Controlling Semantic Consensus and Lexical Density in Image-Based Evaluations for Mobile Input Methods

ANDREAS KOMNINOS, University of Patras, Greece
IOULIA SIMOU, University of Patras, Greece
JOHN GAROFALAKIS, University of Patras, Greece

Mobile input methods (IMEs) for text entry in interactive applications, are typically evaluated in lab settings, using phrase transcription tasks as the de-facto procedural element in the validation stage of their development. Transcription tasks offer strong internal validity but weak ecological and external validity, due to the composition constraints imposed on participants. Free composition tasks using image stimuli have been proposed as an alternative, but they lack standardized control to regulate elicited text volume and quality, therefore limiting evaluation internal validity. In this paper, we propose a methodology for rigorous, computational sampling of image stimuli from large image datasets, with an aim to control lexical and semantic similarity in user-generated text during IME evaluations. We further deploy diffusion generative AI models to methodically derive stylistic variations from the sampled images, allowing us to examine the effects of original images, abstractive styles and distractor elements on user input. Our findings from an online image description study with crowdsourced participants ($N = 100$) and 3,000 captured input samples, demonstrate that stimulus sampling and restyling can be used a methodological control apparatus for evaluating input methods in lab settings, allowing researchers to systematically control participants' input density and diversity, mitigating the internal validity challenges of image task evaluations. We present evidence-based guidelines for selecting image stimuli and release our datasets, experiment application, analysis code, and participant descriptions.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing design and evaluation methods*; **Text input**; **Laboratory experiments**; **Usability testing**.

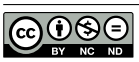
Additional Key Words and Phrases: Input methods, Text Entry, Mobile devices, Lab evaluations, Image tasks, Stimulus selection

ACM Reference Format:
Andreas Komninos, Ioulia Simou, and John Garofalakis. 2026. Controlling Semantic Consensus and Lexical Density in Image-Based Evaluations for Mobile Input Methods. *Proc. ACM Hum.-Comput. Interact.* 10, 4, Article EICS021 (June 2026), 40 pages. https://doi.org/10.1145/3816773

1 Introduction

Mobile input methods (IMES) such as virtual keyboards, are, in and of themselves, a form of complete interactive system, dependent on direct touch, gestural or gaze input, and providing output as visual, auditory or haptic feedback. Beyond translation of user input to alphanumeric characters or other symbols (e.g. emojis), they offer advanced user interface elements such as predictive suggestions, automatic corrections, LLM-driven text transformations. Like all interactive systems, IMEs need to be engineered, i.e. systematically specified, designed, developed and tested, on their own, as per [31], and also as part of interactive systems that depend on it. Regardless of input modality, such as

Authors' Contact Information: Andreas Komninos, Computer Engineering & Informatics Department, University of Patras, Rio, Achaea, Greece, akomninos@ceid.upatras.gr; Ioulia Simou, Computer Engineering & Informatics Department, University of Patras, Rio, Achaea, Greece, simo@ceid.upatras.gr; John Garofalakis, Computer Engineering & Informatics Department, University of Patras, Rio, Achaea, Greece, garofala@ceid.upatras.gr.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2573-0142/2026/6-ARTEICS021
https://doi.org/10.1145/3816773

speech, gestural, or traditional keyboard-based techniques, novel input methods or assistive systems (e.g. autocorrection algorithms, text suggestion or replacement algorithms) need to be evaluated in lab and field settings (in-the-wild), though lab evaluations are dominant in literature, due to costs and privacy concerns inherent in field studies. In the lab environment, text entry evaluations typically consist of *Transcription* tasks, which require participants to copy a short phrase presented to them. Literature suggests these tasks can be complemented (or sometimes replaced entirely) by freeform text entry tasks, namely *Composition* or *Image*-based tasks. These alternative approaches present a textual or image prompt (stimulus) to users, asking them to create freeform text according to the instructions (e.g. "imagine that you are sending an SMS to a friend relating to…", or "please describe what you see in this image") [11, 58]. Transcription tasks use empirically validated phrase sets that are memorable, representative of language or otherwise deemed suitable for the task. In contrast, there are very few validation studies to explore the suitability of various stimulus types in composition tasks, and no studies for image tasks.

Transcription tasks have remained the de-facto evaluation method for lab settings. While composition and image tasks remain attractive due to their higher ecological and external validity, perhaps the strongest reason for the prevalence of transcription-based evaluations is that the other task types have a significant internal validity disadvantage, owing to the unrestricted text that participants contribute [58]. Freeform text-entry tasks like composition and image tasks, will, by nature, result in varied responses across participants, both in terms of input length, and input words used to complete the task. This makes it hard to compare text contributed between the participants. Even when participants are presented with the same stimulus, they are likely to use semantically similar, but lexically different words to describe what they see. Although researchers cannot force participants to use the same wording to describe a concept, experiment design and outcome risks might be mitigated if researchers can predict whether given an image stimulus, participants can a) perceive and focus on the same concepts in the stimulus and b) use semantically similar, if not identical words during their input. Less variation in the semantics (and language) of text being provided by participants would mean that the experiment can be better controlled, while maintaining the increased ecological and external validity inherent in the reflective process, preceding, and also running in parallel with the input activity.

In this paper, we present a methodological contribution that focuses on image-based tasks, and posit that, depending on their content and visual features, some images may be better suited than others in achieving the goal of homogenising participant input. Further, we posit that other image types beyond photos (e.g. drawings, sketches, or otherwise manipulated photos) may elicit different input behaviour in users. Our work covers the systematic exploration of an evaluation method (the Image Task) and its related materials (image stimuli) that can be used to support the validation stage of IME development lifecycles, and to assist research and engineering stakeholders in IME development projects.

In summary, we perform a linguistic and visual feature analysis on the well known Flickr30k image and description dataset [64]. We propose a methodology for the systematic sampling of this dataset, and then use transformer-based generative AI models to derive stylistic variations based on the original sampled images to use as alternative stimuli. We conduct an online experiment with English-speaking participants (N=100) to elicit image descriptions for each type of stimulus and compare the outcomes in terms of linguistic features (e.g. input length, lexical variation, semantic similarity). Our findings show that stylistic variations have a distinct impact on the elicited text and that it is possible to anticipate both the length of contributed text, as well as its semantic and lexical properties, by selecting appropriate image types. We provide our derived stimuli and accompanying phrase sets as new, validated material that researchers may consider in planning and executing future image-based or transcription-based text entry evaluations (see Section 7 for availability).

2 Related Work

We review extant literature across three relevant domains. This section begins with an overview of human visual perception and cognitive factors that affect the formation of internal mental representations (schemata) of visual imagery, as these form the basis of the textual descriptions of images that participants may provide. Next, we discuss findings from large crowdsourced image description experiments with humans, focusing on the diversity of linguistic features used to express the formed schemata in text, as well as relevant metrics. Finally, we discuss the state of the art in the use of image stimuli in IME evaluation experiments. This section concludes with a summary of findings and their implication for our work.

2.1 Visual perception in humans

Cognitive psychology experiments demonstrate that when humans observe visual stimuli, they encode the information into mental schemata (representations), which are formed in short-term memory, shortly after exposure [40]. What will be encoded depends heavily not just on the contents of the visual stimulus, but also on the long-term memory representations already encoded from previous experiences, as well as the associations we make with these encodings and other experiences [9, 10]. As such, our recollection of what we just saw is a process of reconstruction, which may contain omissions (inability to remember details) or additions (recalling seeing things that *should* be there), rather than a pure playback account of the experience [8, 20]. These factors therefore plausibly contribute to the linguistic, and sometimes semantic variability in human descriptions of images. Yet, the determining factor is ostensibly participants perception and attention to salient elements within the same image. Human perception of visual scenes is selective, with observers extracting a high-level interpretation of a scene without processing all elements with equal depth. A key component of visual perception is visual attention, which focuses on a subset of objects and regions within a scene. Computational models of visual saliency have been proposed, to approximate bottom-up attention by predicting which parts of an image are likely to attract attention based on their low-level features [21, 22]. Wolfe and Horowitz [62] provide a broad review of what visual attributes can guide the deployment of attention during visual search tasks. They report that attributes such as colour, orientation and size support attention guidance, whereas more complex properties do not guide attention in the same way.

