Is a Picture Worth a Thousand Words?

CS 6795: Cognitive Science

Adam Davis Cuculich
School of Computer Science
Georgia Institute of Technology
acuculich3@gatech.edu

***Abstract*—Humans represent knowledge through multiple modalities, notably images and words. In this paper, we investigate how each modality encodes underlying "concepts" and whether combining them can compensate for one modality's degradation. We build on the notion that words map to conceptual categories (e.g., "chair," "dog") as a near-isomorphism, whereas images are many-to-one homomorphisms, with numerous visual instances mapping to the same label. By implementing a multimodal classification task using image embeddings and text embeddings, we ask: does high-information text salvage performance when image quality is severely degraded, and vice versa?**

Using a large-scale dataset of dog and cat images, we apply systematic pixel dropout to degrade images and compare classification accuracy with high- or low-detail text descriptions. We fuse text and image embeddings via weighted summation and evaluate how overall accuracy shifts as the weight on each modality varies. Results confirm that high-detail text can "rescue" heavily degraded images, improving classification beyond image-only baselines. Even low-detail text provides some benefit in such cases, though to a lesser extent. Conversely, low-detail text sometimes reduces accuracy when paired with pristine images, reflecting how suboptimal language can dilute valuable visual signals. Our findings provide insight into how each modality contributes to conceptual mapping, offering a clearer understanding of multimodal cognition.

Keywords: Multimodal cognition, homomorphism, isomorphism, image degradation, semantic space, CLIP embeddings, cross-modal compensation, linguistic representation, visual representation, conceptual mapping

I. Introduction

Cognition involves multiple representational modalities—distinct formats or "languages of thought" through which we process information—most notably visual imagery and linguistic descriptions. A fundamental question in cognitive science concerns how these different modalities encode and access conceptual knowledge—and, crucially, how they might interact to compensate for limitations in one another. While traditional views have often treated visual and linguistic processing as largely separate cognitive domains, emerging evidence from studies on multimodal integration suggests far deeper interactions between these representational systems than previously appreciated.

We propose a set-theoretic framing in which words act as symbolic mappings to "equivalence classes," or concepts, that capture shared properties among concrete instances. For example, the label "even numbers" unites integers sharing a specific divisibility property, just as "chair" denotes a set of objects that fulfill a certain function (seating) and share similar physical traits. However, real-world objects often defy strict set boundaries, raising questions about how best to characterize conceptual categories. The concept "chair," for instance, encompasses diverse physical forms that share functional similarities rather than precise physical properties, creating fuzzy category boundaries. Modern vector-based representations in machine learning systems mirror this flexibility, representing conceptual gradations through distance metrics in high-dimensional semantic space rather than through rigid category membership rules. This approach not only powers computational systems but also mirrors how humans perceive graded category membership, where items can be more or less typical examples of a category. This is precisely how CLIP embeddings function, encoding both linguistic and visual information as vectors within a shared embedding space where semantic similarities are captured by vector proximity.

We further suggest that words function as near-isomorphic mappings to these conceptual classes (one word per concept), whereas images, with their multitude of visual variants, serve as many-to-one homomorphisms—multiple images can map to the same concept. This creates a transitive relationship (images $\rightarrow$ words $\rightarrow$ concepts) that characterizes how concrete visual instances relate to abstract conceptual categories. For example, the word "dog" maps directly to the concept DOG, but innumerable photographs of different dogs—varying in breed, posture, lighting, and context—all map to this same concept. Critically, we posit that both modalities converge upon a single underlying "semantic space" despite these different mapping relationships. This view underpins recent multimodal learning models such as CLIP [5], which embed images and text into a unified vector representation. If images and words truly access the same conceptual domain, then in principle, one modality could compensate for degraded signals in the other. In other words, if an image is highly corrupted, a sufficiently rich textual label could help "rescue" classification accuracy—and vice versa.

This compensation mechanism reflects a deeper insight about human cognition: semantics—the meaningful content of our thoughts—can take different representational forms while preserving core conceptual structure. Rather than seeing semantics as necessarily downstream from concrete instances, we propose that semantic content exists somewhat independently of its representation and can manifest in different

forms—whether visual or linguistic. The key question becomes: under what conditions can one representational form effectively supplement or replace another while maintaining access to the same conceptual knowledge?

