Analysing Human-Generated Captions for Audio and Visual Scenes

Irene Martín-Morató, Parthasaarathy Sudarsanam, Tuomas Virtanen

Audio Research Group, Tampere University, Tampere, Finland
{irene.martinmorato, parthasaarathy.ariyakulamsudarsanam, tuomas.virtanen}@tuni.fi

***Abstract*—This work investigates how humans describe audio and visual content by analysing single-sentence captions for each modality. While prior research has focused on improving captioning models and their evaluation, less attention has been paid to how linguistic features differ across modalities. We analyse the distribution of parts of speech and domain-specific vocabulary and examine how a structure-based method and neural network-based model classify captions as audio-based or image-based. The structure-based approach reveals how audio captions include verbs related to sound production (e.g., *heard*, *speaking*, *playing*), while image captions use verbs describing physical actions (e.g., *sitting*, *walking*, *holding*). We also study how the input captions influence neural network predictions using gradient-based attribution. Attribution scores from integrated gradients reveal that words like *growling*, *sounded*, *howling*, and *chirp* strongly support audio classification, while words like *grouped*, *cupcakes*, and *participates* are linked to image captions.**

***Index Terms*—audio captioning, image captioning, classification, natural language processing.**

1. INTRODUCTION

Audio and image captioning have become prominent tasks in multimodal AI, aiming to generate natural language descriptions from audio or visual signals. The task of automatic captioning consists of generating coherent and relevant textual descriptions from a media input, therefore the machine should be able to interpret and communicate perceptual information in a human-readable form. Examples of applications that can benefit from this capability are content retrieval [1], human-machine interaction [2] and automated media annotation [3]. In addition to direct applications, captioning data is also used in training multimodal language models. These models aim to understand and generate textual descriptions from different modalities, such as image [4], audio [5] and video [6].

While captioning systems aim to describe perceptual content, understanding how humans naturally perceive and describe their surroundings can provide valuable insights into the design and evaluation of such systems. Foundational work in visual perception by Gibson [7] and in auditory perception by Gaver [8] explores perception as a direct interaction with the environment. Gaver's approach to auditory events was inspired by Gibson's ecological theory of vision, and both emphasize the idea of the actionable properties of objects and events. However, the nature of these properties differs across modalities: visual perception focuses on spatial layout, motion, and surface characteristics, while auditory perception categorizes sounds based on the events and materials that produce them. These ecological perspectives have been used to better understand scene and event-based perception, and they help explain why captions may differ between audio and image modalities.

Understanding perceptual differences is essential, but equally important is how these perceptions are translated into language during the annotation process, which can vary significantly across datasets. Recent work has examined the impact of dataset characteristics on caption quality. For example, in [9] authors investigate differences across datasets from various modalities (audio, image, and video) with a particular emphasis on the data collection process. Their study highlights how the formulation of annotation instructions significantly influences the nature and quality of the collected captions. This is an important aspect to consider, as these captions are later used to train the models, which tend to reproduce the same linguistic patterns in their outputs. In [10] authors argue that when evaluation relies on standard captioning metrics such as BLEU [11], METEOR [12] or SPIDEr [13], among others, models can achieve high scores simply by exploiting dataset patterns. They also compare the Part-Of-Speech (POS) patterns in machine generated and human-annotated captions, showing that human-annotated captions are more complex and diverse than the machine-generated ones. In [14], authors investigate how human and model-generated image captions vary semantically and expressively across languages. Their results show how captioning models could benefit from a multilingual dataset, to achieve more diverse and semantically richer descriptions. Although they focus on a single modality, their work highlights the difference in perception among different users and languages and how the model is affected by this differences.

Audio and image modalities represent different kinds of information, so the way people describe them also differs. Studying linguistic differences, such as sentence structure and vocabulary, is essential for understanding how each modality is expressed in language and for designing better captioning models. To the best of our knowledge, there are no existing studies that specifically examine the linguistic differences between audio and image captions. Understanding these differences is important because audio and image captions are used to train multimodal models, and each modality may require distinct linguistic structures to capture its unique characteristics.