Beyond general findings on selective attention, several studies have examined visual attention in the context of image description tasks. Elazary and Itti [12] used the LabelMe dataset [45] to treat what people choose to annotate in an image as a proxy for what they find interesting enough to describe. Their results suggest that what humans attend to, and therefore are likely to mention in their descriptions, is shaped by low-level visual features. Berg et al. [5] examined how humans assigned importance to visual elements in images, treating the mention of these elements in produced captions as an indicator of perceived importance. Their results showed that what people chose to describe was influenced not only by low-level features, such as the size and the position of the element, but also by semantic and contextual factors, such as whether an object is animate, an object's category or whether an objects appears in unusual settings. The attentional attraction effect of *unusual* objects was first documented in a 1979 study by Friedman, who found that the placement of out-of-context salient objects in an image scene (e.g. a penguin in a kitchen), is a strong attractor of human attention [15]. Berg et al. further demonstrated that these factors can be used to predict what content will be mentioned in a description, linking image properties to human description behaviour [5]. He et al. [18] introduced a dataset of synchronously recorded eye movements and image descriptions, enabling direct analysis of visual attention during caption

generation. Their results indicated that the described objects were likely to have been fixated, while also reporting different gazing patterns during description generation versus free-viewing tasks.

2.2 Crowdsourced image descriptions and linguistic diversity

Image description tasks are a form of free text-entry task, in which participants generate text in response to a visual stimulus. Such tasks are widely used in image annotation, captioning, and accessibility contexts and have resulted in the creation of several natural-language datasets, where multiple descriptions are collected for each image. Early work by Rashtchian et al. [42] presented a caption dataset created through crowdsourcing, where non-expert participants provided a simple textual description for the presented visual stimulus (image). Subsequent datasets such as *Flickr30k* and *Flickr30k Entities* further established image description as a human text-entry task, by collecting multiple captions per image for photographs depicting people and everyday activities [64] or by providing enriched annotations that link noun phrases to image regions without constraining how descriptions are elicited [39]. Authors in [28] adopted a different stimulus design philosophy to create the Microsoft COCO dataset, focusing on complex, object-centric scenes containing multiple objects. Crowdsourcing has also been employed in other contexts, such as by Vedantam et al. [55], who utilized images from online datasets to create two image caption datasets including cartoon-like imagery, and by Gurari et al. [17], who collected captions for real-world images captured by people who are blind.

Past studies reveal substantial variability across descriptions of the same stimulus. Rashtchian et al. [42] reported high variability in the collected descriptions, despite providing explicit instructions to the participants. In 2014, Young et al. [64] introduced the concept of visual denotations to measure image description similarity, and specifically to handle the fact that lexically different expressions may refer to the same visual entities (notably, this was before the first embedding model techniques started to gain traction e.g. Word2Vec [35]). Image captions collected by Gurari et al. [17] further highlight the lexical variability across captions for images taken in unconstrained settings. They showed that vision-language models trained on curated datasets such as Flickr30k [64] and MSCOCO [28] performed poorly in providing suitable descriptions for these images, highlighting a substantial gap between curated stimuli and images taken "in-the-wild". While this finding is presented in the context of algorithmic performance, it also underscores the broader impact of stimulus characteristics on description generation.

To account for this identified lexical variability several relevant metrics have been proposed. Among them, Vedantam et al. proposed *CIDEr* [54], a consensus-based caption evaluation metric to measure how most people describe an image, rather than relying on exact lexical overlap, computing similarity to multiple reference captions using TF-IDF weighted *n*-grams and rewarding informative words. As an alternative to CIDEr, Anderson et al. [1] proposed *SPICE*, arguing explicitly that *"n-gram overlap is neither necessary nor sufficient for two captions to convey the same meaning"*. Instead, the SPICE metric was designed to compare captions at a semantic level by parsing them into scene-graph tuples and computing an F-score over them, thereby treating lexical variation as incidental when the underlying semantic content is common. The development of language embedding models such as [35], led later researchers to evaluations of their ability to assess image caption similarity compared to previous algorithmic approaches such as CIDEr or SPICE. The highly influential BERTScore paper demonstrated the high effectiveness of embedding models in assessing text similarity in a variety of use cases, including machine translation and image captioning [65]. In [47], Sharif et al. demonstrated the surprising effectiveness of embedding cosine similarity as a metric for image descriptions, compared to SPICE, CIDEr and Word Mover Distance [25] at system-level correlation with human judgments, using a variety of embedding models.

2.3 Free Composition IME evaluation methodologies

IME evaluations are traditionally conducted using transcription tasks. In these, participants are asked to copy pre-selected phrases from carefully curated phrase sets *as quickly and accurately as possible*. Transcription tasks afford high experimental control and facilitate comparability across studies, thus becoming the de facto evaluation method in text entry research [33]. The phrases in the standardized phrase sets are selected to be moderate in length, memorable and representative of the target language, enabling controlled and repeatable evaluation [24, 32, 38, 59]. As a result, evaluation metrics such as Words Per Minute (WPM), Keystrokes Per Character (KSPC) and Corrected and Uncorrected Error Rates (CER, UER) have been widely adopted in text entry studies [2, 34, 51, 61], allowing researchers to attribute observed differences in users' performance to the input method under study.

Despite their advantages, transcription tasks have been criticized for their limited external and ecological validity, as they do not require participants to generate text or engage in text planning [58]. Prior work has also shown that transcription performance can be affected by the language of the phrase set, when the pre-selected phrases are in a language that differs from participants' native language [14]. To address these limitations, *Composition* tasks have been proposed as a complement to transcription tasks [58]. In composition tasks, participants are allowed to freely enter text, better approximating real-world text entry behaviour and improving external validity. Vertanen et al. [57] used a composition task alongside a transcription task to observe more natural typing behaviour, evaluating how different input chunk sizes affected virtual keyboard decoding performance on a smartwatch, including entry speed, error rates and users' preferred strategy. Arnold et al. [4] used an extended composition task, where participants wrote restaurant reviews on a mobile device, rather than simply transcribing predefined phrases, to evaluate how multi-word phrase suggestions influenced suggestion uptake and writing behaviour and how they shaped the resulting written content versus single-word predictions. However, the increased realism of composition tasks introduces additional challenges. Variation in typing speed and accuracy can no longer be attributed solely to the input method under study [32, 58], as users' performance may also be influenced by the cognitive processes involved in text planning and generation. Subsequent work has also shown that participants may adapt their writing strategies in composition tasks, for example by selecting simpler words, thereby affecting their measured performance [16].

A variation of the composition task uses image stimuli to prompt participants free writing (the *Image* task). Dunlop et al. [11] were, to our best knowledge, the first who proposed this variation. In a small-scale experiment, they documented that the Image task elicits longer texts that include out-of-vocabulary language, and consists a more ecologically valid reflection of real-world text entry. Like the heavily contextualised instructions in composition tasks that constrain users' language [4], image tasks may also have a restrictive effect on participant vocabulary, since the likely words to be used in the description are largely dictated by the visual contents of the presented stimulus. Since then, only a handful of text-entry studies appear to employ image-based composition tasks. Expanding on [11], Nicol et al. introduced an image-based composition task as a means to obtain text entry performance metrics from older adults [36]. In Arnold et al. [3], image stimuli were used to elicit the typing of captions with various levels of predictive text assistance, in order to assess the impact of intelligent text entry support methods on typing behaviour. In work by Foley et al. [13], participants were presented with three clip-art images and were instructed to compose a phrase of their own thinking, that would include the words representing these images. Finally, most recently, Li et al. [27] used an image-based composition task as a means to examine effects of visual attention and checking behaviour during typing tasks with text suggestions. All of these studies employ public-domain image sets as a source for their stimuli. However, only [3] and [27] employ

some type of metric to select image stimuli in a non-random manner. Arnold et al. [3] manually selected photos from the Flickr30k set [64] that fall into three categories depending on the number of salient characters (persons or animals) present in the photos (one, two or multiple characters). Li et al. [27] took a more methodical approach, selecting pairs of photos from the COCO dataset [28] where the provided image captions were the most similar, based on the TF-IDF similarity of unigrams in the concatenation of all five of the captions that crowd workers had originally entered for each image. They then selected one of the photos in each pair randomly.