In this paper, we test this hypothesis by systematically degrading images and text, then combining the resulting embeddings to evaluate classification performance. We manipulate image quality through controlled pixel dropout at varying levels (25%, 50%, 75%, and 90%) and contrast minimal textual descriptions with detailed ones. By varying the weighting parameter (α) that determines the relative contribution of each modality to the final representation, where $\alpha = 1$ indicates full reliance on the image and $\alpha = 0$ represents a purely textual representation, we can precisely measure how much a strong signal in one modality compensates for a weak signal in the other. Although the machine-based representations differ from those in human cognition, we propose that these embeddings form a useful analogy: if one modality's degradation is offset by stronger input from the other, it may reveal a more general mechanism of how humans combine partial or noisy signals across modalities.

By examining whether each modality can "pick up the slack," we contribute to understanding how linguistic and visual information might be integrated in the human mind—where either words or images, when sufficiently robust, can compensate for limitations in the other channel. Through this parallel, our work aims to shed light on broader principles of multimodal cognition, offering empirical evidence for how different representational systems might interact to maintain conceptual stability despite noisy or incomplete sensory inputs. Our experimental approach offers a novel means of addressing the age-old question: "Is a picture worth a thousand words?" The answer, we suggest, depends critically on the quality and information content of both the picture and the words—and on how these different forms of representation might complement one another in mapping to the same underlying concepts. Our findings have implications not only for cognitive theory but also for the design of adaptive multimodal systems that can dynamically adjust their reliance on different information channels based on signal quality.

II. Model/Tool Design

A. Cognitive Motivation and Experimental Rationale

To investigate whether a strong linguistic modality could offset severe visual degradation (and vice versa), we adopted OpenAI's CLIP model [5] for its demonstrated ability to align text and image embeddings within a single semantic space. This technological choice reflects our overarching cognitive-science motivation: both words and images can encode conceptual content, albeit through different forms of representation. By systematically degrading one modality while preserving the other, we operationalized the idea that partial or noisy sensory input might still map to a shared conceptual domain, allowing richer information in one channel to "rescue" classification performance in the other. These principles build upon the notion that distinct modalities converge on overlapping mental models or "concepts". In practice, we generated textual descriptions of varying detail (rich vs. minimal) and produced images with increasingly severe pixel dropout, then embedded them using CLIP. This approach directly tested whether linguistic detail compensates for corrupted visual cues, and conversely, whether a clean visual embedding can salvage minimal textual information. Together, these technical and conceptual foundations form the core of our methodology for exploring multimodal "rescue" in a shared embedding framework.

B. Experimental Setup: Constructing Representations

1) Data Selection and Visual Representations: We began with 25,000 labeled images (12,500 cats and 12,500 dogs) [4]. To keep text-generation costs manageable, we randomly selected 500 cat and 500 dog images from this pool. Each chosen image was retained in its *high-information* (original) form. Additionally, we produced *low-information* variants by applying pixel dropout at rates of 25%, 50%, 75%, and 90%. Pixel dropout offered a straightforward, controllable means of simulating reduced visual quality. These specific dropout levels provided a spectrum of degradation, enabling us to investigate how incremental loss of visual detail might influence classification performance.

2) Text Description Generation: For each high-information image, we generated two textual descriptions using ChatGPT:

- **High-information text** (up to 45 words): A richly detailed account of the animal's physical attributes (fur color, pattern, posture, etc.), without explicitly naming the species. The 45-word cap also ensured compliance with CLIP's token-limit constraints.
- **Low-information text** (up to 10 words): A minimal, vague depiction, typically noting only size and primary fur color.

We adopted a strict prompt structure for ChatGPT (see [2] for exact instructions) to avoid species labels and maintain consistent detail levels. Generating both descriptions from the full-quality image ensured textual richness was not artificially limited by image degradation. By pairing detailed or sparse language with images of varying quality, we created distinct multimodal *representations* for subsequent analysis.