In this paper, we present a detailed comparative analysis of caption content across audio and image captioning datasets. Rather than proposing a new model or metric, our goal is to understand the linguistic and structural characteristics of captions and see what makes a caption more likely to belong to one modality. By analysing vocabulary usage, POS distributions, and semantic patterns, we highlight both shared trends and modality-specific differences. Additionally, we apply interpretability methods to identify which words influence a neural network's decision when classifying captions into belonging to one domain or the other. This study aims to deepen our understanding of how humans describe their surroundings in different modalities, and how these linguistic patterns influence the behavior and decision-making of models trained on such captions.

2. DATASETS

In this work, we have used a diverse set of benchmark datasets from both the image and audio captioning domains to support our comparative analysis of caption content. These datasets vary in size, domain, and linguistic characteristics, offering a broad foundation for studying caption structure, vocabulary, and semantics. From the image domain, we include three widely used datasets in image captioning: MS COCO [15], Flickr30k [16] and Conceptual Captions [17]; From the audio domain, we include four datasets that represent a

The funding for this work is supported by Jane and Aatos Erkko Foundation through the CONVERGENCE of Humans and Machines project.

Table 1: Overview of the datasets analysed in this study.

Dataset	Description	Number of captions	Average caption length (in words)
CLOTHO	Mturk annotated, with curated stage, audios from FreeSound	29 614	11.3
AudioCaps	Mturk annotated, audios from Youtube	98 611	8.5
MACS	Volunteers annotators, three acoustic scenes	16 264	9.5
WavCaps	Formed by AudioSet SL, BBC sound effects, FreeSound, SoundBible metadata and tags, use LLM to generate captions	330 701	8.4
MS COCO	Mturk annotated, images from Flickr	593 968	10.5
Flickr30k	Crowdsourcing annotated, images from Flickr	158 439	12.3
Conceptual captions	Raw descriptions are from the Alt-text HTML, images from Google	2 361 004	10.2

Table 2: Part-of-speech (POS) tag distribution in the audio and image datasets, shown as normalized frequencies in parentheses. With examples of the top five most frequent words per POS tag, first from the audio domain, followed by the top five from the image domain.

POS tags	Top five words
Nouns	Audio (0.33): sounds, man, background, noise, wind Image (0.37): person, man, background, woman, people
Verbs	Audio (0.19): heard, speaking, making, playing, talking Image (0.10): sitting, standing, holding, looking, walking
Adjectives	Audio (0.05): human, other, small, male, loud Image (0.09): white, young, black, old, beautiful
Adverbs	Audio (0.02): then, by, nearby, loudly, repeatedly Image (0.02): next, just, together, very, outside
Adpositions	Audio (0.08): in, with, of, on , by Image (0.16): of, in, on, with, at
Pronouns	Audio (0.01): someone, something, there, it, their Image (0.03): his, it, her, that, I
Determinants	Audio (0.12) Image (0.15): a, the, an, this, some
Others	Audio (0.18) Image (0.07)

range of acoustic environments and captioning styles: CLOTHO [18], AudioCaps v2 [19], MACS [20] and WavCaps [21]. An overview of the datasets is presented in Table 1.

For the remainder of this paper, we refer to the three image captioning datasets collectively as the *image dataset*, and the four audio captioning datasets as the *audio dataset*. The linguistic analysis is conducted on these two aggregated datasets, and used for training the caption classification models. To evaluate the proposed methods we use AVCaps dataset [1], an audio-visual dataset that contains separate textual captions for the audio, visual, and audio-visual contents of video clips. This dataset allows for the independent analysis of audio and visual information, as well as the study of how video captions represent a combination of both modalities, which is often imbalanced. For evaluation, we use the training, validation and test splits of the dataset together. The analysis is conducted separately per caption type: audio (captions based only on the audio modality), visual (captions describing silent videos without audio), AV (captions describing both audio and visual content), and GPT_AV (AV captions rephrased using a large language model, providing a balanced representation of the visual and audio content). Please refer to [1] for more details on the dataset creation.