2.4 Overview of Related Work and Research Aims

Overall, prior work suggests that human-generated image descriptions are not arbitrary, but can be shaped by stimulus characteristics that can guide attention and influence what experiment participants may mention in their produced descriptions. As a result, collections of image stimuli may elicit descriptions that vary lexically and semantically based on the entities participants choose to attend to and describe. This highlights the importance of careful design and selection of image stimuli, particularly for text-entry evaluations where control over the produced content might be desired. However, to our knowledge, no research has been carried out to answer whether certain visual stimuli are more suitable than others for eliciting consistent descriptions from human participants, or how stimulus properties may influence text-entry behaviour. Prior work also highlights a clear trade-off in text entry evaluation between the experimental control afforded by the well-established transcription tasks and the external and ecological validity provided by free text-entry tasks, such as composition an image description. While transcription benefits from standardized phrase sets and performance metrics, free text-entry tasks are more suited to eliciting natural writing behaviour. Nevertheless, this happens at the cost of increased variability in the produced text, and a lesser ability to compare outputs and performance across participants. Image description has been adopted both for dataset creation and, more recently, for input method evaluation, but the substantial lexical and semantic variability remains an inherent challenge. Relevant metrics have been proposed to account for lexical variability when comparing captions. However, these metrics do not account for the variability that arises at an earlier stage, where differences in perception and visual attention influence which entities are selected for description.

The reviewed body of work motivates the need for specific approaches to designing and selecting image stimuli for text-entry evaluation tasks, where the goal is to maintain the naturalistic input behaviour of free text-entry tasks while improving control over the lexical and semantic content that is produced. In this paper, we attempt to systematically evaluate image stimulus alternatives, including original images taken from well-known datasets, as well as to explore abstractive or distracting stylistic variations on these images. Our aim is to identify the effect of image stimulus types on the generated text descriptions, and therefore enable informed choice and control over experimental conditions to researchers interested in employing non-transcription task methodologies to investigate natural text entry behaviours in any setting (e.g. mobile, VR, AR etc.).

3 Generating a Diverse Image Stimulus Dataset for Text Entry Evaluations

3.1 Overview of Methodology

To achieve our goal, we begin with a well known dataset (Flickr30k [64]) which contains 31,783 photos sourced from the Flickr.com website, each described by five crowdsourced human workers. The Flickr30k dataset extends an earlier set of approximately 8,000 images [19], using the same crowdsourcing process to obtain 5 descriptions for each image. In this process, independent remote workers were asked to describe the image shown to them, using a concise (single sentence, under 100 characters) description of the activities, people, animals and objects depicted in the image.

Based on this data, we then identify visual and textual features in the original images and their accompanying descriptions, in an attempt to separate these images into categories of description congruency (i.e. stimuli that elicit text descriptions of high, medium or low variability), using appropriate metrics.

We propose a methodology for the rigorous selection of a small but adequately representative number of samples from each category (10 images). The number of images per congruence category needs to be small for two reasons. First, we aim to use these in an experiment with human participants where the target is to obtain 20 descriptions per image, which is significantly more compared to the 5 descriptions per image in the Flickr30K set (Section 4). Second, since we aim to examine the effects of stylistic variations derived from each original image for each participant, exposing them to four stylistic alternatives, the total number of image stimuli presented to participants will grow significantly. Therefore, to balance experiment validity, ability to compare between the baseline original and stylistic variations, as well as experiment duration and participant fatigue, we need a small number of adequately representative original images from each category. To derive stylistic variations, we use state of the art Diffusion generative models [50]. This section details our methodology to compile a final set of 150 images to present to human participants, categorised by original image description congruency and stylistic variation, in order to obtain a new set of descriptions and compare between them. The rest of this section details the final image set generation process. The process is shown diagrammatically in Fig. 1 and is detailed in the next subsections, with implementation details outlined in the paper appendices. All data and code for analysis is provided as a supplement to this paper (see Section 7 for details). Prompts for large language and diffusion models used in our work are provided within the code and we invite the reader to inspect these for clarity.

3.2 Steps 1 & 2: Data separation into congruence categories

Our first target is to examine the degree of similarity between the five descriptions of each image in the Flickr30k set. For this purpose, we employed several metrics, on both lemmatized and unmodified versions of the descriptions for each image, based on its accompanying textual descriptions (see Table 1). All metrics except the last one (embeddings) capture linguistic diversity in a statistical manner, however are unable to capture semantic similarity, which may account for individual differences in explaining the same concept. For example, when shown an image of a girl holding a flower, one participant might describe it as *"A young girl holding a flower"*, while another might write *"A child with a flower in hand"*, or *"A little girl holding a daisy"* and so forth. As is obvious, all of these variations are semantically identical, yet statistical measures would show them to be very different. As is evident from literature, it is unrealistic to expect that participants will use exactly the same words to describe stimuli shown to them during an image-based evaluation task. Instead, we can aspire to guide their attention so that they can provide a limited range of probable words, which we might anticipate (e.g. *[girl, child], [hold, in-hand], [flower, daisy]*). To account for semantic similarity despite lexical variation, we use a state-of-the-art (SOTA) language embedding model (EmbeddingGemma [56]) to obtain description embeddings in a 768-dimensional space. An analysis of these metrics (see Appendix A.1) demonstrated that some of these metrics were well correlated, and that the embeddings-based metric seems to capture overall variability better than other metrics.

Based on the analysis of linguistic metrics, we conclude that embedding similarity might be the most appropriate metric by which to characterise set of images and descriptions (Fig. 1, Step 1). Further, we note from the previous analysis that the correlations do not differ significantly between the unprocessed and lemmatized descriptions, so we select to proceed with only the embedding similarity metrics. We divide the subset into three congruence categories (high, medium and low)