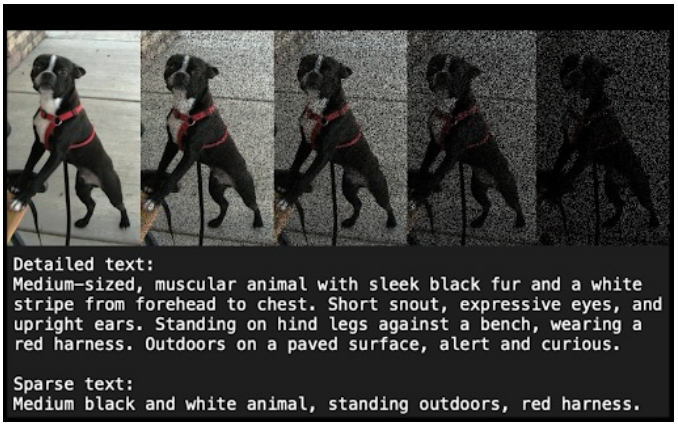

Fig. 1: Progressive pixel dropout on a dog image (from left to right) with paired detailed and sparse text captions used in the multimodal degradation experiment.

3) Embedding Generation with CLIP: We used OpenAI's CLIP model [5] to embed both images and text into a shared semantic space. This allowed us to represent visual and linguistic information using the same dimensional format, enabling direct comparison and combination. By aligning both modalities in this way, we could test whether one could compensate when the other was degraded.

4) Constructing Multimodal Embeddings: After obtaining each image embedding $\mathbf{v}_{\text{image}}$ (intact or degraded) and each text embedding $\mathbf{v}_{\text{text}}$ (high- or low-information), we formed combined vectors by introducing a weighting parameter:

$$\mathbf{v}_{\text{combined}} = \alpha \cdot \mathbf{v}_{\text{image}} + (1 - \alpha) \cdot \mathbf{v}_{\text{text}}, \quad (1)$$
$$\text{where } \alpha \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1.0\}.$$

An α value of 1.0 indicates full reliance on the image (ignoring text), whereas $\alpha = 0$ means the combined embedding depends solely on the text. For each α value, we evaluated four primary pairings:

1) High-information image + high-information text
2) High-information image + low-information text
3) Low-information image + high-information text
4) Low-information image + low-information text

Varying both α and the level of detail in each modality enabled us to systematically test our hypothesis that a more informative modality could *"rescue"* classification when the other modality was degraded.

5) Classification and Baseline Setup: For each combined embedding, we performed a binary classification (cat vs. dog). To ensure a robust evaluation:

- **Cross-validation:** We employed k-fold cross-validation (typically $k = 5$), training on four folds and testing on the remaining fold.
- **Multiple Classifiers:** We tested both logistic regression and a support vector machine (SVM) to check whether our findings depended heavily on the classifier choice.
- **Controlled Random Seed:** A fixed random seed was used for all runs to maintain consistency across experimental variations.

Full experimental details and parameter settings (e.g., train/test splits, classifier hyperparameters) are provided in Section II-C and in the project repository [1].

C. Experimental Methodology

*1) **Experiment 1: Varying Pixel Dropout and** α:* The objective is to examine how different pixel-dropout rates and weighting parameters (α) jointly affect classification performance. This allowed us to observe whether high-quality text could compensate for increasingly degraded images, and vice versa. Four dropout levels ($25\%, 50\%, 75\%, 90\%$) were applied to create low-information images of varying severity. Each degraded image was then combined with its textual counterpart at one of five α values ($0, 0.25, 0.5, 0.75, 1$) in each of the four pairings (e.g., high-information image + high-information text). A logistic regression classifier was trained and evaluated on each combination, following the classification setup described above. By systematically varying α and dropout, we aimed to pinpoint the conditions under which text "rescues" heavily corrupted images (and vice versa).

*2) **Experiment 2: Varying Classifiers**:* The objective is to check whether our findings depended on the choice of classifier, we repeated a subset of conditions using both logistic regression and SVM. We focused on a moderate dropout level (50%) and the same range of α values. For each modality pairing, we trained and tested both classifiers using identical data splits. If both classifiers yielded similar trends, we could conclude that the observed "rescue" effect is robust and not tied to a specific classification approach.

D. Rationale and Replicability

By combining systematically degraded images and purposefully constrained text, this framework makes it straightforward to compare how each modality influences classification under varying conditions. Pixel dropout provides a transparent dial for visual degradation, while strict word-count limits create distinct levels of linguistic detail. All components—pixel dropout scripts, ChatGPT prompts [2], and CLIP-based embedding—are readily accessible and reproducible with open-source tools. This ensures that other researchers can replicate our methodology or adapt it for larger datasets and different semantic categories.

