural multiplicity. This will be an important goal for both reliable evaluation and stable training. In the long term, we may also need an iterative development cycle for multimodal systems that includes dataset collection, filtering, modeling, and evaluation (all under multiplicity-aware framework). As shown in the unimodal dataset construction [81], such iteration will significantly improve dataset quality and system robustness over time, despite its inherent complexity.

Multiplicity is not just a minor issue, but it is a core part of how real-world multimodal data works. To make progress, we need to build datasets, models, and evaluations that take it into account. Only then can we develop multimodal systems that truly understand the richness of the world.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 1, 3, 4, 5, 6, 8, 9

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR, 2021. 1, 8, 9

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023. 1, 5, 6

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*, pages 1–5. IEEE, 2023. 1, 3

[6] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1, 3, 6

[7] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*, 2023. 1, 3

[8] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1

[9] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020. 1, 2

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 3, 4, 6, 7, 8

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 3

[12] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *AAAI Conference on Artificial Intelligence*, 2025. 3, 9

[13] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020. 3, 6, 8, 9

[14] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, pages 8634–8644. PMLR, 2020. 3, 6, 8

[15] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 6, 8

[16] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000. 3

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 3, 8

[18] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 6, 7, 9

[19] Allan Paivio. *Mental representations: A dual coding approach*. Oxford university press, 1990. 4

[20] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision (ECCV)*, 2020. 4, 8

[21] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[22] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725. IEEE, 2020. 4

[23] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 4

[24] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413, 2016. 4

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019. 4, 5

[26] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*, 2021. 4, 5

[27] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 4, 5, 6

[28] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568, 2021. 4, 8

[29] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Dataset and Benchmark (NeurIPS D&B)*, 2021. 4

[30] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022. 4, 8

[32] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 4, 8, 9

[33] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*, 2018. 5

[34] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15789–15798, 2021. 5

[35] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17612–17625, 2022. 5

[36] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2785–2795, 2022. 5

[37] Jaeseok Byun, Dohoon Kim, and Taesup Moon. Mafa: Managing false negatives for vision-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27314–27324, 2024. 6

[38] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations (ICLR)*, 2022. 6

[39] Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*, 2024. 6

[40] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 6

[41] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 6

[42] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, hua wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6

[43] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022. 6

[44] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020. 6

[45] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1979–1988, 2019. 6, 9

[46] Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23422–23431, 2023. 6

[47] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 6

[48] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. 6

[49] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *International Conference on Computer Vision (ICCV)*, pages 1899–1910, 2023. 6

[50] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:24564–24585, 2023. 6

[51] Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. Post-hoc probabilistic vision-language models. *arXiv preprint arXiv:2412.06014*, 2024. 6

[52] Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2025. 6, 9

[53] Sanghyuk Chun and Sangdoo Yun. LongProLIP: A probabilistic vision-language model with long context text. In *ICLR Workshop on Quantify Uncertainty and Hallucination in Foundation Models*, 2025. 6

[54] Anonymous authors. Seeing what you say: Expressive image generation from speech. *Under review*, 2025. 6

[55] Nikita Kisel, Illia Volkov, Kateřina Hanzelková, Klara Janouskova, and Jiri Matas. Flaws of imagenet, computer vision's favorite dataset. In *ICLR Blogposts 2025*, 2025. https://d2jud02ci9yv69.cloudfront.net/2025-04-28-imagenet-flaws-135/blog/imagenet-flaws/. 6

[56] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang Chang, and Seong Joon Oh. ECCV Caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO. In *European Conference on Computer Vision (ECCV)*, 2022. 7, 8

[57] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020. 7

[58] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:69981–70011, 2023. 7

[59] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 7

[60] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*, pages 311–318, 2002. 7

[61] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7

[62] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7

[64] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 7

[65] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 8

[66] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3650–3660, 2021. 8

[67] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021. 8

[68] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023. 8, 9

[69] Pratyush Maini, Sachin Goyal, Zachary C Lipton, J Zico Kolter, and Aditi Raghunathan. T-mars: Improving visual representations by circumventing text feature learning. In *International Conference on Learning Representations (ICLR)*, 2024. 8

[70] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations (ICLR)*, 2024. 8

[71] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. 8

[72] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! learning visual tasks using only language supervision. In *International Conference on Computer Vision (ICCV)*, pages 2672–2683, 2023. 8

[73] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. In *International Conference on Learning Representations (ICLR)*, 2023. 8

[74] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[75] Haichao Yu, Yu Tian, Sateesh Kumar, Linjie Yang, and Heng Wang. The devil is in the details: A deep dive into the rabbit hole of data filtering. *arXiv preprint arXiv:2309.15954*, 2023. 9

[76] Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2024. 9

[77] Wonjae Kim, Sanghyuk Chun, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. HYPE: Hyperbolic entailment filtering for underspecified images and texts. In *European Conference on Computer Vision (ECCV)*, 2024. 9

[78] Geonmo Gu, Sanghyuk Chun, HeeJae Jun, Yoohoon Kang, Wonjae Kim, and Sangdoo Yun. CompoDiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research (TMLR)*, 2024. 9

[79] Jungbeom Lee, Sanghyuk Chun, and Sangdoo Yun. Toward interactive regional understanding in vision-large language models. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. 9

[80] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921, 2023. 9

[81] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11700–11709, 2019. 9