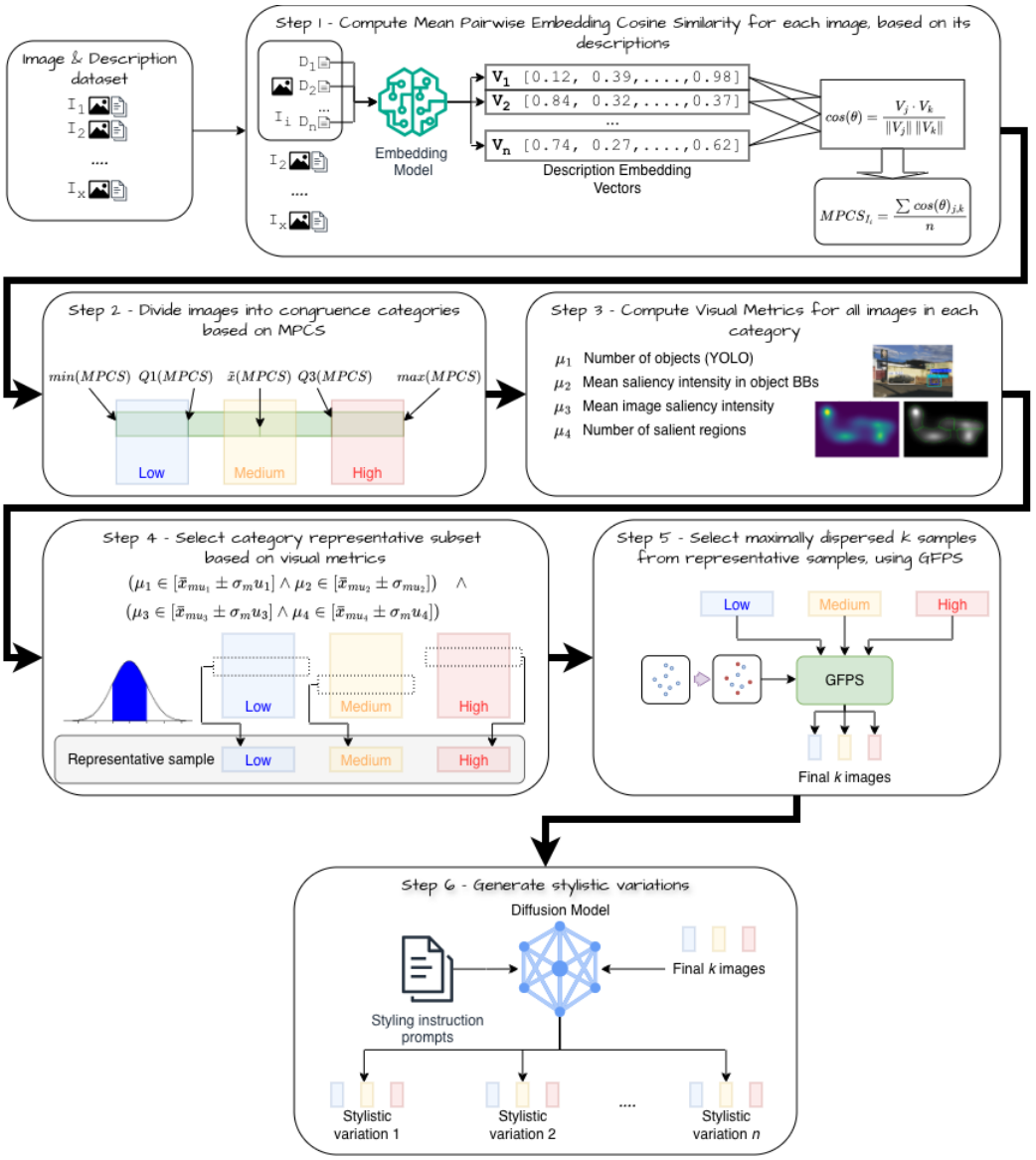


Fig. 1. Overview of the dataset sampling and stylistic variation generation process.

by taking the first quartile (0-25%) of the unprocessed description average embedding similarities range, the inner half of the IQR (37.5-62.5%) and fourth quartile (75-100%), as shown in Table 2 (Fig. 1, Step 2).

3.3 Steps 3 - 5: Data sampling based on visual and linguistic metrics

As the number of images in each congruence category is large (see Table 2), we employed two further techniques to derive four more metrics based on visual features in the images, so as to extract *representative* samples from each category (Table 3 and Fig. 1, Step 3). First, we used the

Table 1. Linguistic Metrics for Flickr30k Image Descriptions

Metric	Description
TF-IDF Similarity	Average of all pairwise cosine similarities between caption unigram vectors within a single image's description set [41].
Jaccard Similarity	Average of all pairwise Jaccard similarity coefficients for the image descriptions within a single image's description set [41].
Entropy	A measure of the unpredictability or information content within a specific description set [38].
KL Divergence	Relative entropy of an image's description set measured against the global corpus distribution of the subset [38].
Perplexity	A measurement of how well the probability distribution of the caption set predicts the sample [38].
Embedding Similarity	Average of all pairwise cosine similarities between caption embedding vectors within a single image's description set [43, 56].

	Low		**Medium**		**High**	
Quartile Bounds	0 (min value)	0.25	0.375	0.625	0.75	1 (max value)
Embedding Value	0.360	0.629	0.661	0.714	0.743	0.968
***N** Samples*	7951		8002		7917	

Table 2. Congruence categories by average description embedding value.

YOLO object detection model [23] (version 11x, latest version at the time of writing), to detect objects present in the scene depicted in each image. YOLO11x only offers recognition of about 80 types of object without additional fine-tuning, which is clearly not enough to capture the full breadth of objects in the Flickr30k dataset. Yet, it can offer an indication of scene complexity, as busier scenes will likely contain more recognisable objects. Further, we used TranSalNet, a state-of-the-art transformer-based model to estimate visual saliency in images, trained on human visual attention data [29]. TranSalNet generates a saliency map for any image input, therefore we can assess the visual complexity of a scene by estimating the mean map intensity over the entire image. We also measured the mean saliency map intensity in the bounding boxes of objects detected by YOLO. Finally, to overcome the limitation of YOLO classes, we use the Watershed algorithm [6] to segment the saliency map into regions, centered on areas of intense saliency. This algorithm treats the saliency map as a topographical map, where areas of high intensity form local peaks, around which image segments can be identified. It is particularly effective at overcoming the limiting effects of local maxima, therefore can identify segments of low intensity as well as high intensity ones. In this manner, we are able to obtain a better estimate of the business of the image, not in terms of actual objects present in it, but instead, attention-drawing objects and features that are semantically important to humans, thus likely to be mentioned in the image description. Figure 2 demonstrates these concepts in an example image from Flickr30k. While YOLO is able to detect only three distinct objects, the watershed-based segmentation identifies five areas of interest. As can be seen in Table 4, the focus on the sign (left of the image) and wording on the building (right of the image) leads to explicit mentions of the sign (id 1) and deduction this might be a gas station (id 3). Both the saliency segments and YOLO identify the bicycle as a central focal point in 4 out of

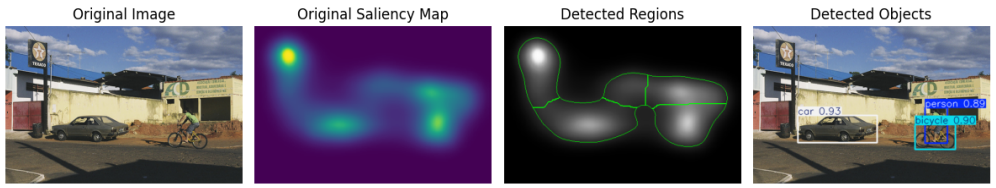


Fig. 2. Example of an original Flickr30k image, its saliency map, watershed segmentation of the saliency map, and YOLO11x object detection.

5 descriptions, while the car features only in one description. Notice also how the same object is described differently across captions ([person, someone, man] and [bicycle, bike]).

Table 3. Visual Metrics for Flickr30k Images

Metric	Description
YOLO objects	Number of detected objects based on YOLO11x default categories [23].
YOLO saliency	Mean saliency map intensity across bounding boxes of all YOLO11x detected objects [23, 29].
Image saliency	Mean saliency map intensity across entire image [29].
Salient regions	Number of salient regions detected using the watershed algorithm across entire image [6].

Consequently we calculate, for each image, the number of YOLO-detected objects, the number of identified salient regions, the mean salience intensity across the entire image, and the mean intensity (salience) across the YOLO-detected object bounding boxes. We then employed these visual metrics to extract a small number of adequately representative samples from each congruence category. We selected from each congruence category a subset of images whose visual metrics are within one standard deviation of the mean of the respective category metrics (Fig. 1, Step 4). This process yields 90 images for the low congruence category, 171 images for the medium category and 89 for high. To select the final 10 images from each of these subsets, we first employ a SOTA vision-based LLM (QWEN3-VL 8B Instruct 4-bit quantized model [63]) to derive a generic description of each image and compute its embedding vectors using the EmbeddingGemma model [56]. Finally, in each category, we use greedy farthest point sampling (GFPS) with multi-start optimization ($n = 1000$ random initializations), to obtain a set of 10 images, so that they are maximally dispersed in the

id	Description
1	Someone wearing a green shirt and jean knee high jeans riding a bike on the side of the road about to pass a gray car and a texaco sign
2	A person is riding a bike along a street in a desolate city setting .
3	A vacant gas station among the shadows of the street corner .
4	A person riding a bicycle wearing a green shirt .
5	A man on his bike pedals down the street

Table 4. Descriptions for the original Flickr30k image shown in Figure 2.

768-dimension embedding space [60] (Fig. 1, Step 5). The process in Step 5 is fully outlined in Appendix B.

3.4 Step 6: Generating image variations

Once the final stimulus set of 10 images per congruence category was selected, we set out to perform a series of image transforms, to examine how abstraction in stimulus rendering affects perception and therefore description (Fig. 1, Step 6). For this, we deploy a SOTA open source image-to-image diffusion model (FLUX.2-klein-9B, 8-bit quantized [26]) and instruct it to perform the following stylistic changes (image variations): a) apply a *plakatstil* poster style to the image (V0); b) apply a *pencil* drawing style to the image removing unimportant details and focusing details on salient objects or persons only (V1), and c) apply a simplified black and white *cartoon* drawing style to the image removing unimportant details and focusing details on salient objects or persons only (V2). Additionally, we implemented a distractor transformation, where an unusual object was in-painted within the original image (V3).