III. Results

Experiment 1 Results

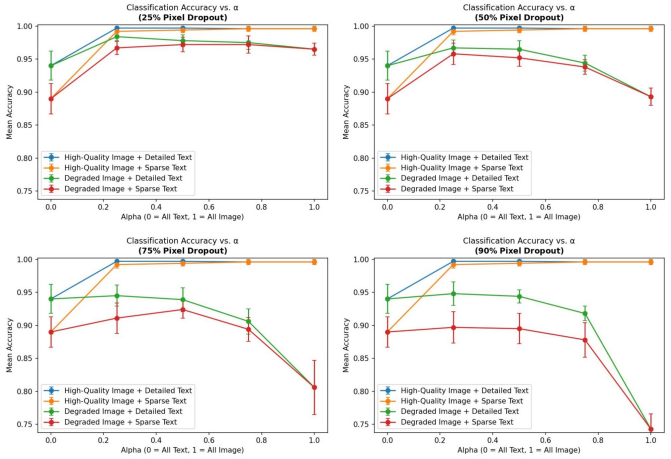

Fig. 2: Classification accuracy at different alpha values (0 = all text, 1 = all image) under four pixel dropout levels (25%, 50%, 75%, 90%).

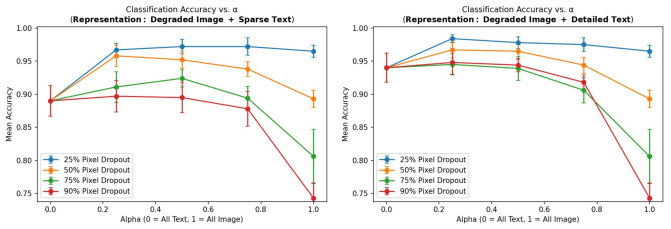

Fig. 3: Classification accuracy at different alpha values for two representation types (Degraded Image + Sparse Text vs. Degraded Image + Detailed Text), each tested under four pixel dropout levels (25%, 50%, 75%, 90%).

In the first set of graphs, classification accuracy is plotted as a function of the mixing parameter α (where $\alpha = 0$ corresponds to pure text and $\alpha = 1$ corresponds to pure image). Notice that when $\alpha = 0$, all plots start at the same level—this reflects the fact that the textual descriptions (regardless of their detail level) are standardized across all conditions. Likewise, at $\alpha = 1$, the curves converge to the same point in cases where the image is of high quality (i.e., no pixel dropout), establishing a common baseline. As the level of pixel dropout increases, the image-only (low-quality image) classification accuracy declines markedly. Yet, in every dropout condition, adding detailed text consistently boosts accuracy more than adding sparse text, demonstrating that richer linguistic information helps compensate for degraded visual input.

In comparing across representations, the data further reveal that high-information text not only yields a higher baseline (when used alone, $\alpha = 0$) but also robustly rescues performance in multimodal combinations. Specifically, adding detailed text to images—whether high-quality or degraded—results in classification outcomes that are at least as good as, if not superior to, using the image alone. Conversely, incorporating low-information text into high-quality images tends to worsen performance as text contribution increases, underscoring that insufficient textual detail can dilute valuable visual cues. Moreover, with severely degraded images (e.g., 90% pixel dropout), there is an optimal balance: increasing text weight initially improves accuracy up to about $\alpha = 0.25$, beyond which further reliance on text becomes counterproductive. Overall, these trends highlight that as image quality deteriorates, the benefit—and the disparity—of detailed versus sparse textual augmentation becomes increasingly pronounced.

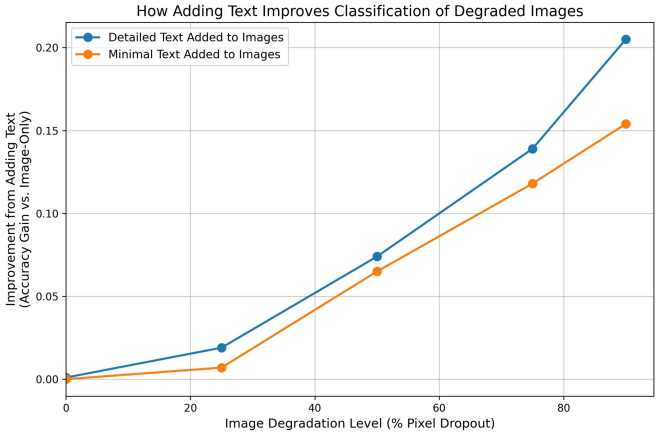

Fig. 4: Improvement in classification accuracy when combining images with detailed versus minimal text descriptions across varying levels of image degradation (pixel dropout rates from 0% to 90%). Higher image degradation levels benefit more significantly from the inclusion of detailed textual information, indicating that detailed text provides stronger compensation as visual quality deteriorates.