3. LINGUISTIC ANALYSIS OF CAPTIONS

This section presents a detailed linguistic analysis of the image and audio captioning datasets, which share a total of 17699 common words. Part-of-Speech (POS) tags are an important part of the analysis and evaluation process in captioning tasks, as they help assess the grammatical structure and linguistic quality of generated captions. POS tagging helps to examine whether models produce syntactically coherent sentences and to what extent they capture the diversity of natural language. For instance, analysing the distribution of nouns, verbs, adjectives, and adverbs in generated captions can reveal whether a model tends to overuse object-centric descriptions (e.g., nouns) while neglecting actions or attributes (e.g., verbs and adjectives) [22], [23]. This kind of analysis is particularly useful in both image and audio captioning, where the goal is not only to identify content but also to describe it in a way that is informative and human-like.

Table 2 shows the distribution of POS tags in the audio and image datasets. The values are normalized frequencies, meaning they represent the proportion of each POS tag in the total number of words. The last column gives examples of the most frequent words for each POS tag. Nouns are the most common POS tag in both datasets, with slightly higher frequency in the image domain (0.37) than in audio (0.33). Verbs are more frequent in the audio dataset (0.19) compared to the image dataset (0.10), which reflects how humans typically describe generic audio using nouns, describing the sound sources and verbs, describing their action [24]. Adjectives appear more often in the image dataset (0.09) than in audio (0.05), likely because visual descriptions often rely on adjectives.

When we take a closer look at the top three POS tags, nouns, verbs, and adjectives, we can observe clear differences in vocabulary between the audio and image domains. In the audio domain, captions often include sound-related nouns such as sounds, noise, wind, birds, and music. In contrast, image captions tend to use nouns that refer to people, objects, or places, such as person, man, woman, people, player, view, city, and street. A similar pattern appears with verbs. Audio captions frequently include verbs related to sound production, such as heard, speaking, making, playing, talking, recorded, and singing. On the other hand, image captions often use verbs that describe physical positions or actions, like sitting, standing, holding, looking, walking, and wearing. Finally, adjectives also reflect these domain-specific differences. In the audio domain, adjectives such as male, loud, female, distant, present, high, and low are commonly used to describe sound characteristics. In contrast, image captions more often include adjectives related to visual appearance, such as white, black, red, young, old, and beautiful. Adverbs and adpositions show similar patterns in both datasets, although adpositions are more frequent in the image domain (0.16 vs. 0.08). Pronouns and determiners are relatively low in both datasets, but determiners are more common than pronouns. The ”Others” category is higher in the audio dataset (0.18), which include interjections, numerals, symbols or unclassified words.

4. STRUCTURE-BASED CAPTION CLASSIFICATION

To understand whether main semantic content of a caption can reveal its modality, we construct Subject-Verb-Object (SVO) triplets for both audio and image captions. The goal is to determine whether a caption can be classified as audio-based or image-based by analysing

its SVOs. An SVO triplet represents a basic syntactic structure in which a subject performs an action (verb) on an object, for example, "people-make-noise". This approach is conceptually aligned with the use of scene graphs in visual understanding, as introduced in [25]. In their work, scene graphs are used to represent visual content through structured relationships between objects, typically in the form of subject-predicate-object triplets (e.g., "man-riding-horse"). Similarly, our use of SVO triplets captures the core semantic structure of a more general caption, not being modality-specific.

These triplets were extracted using the open-source Natural Language Processing library spaCy[1]. Prior to extraction, we preprocessed the captions by removing punctuation and applying lemmatization to ensure consistency in the vocabulary. Additionally, we filtered out non-informative or malformed triplets such as "that-have-it," "this-have-it," "that-have-be," "this-have-be," and "it-be-there," which did not contribute meaningful semantic content.

Table 3: Top 5 SVO triplets for audio and image modalities. The values in parentheses represent normalized frequency, calculated as the count of each triplet divided by the total number of triplets.