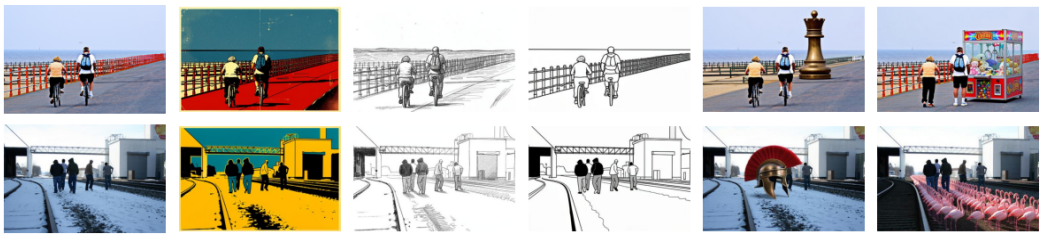
The distractor (V3) variation was inspired by a well known effect in visual processing where human attention focuses strongly on infrequently encountered objects, or objects that are unexpected for the setting displayed in an image stimulus (e.g. a penguin in a kitchen), and that objects naturally related to the scene do not receive much attention (therefore unlikely to be encoded in memory and recalled as part of a description exercise) [5, 15]. We posited that the manipulation of original images to include unexpected objects might strongly influence the participants description and lead to more controlled text entry outcomes as per the original paper by Friedman [15]. The distractor images were generated with the same diffusion model (FLUX.2-klein-9B, 8-bit quantized [26]) and the process is fully described in Appendix C. Examples of the resulting transformations are shown in Fig. 3. For the rest of the paper, we use the symbols V[0-3] for the image variations as described above, and the symbol V4 for the original unmodified image. The full set of images (originals and derivations) is provided with this paper (see Section 7).

4 Image Task Online Study

As our goal is to assess whether our stimulus selection pipeline can be used to control the qualitative and quantitative attributes of text contributed by participants in IME evaluation studies, we proceed with a crowdsourced image description task study, using $N = 100$ native English-speaking participants, recruited via the Prolific platform. In this study, we are not interested in evaluating new IMEs, but rather to confirm that our stimulus selection (and stylistic variation) can be used to support to the validation stage of the IME development lifecycle. Hence we allow users to employ their own habitual IME as installed on their mobile device, and focus the analysis on the produced text.

The study was set up so as to capture 20 descriptions from unique participants for each image in the full set of originals and stylistic variations. Each participant provided descriptions for 30 images (3 blocks of 10 images, one block per stylistic variation, including original images). We limited participant's exposure for two reasons: first, to prevent participants from viewing a single image in more than one of its stylistic variations, and, second, to prevent participant fatigue and likely degradation of text quality and quantity. As a result, and because participants did not experience all possible stylistic variation styles, we use mixed effects models to analyse the results.

The study environment consisted of a mobile web app to present the stimuli (see Fig. 4). Each stimulus was presented to participants for 10 seconds, giving ample time for participants to move visual content from gist memory into conceptual short-term memory [10, 40] and to form a structured visual representation [9] (Fig. 4a). The image was then hidden, and participants were shown instructions on how to provide a description. We hid the stimulus instead of permanently



(a) High congruence stimuli



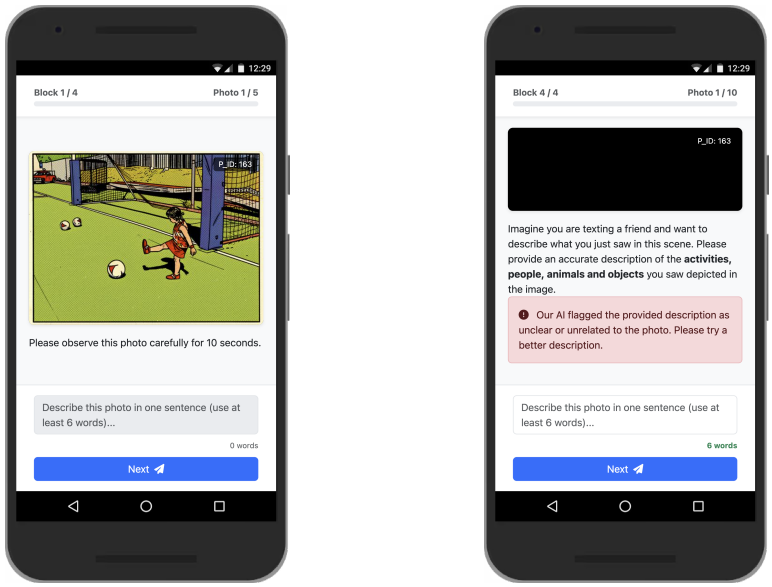
(b) Medium congruence stimuli



(c) Low congruence stimuli

Fig. 3. Image transformations applied to original Flickr30k images. From left to right: original image, plakatstil, pencil, cartoon, QWEN3 inspired distractor, Gemma-3 inspired distractor.

displaying it, in order to capture the internal schema formed by participants [8] without risking artificial inflation of the text due to prolonged exposure (i.e. submitting secondary non-salient details). Participants' input was bound programmatically to a length between 6 and 50 words. On our server, a small vision-enabled LLM (Gemma-3-4b LLM, 8-bit quantized) checked the submitted description against the actual image, and if a very low consistency or relevance was detected (e.g. entering gibberish, or a description that did not match the image at all) participants were asked to provide a more suitable one (see Fig. 4b). Participants used their habitual smartphone keyboard and we did not restrict use of advanced entry support features (e.g. autocorrect, glide typing, word suggestions etc.) so as not to impose additional cognitive burdens that might artificially impact the core of the experiment (visual memory schema recall while typing) and threaten its internal validity with confounding factors [58]. Finally we should mention that participants took a guided



(a) Sample stimulus presentation (practice run) (b) Feedback after AI-detected inconsistent input.

Fig. 4. Online application experiment interface.

practice run of 5 images, with instructions to demonstrate the study conditions, before the actual blocks of images. In the rest of this section, we present summaries of the analyses which lead to our guidelines for sampling and using image stimuli to support the IME validation stages in lab experiments. Full details of the study, procedure and analyses of results can be found in Appendices D (online study description) and E (online study data analyses).

4.1 Description lengths

We first examined description length in terms of character counts and word counts, after basic pre-processing including punctuation and stopword removal. Details of this analysis can be found in Appendix E.2. Summarily, we observe that participants contribute relatively long text descriptions for each image (on average, 80 characters or 16 words). More than half of this content consists of *interesting* words (i.e. non-stopwords). Taking the High congruency original image (V4) as the baseline, we observe that it elicited an average meaningful text length of 49.39 characters ($SE = 2.14$) and 7.89 ($SE = 0.337$) words. A mixed effects model revealed a significant main effect for category: Low Congruence images naturally elicited significantly more text (+19.11 chars, +2.86 words; $p < .001$) compared to the baseline. We also find that in High congruence images, when the scene was deemed semantically clear according to original descriptions, in our experiment, mild abstraction (Plakatstil V0, Pencil V1) had no significant effect on text length. However, introducing strong stylistic abstraction via Cartoon (V2) or focal distractions (V3) significantly increased text entry (V2: +15.83 chars / +2.32 words; V3: +9.03 chars / +1.48 words; all $p < .001$). Medium Congruence images largely mirrored the baseline, with the exception of the Cartoon (V2) variation, which significantly reduced text length ($\beta = -19.94$ chars, $p < .001$). Finally, for the Low congruence category, every stylistic transformation (V0–V3) significantly reduced the amount of text entered compared to the original photo ($p < .001$ for all). In scenes where participants usually struggle to agree on what to describe (leading to verbosity), abstraction (V0-V2) appears to act as a constraint.

The stylistic filters likely remove the background details or ambiguities that caused the original divergence, effectively pruning the scene down to its essential elements only. This guides participants to focus on fewer, more salient elements, resulting in shorter, more concise descriptions. The same effect is also emergent when introducing a distractor element, without abstracting visual details from the scene.