This graph shows how much classification accuracy improves when text is added to images at various levels of pixel dropout. The x-axis indicates the degree of image degradation (from 0% to 90%), while the y-axis plots how much adding text boosts accuracy compared to using images alone. As the pixel dropout increases, the benefit from including text becomes more pronounced, reflecting how linguistic information can "fill in" lost visual details. Notably, detailed text (blue line) provides a stronger boost than minimal text (orange line) because richer descriptions convey more semantic cues that help compensate for degraded imagery. This demonstrates an important point in multimodal cognition: the more an image is corrupted, the more valuable comprehensive language input becomes, offering a "rescue" effect that preserves overall classification performance.

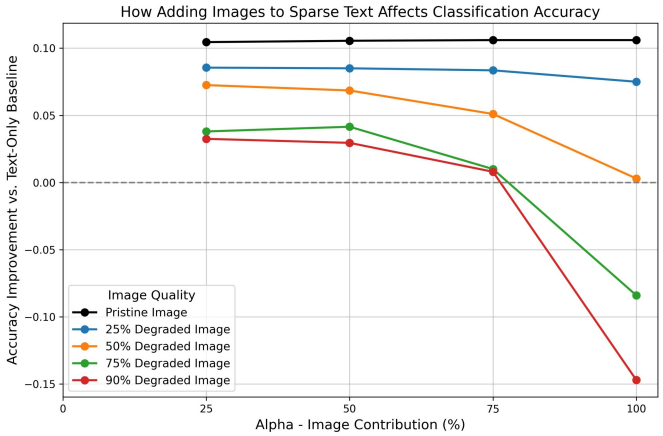

Fig. 5: Impact of image quality on classification accuracy when images are combined with sparse (low-detail) text descriptions. Classification accuracy improvement is measured relative to a low-detail, text-only baseline across varying levels of image degradation. Pristine images significantly enhance accuracy at all levels of image contribution, whereas heavily degraded images (75% and 90% pixel dropout) reduce accuracy when image contribution exceeds approximately 50%.

This figure illustrates how varying the proportion of image contribution (α) affects accuracy relative to a text-only baseline when the text is sparse and images differ in quality. The y-axis shows the gain (or drop) in accuracy over using only low-detail text, and each line corresponds to a different level of pixel dropout. The black line (pristine images) consistently remains above zero, indicating that high-quality visuals always bolster sparse text. In contrast, heavily degraded images (green and red lines) degrade performance once the model relies too heavily on them (i.e., $\alpha > 0.5$), ultimately falling below the text-only baseline. This pattern highlights that while even a low-quality image can offer a slight benefit at moderate weighting, overly relying on severely corrupted visuals can overshadow and diminish the gains provided by text.

Experiment 2 Results

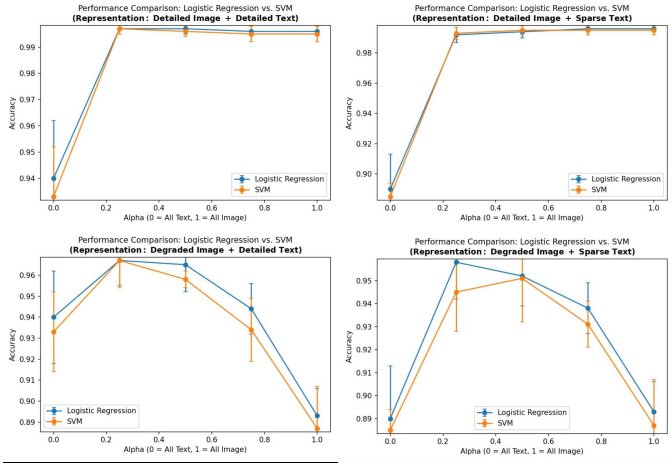

Fig. 6: Comparison of logistic regression vs. SVM classification accuracy across four image–text representation strategies as α varies ($0 =$ all text, $1 =$ all image). Each subplot corresponds to a distinct representation (e.g., "Detailed Image + Detailed Text"), with error bars indicating the standard deviation of accuracy.