Audio triplets (freq)	Image triplets (freq)
Bird-make-call (0.0048)	Actor-attend-premiere (0.0010)
Man-speak-noise (0.0041)	Man-stand-next (0.0007)
Bird-chirp-background (0.0029)	Man-stand-front (0.0006)
Car-pass-by (0.0026)	Actor-arrive-premier (0.0005)
There-be-sound (0.0024)	Man-sit-bench (0.0005)

By analysing each caption from both the image and audio datasets, we extracted Subject-Verb-Object pairs/triplets and compiled domain-specific SVO lists, referred to as image-SVO and audio-SVO throughout the rest of the paper. Example: "a trolley train is approaching and it is ringing a bell", extracted SVO: ('trolley train', 'approach') and ('it', 'ring', 'bell'). Once all the SVOs are extracted from all the captions, the number of each triplet/pair is counted and normalized based on the total size of the audio/image dataset. Table 3 shows top 5 collected triplets extracted after analysing all the audio and image datasets.

Table 4: Structure-based caption classification into audio/image by caption type.

Caption type	Audio (%)	Image (%)
audio_captions	3690 (76.80%)	1115 (23.20%)
visual_captions	929 (16.85%)	4584 (83.15%)
AV_captions	1121 (20.07%)	4464 (79.93%)
GPT_AV_captions	2208 (41.97%)	3053 (58.03%)

We collected all the splits (train, validation, and test) from the AVCaps dataset and combined them into a single set for evaluation. Then, we separated the captions by caption type: audio, visual, AV (audio visual) and GPT_AV (rephrased using LLM). Classifying AVCaps dataset by modality (audio/image) based on their linguistic content allows us to explore which actions and entities are more typical of each modality.

Steps to classify a given caption as audio or image caption based on their grammatical structure.

1) Preprocess the caption: remove punctuation and apply lemmatization to standardize word forms.
2) Extract SVO structure: identify the Subject-Verb-Objects from the caption.
3) Compare with reference list: find in which list, audio-SVO or image-SVO, the extracted SVO triplets/pairs is.
4) Determine the domain: classify the caption to the domain (audio or image) based on the audio-SVO and image-SVO counts. In case of the extracted SVO belonging to both list, the one with higher normalized frequency will be selected.

Table 4 shows the average classification results by caption type. Our structure-based classification method achieved high accuracy, correctly identifying 76.80% of audio captions and 83.15% of visual captions. On the AV_captions type, almost 80% are classified as image, similar to the visual captions case. These results align with the observations in [1] that AV_captions are visual-centric, while the GPT_AV_captions are more balanced.

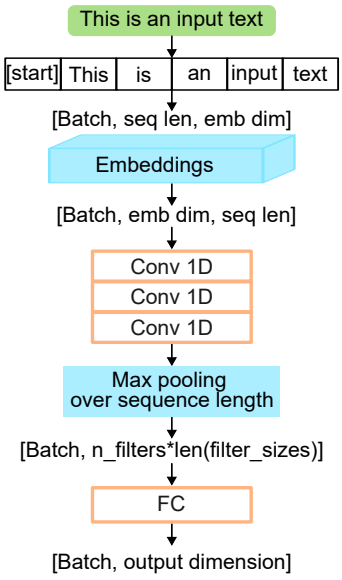

Fig. 1: CNN model for sentence classification. The input is a rearranged to fit 1D convolution layers, each convolutional layers has 100 filters with sizes 3, 5 and 7. After ReLU activation and max pooling, the outputs are combined and sent to a fully connected layer with one output neuron for binary classification.

5. NEURAL NETWORK-BASED CAPTION CLASSIFICATION

In this section, we present a neural network-based classification system to compare against the structure-based classification method presented in Section 4. The model was trained to classify whether a caption is from the audio or image domain, based on its textual content.To accommodate for a large corpus and improve generalization, we used pre-trained word vectors from the wiki-news-300d-1M.vec model [26]. These embeddings provide rich semantic information learned from a large text corpus, which helps the model understand word meanings and relationships more effectively.

The model architecture was selected to be a simple convolutional neural network (CNN) model. The architecture is based on the model proposed by Kim [27], originally designed for sentiment analysis of text (e.g., classifying positive vs. negative reviews). The structure of our model is illustrated in Figure 1. For training, we split both the *image dataset* and *audio dataset* into training and validation sets, maintaining the same proportion for each. Since the datasets are imbalanced, having more image than audio captions, we applied a resampling strategy to balance the classes during training. Finally, input to the model has a dimensionality of [number of tokens, embedding dimensionality], where the embedding dimensionality is fixed at 300, matching the size of the pre-trained word vectors.