4.2 Description Similarity

For description similarity, we employ the previously used Jaccard similarity, TF-IDF similarity and Embedding similarity measures, reporting again for descriptions after stopword removal. Jaccard and TF-IDF based metrics represent the lexical consensus in participants' descriptions, while the Embeddings metric represents semantic consensus. Details of this analysis can be found in Appendix E.3. Overall, the baseline condition (High Congruence, Original Photo V4) revealed a fundamental characteristic of free-text entry. While semantic consensus was high (MECSS-S = 0.722, matching the original Flickr30k descriptions), lexical consensus was notably low (Jaccard = 0.218; TF-IDF = 0.184). This indicates that even when participants perfectly agree on what they are seeing (i.e. the image concept), they naturally select diverse vocabulary to describe it. Both Medium and Low Congruence original images (V4) exhibited significantly lower similarity scores across all three metrics compared to the baseline, confirming that the ambiguity in descriptions from the Flick30k dataset is also present in our own collected data. The effects of stylistic variations are more nuanced and depend on the congruence category of the image. In the High congruence category, abstractive styles (V0–V2) resulted in small but statistically significant decreases across all three metrics. We could assume that by removing the high-fidelity visual cues of the original photo, these styles forced participants to interpret the abstraction, introducing noise into both word choice and semantic focus. Notably, the Distractor variation had no statistically significant effect on similarity in High Congruence scenes. Unlike the abstractive styles (V0–V2), the addition of a distinct object did not disrupt the baseline consensus.

In Medium and Low Congruence scenes, this relationship inverted. Stylistic transformations acted as a unifying force, mitigating the natural ambiguity of the images, particularly for the Low category, where we observe the most statistically significant effects. In the Low Congruence category, all stylistic variations (V0–V3) showed statistically significant positive effects in lexical alignment metrics, increasing similarity. However, the magnitude of this lexical rescue was small ($\beta \approx 0.04$ to 0.06). This suggests that while abstraction helps converge vocabulary slightly, participants still rely on diverse synonyms. Further, in the Low category, all variations (V0–V3) improved semantic consensus, though the effect is more pronounced for V3, particularly when considering the level of improvement across all three metrics. Finally, for the Medium category, the Plakatstil (V0) style stood out as a powerful semantic homogenizer, showing a large increase in semantic similarity ($\beta = 0.102, p < .05$).

4.3 Language representativeness

Finally, we examined the representativeness of the descriptions with regard to the English language, against the Mackenzie and Soukoreff [32] and Vertanen & Kristensson [59] phrase sets, typically employed in transcription tasks. The main metric for this was Kullback–Leibler divergence (KLD) on character-level bigrams, using the Google Books nGrams Data (version 20120701) as a reference corpus, A non-parametric bootstrap (B=500, sample size=50) was used to estimate the 95% confidence intervals of the KLD for all 3 phrase datasets (ours and [32, 59] relative to the reference corpus. Our dataset demonstrated a divergence that rests between that of the two other phrase sets, confirming adequate linguistic representativeness with little variation across image category and version (see Fig. 5). This intermediate score indicates that our captured descriptions maintain a high degree of

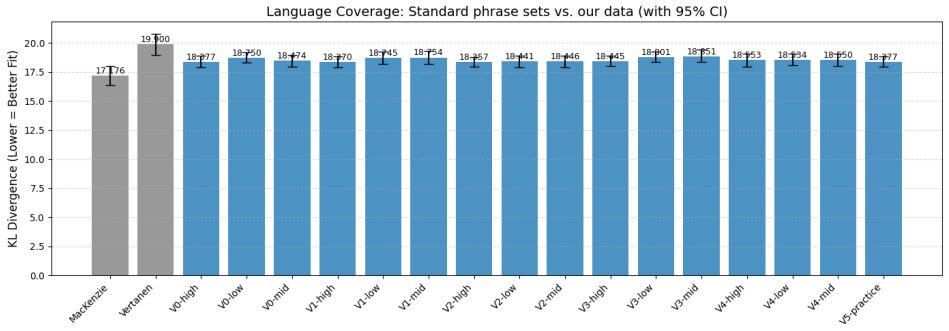


Fig. 5. Representativeness of bigram frequency distribution against the Google Books corpus, for standard transcription task phrase sets, and elicited image descriptions per congruence category and image version, measured in KL-Divergence (relative entropy).

lexical diversity and grammatical structure (outperforming the Vertanen set in standard language coverage) while capturing more naturalistic variation than the strictly formal MacKenzie set. This suggests that the image tasks are successful in providing a balanced and adequate challenge to participants, during text entry evaluation. More details of this analysis can be found in Appendix E.4.

5 Discussion

Motivated by the limitations of current methodological practice in mobile text entry evaluations, and notably the transcription task, we attempted a thorough investigation into the selection of image stimuli for the alternative, less-researched image-based task. Our aim was to examine the potential for controlling the internal validity aspect of this alternative task, hoping to find evidence that it can be considered as a complementary, if not main, lab based evaluation methodology for novel input methods and assistance tools. In the process, we made innovative use of state-of-the-art capabilities afforded by recent developments in Generative AI, and demonstrate how these can be useful in helping future researchers shape methodological practice and answer difficult questions that might have been impractical to address a decade ago. Our work provided some useful contributions which we outline next, but is not without limitations. We discuss these in the following section, along with open issues raised by our findings.

5.1 Contributions

5.1.1 Implications for Text Entry Experiment Designs. Our core contribution is to inform experiment design in text entry studies, where the Image Composition task might be chosen as a more ecologically valid choice compared to traditional transcription tasks. Given a large dataset like Flickr30k as a source from which to select visual stimuli for an image composition task, the question we tried to address was how to select the most appropriate ones. The results from our online study confirm that the stimulus selection process was successful, since it actually caused the desired effects when using original images during a text entry task. The behaviour during stylistic images is an additional contribution showing how original sets can be used as a basis to elicit additional behaviours.

The findings are nuanced according to the desired experiment outcomes. If the target of the task is to elicit text entry that is as homogeneous as possible, our results show that the most controlled output (i.e., highest semantic and lexical consensus with predictable text volume) is obtainable

		Text Similarity	
		High	**Low**
Text Volume	**High**	High Congruency + Distractor	Low Congruency + Original
	Low	High Congruency + Original	Low Congruency + Abstractive

Table 5. Recommendations for image selection and transformation applications, according to desired experiment outcomes, based on the Flickr30k dataset.

through selecting images using the criteria we apply for assignment into High Congruency, using the mean pairwise embedding similarities of the Flickr30k dataset as a selection criterion, according to the thresholds provided in Table 2. These photos naturally elicit the highest lexical and semantic similarity ($\approx$ 0.77 embedding score) and the briefest text entry ($\approx$ 49 characters), keeping the tasks short. Adding a visual distractor to these images maintains the high lexical and semantic similarity, while having the effect of also increasing the volume of submitted text, making the tasks longer. If the experiment might benefit from more diverse participant input, the original (V4) Low congruence photos are particularly effective at eliciting both vocabulary sprawl and additional length per task. Otherwise, to maintain vocabulary and semantic diversity while keeping tasks short, we recommend applying an abstractive style to the Low congruency images. These recommendations are summarised in the Table 5.

5.1.2 Further contributions. In this paper, we present a complete processing pipeline that allows researchers to sample image stimuli from the Flickr30k dataset. One advantage of this pipeline is that all steps are automated, with manual inspection only necessary at the end of Step 6, to inspect the quality of derived variations (see Appendix C), which requires minimal effort. This method can be applied to any other dataset similar to Flickr30k, including extant datasets for specific application domains, e.g. to investigate text entry for medical doctors typing a diagnosis (e.g. [28, 30, 44, 48, 49, 52]). Apart from the processing pipeline and code, we provide the full, ready-to-use set of image stimuli in our experiment as a supplement to this paper (see Section 7). Researchers are invited to consider these for further use in image description tasks for mobile text entry studies, as a validated choice of stimuli, along with the elicited descriptions that can provide sufficient indication of the text that can be expected from their presentation.

Our experiment produced a significant corpus of human-sourced text entry. The phrases we collected can be used as a source (target stimulus) for traditional text entry evaluations using the transcription task (see Section 7 for availability of this phrase set). Since the language coverage based on bigram distribution is between the more formal phrases in the Mackenzie set and the more mobile-oriented phrases in the Vertanen set, these stimuli present a balanced benchmark covering a wider spectrum of use-cases than either set, particularly using phrases from the Low-V1 subset that exhibit the shortest character and word-wise lengths to keep tasks short. The captured phrases represent the structure of formal English without being archaic, and the brevity of mobile text without being broken (in the sense that they present coherent phrases, not out-of-context word sequences as in some of the Vertanen set (e.g. *"Get with Mary for format", "What is up with ENE?"*). This is critical for wearables (smartphones, smartwatches, AR glasses) where interaction is short but often commands structured replies. Further, our captured phrases may present a better benchmark for stress-testing predictive text algorithms, compared to the Mackenzie set (which, composed of mostly standard English, is easy to train for and predict) or the Vertanen set (which contains many out-of-vocabulary words or breaks grammar rules), although a formal analysis of the syntactic and grammatical patterns in these datasets (and ours) would be needed to confirm this. Finally, our

dataset might be useful in assessing text entry evaluations under variable cognitive load conditions. The High/Medium/Low congruency segmentation suggests a variable cognitive load for human participants while examining and forming memory schemata of the stimuli. Therefore the stimuli could provide future researchers with a granular instrument to evaluate how input techniques degrade under varying levels of linguistic entropy, which could be a critical factor obscured by monolithic average WPM scores.