In this second experiment, we compare how logistic regression and SVM perform across four distinct image–text representation strategies, as shown in the subplots. Each subplot tracks accuracy as the mixing parameter α shifts from text-only ($\alpha = 0$) to image-only ($\alpha = 1$). Overall, logistic regression tends to outperform SVM in most conditions, but both classifiers follow similar trends: when either the image or text is high quality, accuracy remains high across α, whereas if both modalities are limited, performance dips for intermediate α values. These parallels between logistic regression and SVM confirm that our observed "rescue" effect—where detailed text helps weak images and pristine images compensate for sparse text—is robust to the choice of classifier, though logistic regression generally has the slight advantage.

IV. Discussion and Analysis

Our experimental findings offer several insights into how—and why—one modality can "rescue" classification when the other is compromised. First, Experiment 1 shows that detailed text descriptions substantially boost accuracy for images suffering heavy pixel dropout, consistent with the idea that linguistic detail helps restore or clarify conceptual mappings otherwise lost to poor visual fidelity. Notably, as pixel dropout increases, the gap between adding high-detail versus sparse text widens, underscoring that not all language inputs are equally beneficial. This highlights an important mechanism: richer text not only complements deteriorating images but can surpass image-only baselines by supplying semantic clarity otherwise missing in the corrupted visual channel. Conversely, low-information text can hamper performance when paired with pristine images, suggesting that insufficiently descriptive language may overshadow a more reliable modality—

a phenomenon seen in real-world scenarios where minimal or misleading verbal cues sow confusion rather than enhance understanding.

Experiment 2 reinforces this view by showing that both logistic regression and SVM manifest the same fundamental rescue patterns, with logistic regression often yielding slightly higher accuracies. The consistency across distinct learning algorithms indicates that our "rescue" effect is not just a quirk of classifier choice but reflects broader principles of multimodal data integration. Even small performance gains in logistic regression could be pivotal in fields such as medical imaging or safety-critical surveillance, where incremental improvements have substantial impact. For instance, a telemedicine system dealing with noisy scans might weight textual reports more heavily when visual fidelity degrades, while an automated security platform could rely on textual object descriptors in dim or partially occluded settings, provided those descriptors convey genuinely useful information.

From a cognitive-science standpoint, these results support the notion that images and words can map onto a shared conceptual space in which each modality can reinforce or dilute the other's signal. When language is robust, it fills gaps left by partial or corrupted visuals, much like a detailed verbal explanation can clarify ambiguous perceptual input. Conversely, if text is too sparse, a high-fidelity image may dominate, unless minimal language introduces additional uncertainty. Our findings specifically highlight that, for severely degraded images, an α value around 0.25 is often optimal—indicating that leaning moderately toward text, yet still preserving some image signal, yields the best balance. This mirrors everyday human behavior, where we instinctively favor the most reliable information channel but still glean value from complementary signals.

Finally, these insights have real-world and theoretical implications for multimodal AI design. Technologically, adapting each modality's weight based on signal quality could greatly enhance robustness, particularly in domains prone to noise or partial data loss. Conceptually, seeing that a single embedding framework (CLIP) can replicate the capacity for cross-modal compensation illustrates how richer data in one modality can salvage concept recognition when another modality fails. Future multimodal systems might incorporate dynamic weighting mechanisms at training or inference time, ensuring they allocate attention to whichever channel remains most trustworthy under changing conditions. In this sense, our work highlights the value of flexible cross-modal methods for achieving resilient perception, echoing how human cognition integrates diverse inputs to maintain reliable concept formation in a sometimes noisy world.

V. Conclusion

Our experiments show that the synergy between images and text hinges critically on the fidelity of each modality. When images are degraded, detailed text provides a "rescue" that preserves or even surpasses accuracy relative to image-only baselines; however, sparse text can undermine performance if the accompanying images are pristine. These effects were observed across different classifiers, indicating they likely reflect broader principles of multimodal data integration rather than model-specific artifacts.

From a theoretical standpoint, our results support the view that linguistic and visual information converge on a shared conceptual space, allowing richer information in one modality to compensate for weaknesses in the other. Practically, this implies that future multimodal AI systems should dynamically assess and weight each modality's contribution based on current signal quality, thus ensuring robust classification under real-world conditions. By extending these techniques to larger, more diverse datasets and additional modalities (e.g., audio), researchers can further explore and refine the principles governing cross-modal compensation, ultimately leading to more resilient and human-like approaches to multimodal perception.