[1]https://spacy.io/models/en#en_core_web_trf

Table 5: Neural Network-based caption classification into audio/image by modality.

Caption type	Audio (%)	Image (%)
audio_captions	4386 (71.78%)	1724 (28.22%)
visual_captions	787 (10.36%)	6812 (89.64%)
AV_captions	1001 (12.92%)	6745 (87.08%)
GPT_AV_captions	1895 (30.65%)	4288 (69.35%)

Table 5 shows the average classification results by caption type. Our classification model achieved high accuracy, correctly identifying 71.78% of audio captions and 89.64% of visual captions.

When comparing the results with the structure-based classification, we observe that it assigns slightly more captions to the audio category than the neural network-based method. In contrast, the neural-network based classifier consistently labels a higher percentage of captions as visual across all caption types. This suggests that the neural-network based model is more confident, or biased, towards the visual modality. One possible explanation for this behavior is overfitting in the CNN model. Although we applied a resampling strategy to balance the training data, the overall dataset still contains significantly more visual captions than audio ones. This imbalance may have influenced the model to favor the visual class during prediction, especially in ambiguous cases.

6. MODEL INTERPRETABILITY

In the structure-based method, is possible to interpret the output because classification is based on SVO triplets. In contrast, the neural network-based classification behaves more like a black box, making it harder to understand why a sentence is classified as an audio or image caption. To address this, we use integrated gradients, introduced in [28], a method to understand which input features contribute most to the model's prediction. It works by comparing the model's output on the actual input to a baseline input (e.g. all zeros), which represents the absence of features. The method computes gradients of the model's output with respect to the input. Integrated gradients satisfy two key axioms: Sensitivity, which ensures that features affecting the output receive non-zero attribution, and implementation invariance, which guarantees consistent attributions for functionally equivalent models. This makes it a theoretically sound and interpretable method for explaining deep learning predictions.

For our experiments we use Captum[2], an open source library for model interpretability built on PyTorch. To obtain word-level attribution scores, we apply the Integrated Gradients method and sum the attribution scores across all embedding dimensions for each word. This gives us a single attribution score per word, indicating how much that word contributed to the model's final decision. The overall attribution score for a sentence is then the sum of its word-level attributions. In our case, illustrated in Figure 2, the model is trained to classify sentences into two modalities, class 0 corresponds to image captions; class 1 corresponds to audio captions.

The attribution scores help us understand how much each word pushed the model toward or away from predicting a specific class. A positive attribution score means the word contributed toward predicting class 1 (audio), while a negative score indicates a push toward class 0 (image). Table 6 shows the top ten words with highest and lowest attribution scores, with words like growling, sounded, howling, chirp, and chirping as strongly associated with the audio modality. In contrast, words with strong negative attribution scores include grouped, cupcakes, administers, powerful, and participates are more aligned with the image modality. If we look at the top 10 most frequent words in the AVCaps dataset, we have words like "speaking" (0.64), "talking" (0.58), and "singing" (0.42) have high positive attribution scores, indicating that they strongly support the model's prediction of the audio class. These words are semantically related to sound actions. In contrast, words such as "sitting" (-0.54), "man" (-0.24), and "child" (-0.19) have negative attribution scores, suggesting they are more indicative of the image class. These terms typically describe visual scenes or entities, which are more likely to appear in image captions. Interestingly, some high-frequency words like "playing" (0.09) and "baby" (-0.11) have attribution scores closer to zero, indicating a more neutral or ambiguous role in the classification task. This could be due to their presence in both audio and image contexts, making them less discriminative.