5.2 Limitations and Future Work

We enforced the use of mobile devices in this experiment, and as such, the effects of abstraction might be magnified by the motor constraints of mobile typing. We cannot discount the possibility that on a physical keyboard, participants might have typed long descriptions regardless of the image style because typing is more fluid. Nevertheless, given the length of text provided by participants in other image styles, we believe the observed effects are largely attributable to the abstraction, not motor constraints or fatigue from experiment duration. However, the results might not fully generalize to desktop-based text entry experiments. Further, our experiment was conducted solely in English. The lexical consensus we measured (Jaccard, TFIDF) is highly dependent on the English language's synonym density. We are aware that in languages with higher morphological complexity (e.g., Finnish) or different adjective structures, the impact of visual abstraction on word choice might look very different. Our definition of *congruence* relies on the semantic consensus of the original Flickr30k annotators. While robust for this dataset, this metric is linguistically bound to English descriptions and Western visual norms. The ambiguity we measured could therefore be socio-linguistic construct as much as a visual one.

While we manually inspected the Distractor (V3) images used in the experiment to ensure the set of pre-selected anomalous objects (e.g., penguins, toasters) did not occupy too much space or occlude important salient features, we did not systematically control for the semantic distance or pixel-occupancy ratio of these distractors. Future research should quantify whether larger or more semantically shocking distractors yield stronger focalization effects.

We used an LLM to check for gibberish or otherwise incoherent input. While this was an important factor to ensure quality in a crowdsourced task, this might have artificially affected our participants' input or behaviour, and thus resulting metrics. Although we believe the minimum word count we enforced (6) and the low AI-detected consistency threshold ($t \leq 2/5$) were not unrealistic or overly strict, by rejecting very short or lazy inputs, we effectively clipped the tail of the distribution and as such, average lengths are likely higher than a completely unmoderated crowdsourcing task. On the other hand, AI coherence scores had a very high mean value ($\bar{x} = 4.669, \sigma = 0.670$). This could mean that the participants were forced to submit what the AI also saw in the image, not what *they* actually saw. To mitigate the risk of the AI guardrail 'forcing' a specific description, we iterated on the AI prompt design until we were satisfied with the result, and permissively calibrated the acceptance threshold, filtering only for complete irrelevance (e.g., gibberish or salient elements not present in the image). While description rejection logs were not retained, the impact of the AI guardrail on participant creativity appears minimal. Post-hoc analysis of Jaccard similarity confirms that this mechanism did not artificially homogenize participant input, as lexical diversity remained high (≈ 0.2) across all conditions. Particularly for the most abstractive style (V2), the AI coherence means remain consistently high (V2-High $\bar{x} = 4.790, \sigma = 0.536$; V2-Medium $\bar{x} = 4.415, \sigma = 0.834$; V2-Low $\bar{x} = 4.685, \sigma = 0.639$). Therefore we believe this concern was sufficiently mitigated.

Finally, in our experiment, as inherent to crowdsourcing tasks, participants used their own device. This introduced variability in screen size and keyboard dimensions, it ensures high ecological validity as participants performed the task using the hardware and muscle memory they are most accustomed to. We deliberately did not restrict participants' use of advanced IME features such as

autocorrect, predictive text, or glide typing. While this introduces some variability in input speed and potential lexical priming (via predictive suggestions), it preserves the ecological validity of the study and ensured a fair cognitive load in memorizing the image and recalling its details while providing a description. While advanced features were not explicitly disabled or prohibited, the nature of the crowd-sourcing task (often performed in public or semi-public spaces) and the textual data characteristics suggest that the majority of input was manual. Future studies could explicitly control for modality to isolate the effects of speech-to-text or word suggestion or corrections on image description verbosity.

In this paper, we do not present any analyses of text entry metrics (e.g. WPM, error rates etc.). Although we collected the pertinent data, the scope of such analysis is outside the target of this paper, which was to examine the final outcomes of the text entry process, rather than the process itself. A careful analysis of text entry metrics may yield useful in-the-wild typing behaviour insights particularly in comparing participants who used advanced text entry support against those who didn't. We leave this as future work.

The larger generative AI models used for this paper were ran on a departmental GPU cluster using NVIDIA A100 40Gb GPUs. Smaller models were ran on a desktop workstation using an NVIDIA 3070 8Gb GPU. In this research, we use exclusively open-weight models limited to 14 billion parameters. This is in order to ensure maximum replicability by other researchers using commodity workstation-class GPUs only (e.g. >=16Gb VRAM). Naturally, different generative models might produce other results, particularly in generating image variations. Even if the same models are employed, due to the stochastic nature of generation, the work is not strictly reproducible (each run will produce slightly different results). As such, we release all the data produced in every step, so the findings can be validated by others using our own data. Future work could also investigate more thoroughly the performance of other open or proprietary models in their ability to create image versions with better fidelity. Additionally, future work could investigate if our findings hold while using alternative source datasets (e.g. [28, 30, 44, 48, 49, 52]).

6 Conclusion

As mobile text entry research continues to rely on ecologically valid methodologies to evaluate keyboard performance and predictive models, the need for controlled, replicable data collection methods becomes paramount. Going beyond the traditional transcription task, and avoiding the chaotic nature of low internal validity in-the-wild studies, alternative lab methods such as the Image task can provide naturalistic behaviour within controllable boundaries that do not undermine the internal validity of studies. This study demonstrates that the visual stimulus itself is a critical variable that can be engineered to shape participant behavior.

Through our extensive analysis of the subject, we conclude that the preference for stimulus image fidelity is context-dependent. High-fidelity photography is optimal only for simple, unambiguous scenes. In complex environments, realism invites vocabulary sprawl, whereas stylistic abstraction acts as a necessary constraint to ensure data quality. The introduction of anomalous focal points (visual distractors) is a highly effective method for imposing semantic order on ambiguous scenes, effectively rescuing consensus in otherwise messy datasets. Researchers can and should treat image style as a methodological tool. By selecting the appropriate combination of congruence and abstraction, experimenters can program the expected density and diversity of user input, tailoring the task to the specific needs of the keyboard or language model being evaluated. Future work should extend this framework to non-English languages or even specialized linguistic domains (e.g., medical imaging), further validating the use of visual design as a mechanism for cognitive control in human-computer interaction.

7 Data and Code

We release all our code for Flickr30k dataset analysis, image generation, experiment application interface and experiment data analysis openly as a supplement to this paper. All code is written in Python and released as Jupyter Notebooks and Python scripts. All code ran using the latest PyTorch version (2.10) with CUDA13 support, and SentenceTransformers v5.2.2. Additionally, all data gathered in the experiment is released.

Acknowledgments

This work was supported by [anonymised department] computer centre, we are grateful for the access provided to its GPU cluster that enabled execution of larger LLMs. [Further acknowledgemnts anonymised for review].

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. [n. d.]. SPICE: Semantic Propositional Image Caption Evaluation. https://doi.org/10.48550/arXiv.1607.08822 arXiv:1607.08822 [cs]

[2] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. [n. d.]. Analysis of text entry performance metrics. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)* (2009-09). 100–105. https://doi.org/10.1109/TIC-STH.2009.5444533

[3] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. [n. d.]. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2020-11-17) *(IUI '20)*. Association for Computing Machinery, 128–138. https://doi.org/10.1145/3377325.3377523

[4] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. [n. d.]. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo Japan, 2016-10-16). ACM, 603–608. https://doi.org/10.1145/2984511.2984584

[5] Alexander C. Berg, Tamara L. Berg, Hal Daumé, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, and Kota Yamaguchi. [n. d.]. Understanding and predicting importance in images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012-06). 3562–3569. https://doi.org/10.1109/CVPR.2012.6248100 ISSN: 1063-6919.