VI. Limitations and Future Research

Our study highlights several areas that warrant caution and invite future exploration. First, we focused on a binary classification task involving cats and dogs; it remains an open question whether the observed "rescue" effects generalize to more diverse datasets, semantic categories, or higher-level visual concepts. Testing our approach on more complex classification tasks involving hundreds of categories or abstract concepts would provide stronger evidence for the generalizability of our findings. Second, while CLIP is well-suited for aligning image and text embeddings, relying on a single embedding model may limit generalizability; alternative architectures or more recent multimodal frameworks could produce different interaction dynamics between modalities. Comparative analyses across multiple embedding frameworks would help identify which aspects of our findings reflect general principles of multimodal integration versus model-specific behaviors.

Third, all textual descriptions were generated directly from the original images, meaning the text was downstream from a richer, concrete visual source. This conversion from high-fidelity visual input to abstract language may introduce semantic compression or loss—potentially explaining why high-information images often outperform even the most detailed descriptions. Future work might explore independently sourced text and images that reference the same concepts but were not derived from one another. Fourth, we restricted degradation methods to pixel dropout for images and word-count limits for text; other techniques such as image blurring, random word insertion, or syntactic corruption could offer complementary perspectives on robustness and recoverability. Additionally, our controlled degradation might not reflect real-world conditions where image corruption follows specific patterns (e.g., motion blur, occlusion) rather than random pixel dropout.

Fifth, the linear weighting parameter (α) used to combine modalities may oversimplify how different information channels should interact. More sophisticated fusion techniques—such as attention mechanisms that dynamically emphasize different aspects of each modality—could potentially

yield even stronger rescue effects and better mirror how humans selectively attend to the most reliable parts of a signal. Sixth, our study lacks direct comparisons with human performance on similar tasks. Future research could involve parallel experiments with human participants to determine whether the optimal weighting parameters we identified computationally align with human multimodal integration strategies when faced with degraded inputs.

Finally, although our work draws on analogies to human cognition, these findings ultimately reflect machine learning behavior. The structure of CLIP's embedding space—and its capacity for cross-modal compensation—may not correspond to how humans internally integrate language and vision. Individual differences in perceptual abilities, linguistic competence, and cognitive strategies further complicate any direct mapping between our computational findings and human cognition. Cross-cultural and linguistic differences might also influence how text and images are optimally weighted, as languages vary in their descriptive precision for different domains. As such, care must be taken when drawing cognitive implications from machine-based results. Looking forward, our findings point toward several promising research directions. The development of adaptive weighting systems that can dynamically adjust reliance on different modalities based on signal quality could enhance robustness in real-world applications like medical imaging, autonomous navigation, and accessibility technologies. Extending our approach to temporal data (video paired with narration) could illuminate how cross-modal rescue effects operate over time. Finally, incorporating more sophisticated models of human attention and perception could help bridge the gap between computational findings and cognitive theory, potentially leading to more human-like multimodal systems and deeper insights into the fundamental nature of conceptual representation across different forms of sensory input.

APPENDIX

Project Artifacts

Project code: The complete code repository is publicly available at: https://github.com/cucupac/cognitive-science-project
AI instructions: Detailed instructions for generating AI text descriptions are provided in: `ai_instructions.py` (see repository above).
Project guidelines: Original project instructions from Georgia Tech can be found at: https://gatech.instructure.com/courses/436436/files/57844847?wrap=1
These materials support transparency and reproducibility of the work presented.

REFERENCES

[1] Cucupac, A.D. (2025). *Cognitive Science Project GitHub Repository*. Retrieved from https://github.com/cucupac/cognitive-science-project
[2] Cucupac, A.D. (2025). *AI Instructions for Description Generation*. Retrieved from https://github.com/cucupac/cognitive-science-project/blob/main/preprocessing/descriptions/ai_instructions.py
[3] Georgia Institute of Technology. (2025). *Project Instructions*. Retrieved from https://gatech.instructure.com/courses/436436/files/57844847?wrap=1
[4] Dogs vs. Cats Redux: Kernels Edition. (n.d.). Retrieved from https://www.kaggle.com/c/dogs-vs-cats-redux-kernels-edition/data
[5] OpenAI. (n.d.). *CLIP: Connecting Text and Images*. Retrieved from https://openai.com/index/clip/