[2]https://captum.ai/docs/extension/integrated_gradients

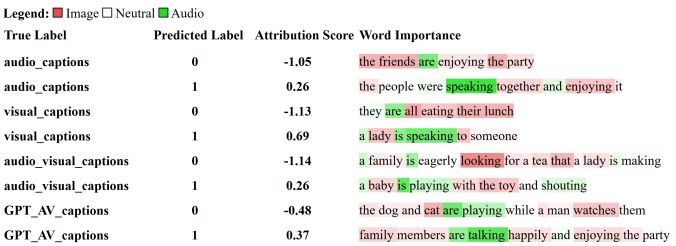

Fig. 2: Visual representation of the word-level attribution in the model output prediction. The green and red color represents audio (class 1) and image (class 0) respectively, while the intensity indicates attribution strength. The first column is the true label, the second column is the predicted label and the third column is the attribution score.

Table 6: Top 10 words with the highest and lowest average attribution scores in the AVCaps dataset. Positive scores indicate stronger association with audio captions, while negative scores suggest stronger association with image captions.

Word	Average attribution	Word	Average attribution
growling	0.88	badge	-0.76
sounded	0.83	attracts	-0.76
howling	0.82	yogurt	-0.76
chirp	0.81	celebrity	-0.76
mumbling	0.77	purchased	-0.79
requests	0.75	participates	-0.80
murmurs	0.72	powerful	-0.81
conversed	0.71	administers	-0.83
meows	0.69	cupcakes	-0.85
hums	0.69	grouped	-0.89

7. CONCLUSION

In this work we have studied the linguistic and structural differences between audio and image captions, highlighting how these differences and similarities affects the model outputs. The linguistic analysis revealed clear modality-specific patterns, with audio captions favoring sound-related vocabulary and image captions focusing on visual elements such as people and scenes. Through the Subject-Verb-Object structures, we created a modality-specific references, which were then compared to an CNN classifier. While the CNN model performed well, it showed a consistent bias toward the visual modality, likely influenced by dataset imbalance. To further interpret model behavior, we applied integrated gradients, which confirmed that words strongly associated with sound (e.g., growling, chirping) positively contributed to audio predictions, while visually descriptive terms (e.g., grouped, cupcakes) supported image predictions. These findings emphasize the importance of understanding linguistic patterns in multimodal datasets, as they directly shape model outputs and can inform the design of more balanced and interpretable captioning systems.

REFERENCES

[1] P. Sudarsanam, I. Martín-Morató, A. Hakala, and T. Virtanen, "Avcaps: An audio-visual dataset with modality-specific captions," *IEEE Open Journal of Signal Processing*, pp. 1–15, 2025.

[2] Z. Meng, L. Yu, N. Zhang, T. Berg, B. Damavandi, V. Singh, and A. Bearman, "Connecting what to say with where to look by modeling human attention traces," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 674–12 683.

[3] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6847–6857.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139, 18–24 Jul 2021, pp. 8748–8763.

[5] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid, "Learning audio-video modalities from image captions," in *17th European Conference on Computer Vision (ECCV) European Conference*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 407–426.

[7] J. J. Gibson, *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, 2014.

[8] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[9] L. Hekanaho, M. Hirvonen, and T. Virtanen, "Language-based machine perception: linguistic perspectives on the compilation of captioning datasets," *Digital Scholarship in the Humanities*, vol. 39, no. 3, pp. 864–883, 06 2024.

[10] M. Dai, S. Grandic, and J. C. Macbeth, "Linguistic variation and anomalies in comparisons of human and machine-generated image captions," *Advances in Cognitive Systems*, vol. 8, p. 335, 2019.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318.

[12] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 228–231.

[13] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 873–881.

[14] A. Ye, S. Santy, J. D. Hwang, A. X. Zhang, and R. Krishna, "Semantic and expressive variations in image captions across languages," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 29 667–29 679.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *13th European Conference Computer vision (ECCV), zurich, Switzerland, September 6-12*. Springer, 2014, pp. 740–755.

[16] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 02 2014.

[17] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, Jul. 2018, pp. 2556–2565.

[18] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[19] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132.

[20] I. Martín-Morató and A. Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 76–80.

[21] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[22] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] I. Martín-Morató and A. Mesaros, "Diversity and bias in audio captioning datasets," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 90–94.

[24] B. Gygi and V. Shafiro, "Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–12, 2010.

[25] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[26] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2018.

[27] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751.

[28] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70, 06–11 Aug 2017, pp. 3319–3328.