[6] Serge Beucher. 1992. The Watershed Transformation Applied to Image Segmentation. *Scanning Microscopy* 1992, 6 (Jan. 1992). https://digitalcommons.usu.edu/microscopy/vol1992/iss6/28

[7] Aysecan Boduroglu, Priti Shah, and Richard E. Nisbett. 2009. Cultural Differences in Allocation of Attention in Visual Information Processing. *Journal of Cross-Cultural Psychology* 40, 3 (May 2009), 349–360. https://doi.org/10.1177/0022022108331005

[8] Timothy F. Brady. 2024. Visual Memory. *Open Encyclopedia of Cognitive Science* (2024). https://doi.org/10.21428/e2759450.9aa55a0f

[9] Timothy F. Brady, Talia Konkle, and George A. Alvarez. 2011. A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision* 11, 5 (May 2011), 4. https://doi.org/10.1167/11.5.4

[10] Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* 105, 38 (2008), 14325–14329. https://doi.org/10.1073/pnas.0803390105

[11] Mark D. Dunlop, Emma Nicol, Andreas Komninos, Prima Dona, and Naveen Durga. [n. d.]. Measuring inviscid text entry using image description tasks. In *In Proceedings of the 2016 CHI Workshop on Inviscid Text Entry and Beyond* (2016).

[12] Lior Elazary and Laurent Itti. [n. d.]. Interesting objects are visually salient. 8, 3 ([n. d.]), 3. https://doi.org/10.1167/8.3.3

[13] Margaret Foley, Géry Casiez, and Daniel Vogel. [n. d.]. Comparing Smartphone Speech Recognition and Touchscreen Typing for Composition and Transcription. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2020-12-23) *(CHI '20)*. Association for Computing Machinery, 1–11. https://doi.org/10.1145/3313831.3376861

[14] Marc Franco-Salvador and Luis A. Leiva. [n. d.]. Multilingual phrase sampling for text entry evaluations. 113 ([n. d.]), 15–31. https://doi.org/10.1016/j.ijhcs.2018.01.006

[15] Alinda Friedman. 1979. Framing Pictures: The Role of Knowledge in Automatized Encoding and Memory for Gist. *Journal of Experimental Psychology: General* 108, 3 (1979), 316–355. https://doi.org/10.1037/0096-3445.108.3.316

[16] Dylan Gaines, Per Ola Kristensson, and Keith Vertanen. [n. d.]. Enhancing the Composition Task in Text Entry Studies: Eliciting Difficult Text and Improving Error Rate Calculation. In *Proceedings of the 2021 CHI Conference on Human*

Factors in Computing Systems (New York, NY, USA, 2021-02-07) *(CHI '21)*. Association for Computing Machinery, 1–8. https://doi.org/10.1145/3411764.3445199

[17] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. [n. d.]. Captioning Images Taken by People Who Are Blind. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII* (Berlin, Heidelberg, 2020-05-23). Springer-Verlag, 417–434. https://doi.org/10.1007/978-3-030-58520-4_25

[18] Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. [n. d.]. Human Attention in Image Captioning: Dataset and Analysis. https://doi.org/10.48550/arXiv.1903.02499 arXiv:1903.02499 [cs]

[19] M. Hodosh, P. Young, and J. Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research* 47 (Aug. 2013), 853–899. https://doi.org/10.1613/jair.3994

[20] Ira E. Hyman. 1993. *Chapter 4 imagery, reconstructive memory, and discovery*. Advances in psychology, Vol. 98. North-Holland, 99–121. https://doi.org/10.1016/S0166-4115(08)60140-5 citation-key: HYMAN199399.

[21] L. Itti and C. Koch. [n. d.]. Computational modelling of visual attention. 2, 3 ([n. d.]), 194–203. https://doi.org/10.1038/35058500

[22] L. Itti, C. Koch, and E. Niebur. [n. d.]. A model of saliency-based visual attention for rapid scene analysis. 20, 11 ([n. d.]), 1254–1259. https://doi.org/10.1109/34.730558

[23] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics

[24] Per Ola Kristensson and Keith Vertanen. [n. d.]. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (New York, NY, USA, 2012-10-14) *(IUI '12)*. Association for Computing Machinery, 29–32. https://doi.org/10.1145/2166966.2166972

[25] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 957–966. https://proceedings.mlr.press/v37/kusnerb15.html

[26] Black Forest Labs. 2025. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2.

[27] Yang Li and Anna Maria Feit. [n. d.]. How We Type with Word Suggestions: Understanding Visual Attention and Checking Behavior during Mobile Text Input. 9, 3 ([n. d.]), 105:1–105:29. https://doi.org/10.1145/3749520

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. [n. d.]. Microsoft COCO: Common Objects in Context. https://doi.org/10.48550/arXiv.1405.0312 arXiv:1405.0312 [cs]

[29] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. 2022. TranSalNet: Towards Perceptually Relevant Visual Saliency Prediction. *Neurocomputing* 494 (July 2022), 455–467. https://doi.org/10.1016/j.neucom.2022.04.080

[30] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2018. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing* 56, 4 (April 2018), 2183–2195. https://doi.org/10.1109/TGRS.2017.2776321

[31] Víctor Manuel López Jaquero, Radu-Daniel Vatavu, Jose Ignacio Panach, Oscar Pastor, and Jean Vanderdonckt. 2019. A Newcomer's Guide to EICS, the Engineering Interactive Computing Systems Community. *Proc. ACM Hum.-Comput. Interact.* 3, EICS (2019), 1:1–1:9. https://doi.org/10.1145/3300960

[32] I. Scott MacKenzie and R. William Soukoreff. [n. d.]. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2003-12-05) *(CHI EA '03)*. Association for Computing Machinery, 754–755. https://doi.org/10.1145/765891.765971

[33] I. Scott MacKenzie and Kumiko Tanaka-Ishii. [n. d.]. *Text Entry Systems: Mobility, Accessibility, Universality*. Elsevier. Google-Books-ID: XWSc3b_gkX8C.

[34] I. Scott MacKenzie, View Profile, Shawn X. Zhang, and View Profile. [n. d.]. The design and evaluation of a high-performance soft keyboard. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 25–31. https://doi.org/10.1145/302979.302983

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations (ICLR2013)*. arXiv. https://doi.org/10.48550/arXiv.1301.3781 arXiv:1301.3781 [cs].

[36] Emma Nicol, Mark D Dunlop, and Andreas Komninos. [n. d.]. A Participatory Design and Formal Study Investigation into Mobile Text Entry for Older Adults. 8, 2 ([n. d.]), 20–46. https://doi.org/10.4018/IJMHCI.2016040102.oa

[37] Richard E. Nisbett and Yuri Miyamoto. 2005. The Influence of Culture: Holistic versus Analytic Perception. *Trends in Cognitive Sciences* 9, 10 (Oct. 2005), 467–473. https://doi.org/10.1016/j.tics.2005.08.004

[38] Tim Paek and Bo-June (Paul) Hsu. 2011. Sampling Representative Phrase Sets for Text Entry Experiments: A Procedure and Public Resource. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2477–2480. https://doi.org/10.1145/1978942.1979304

[39] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. [n. d.]. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. https:

//doi.org/10.48550/arXiv.1505.04870 arXiv:1505.04870 [cs]

[40] Mary C. Potter. 2012. Conceptual Short Term Memory in Perception and Thought. *Frontiers in Psychology* 3 (April 2012). https://doi.org/10.3389/fpsyg.2012.00113

[41] Dimas Wibisono Prakoso, Asad Abdi, and Chintan Amrit. 2021. Short text similarity measurement methods: a review. *Soft Computing* 25, 6 (March 2021), 4699–4723. https://doi.org/10.1007/s00500-020-05479-2

[42] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. [n. d.]. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (USA, 2010-03-06) *(CSLDAMT '10)*. Association for Computational Linguistics, 139–147. https://dl.acm.org/doi/10.5555/1866696.1866717

[43] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[44] Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. 2024. ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset. *Scientific Data* 11, 1 (June 2024), 688. https://doi.org/10.1038/s41597-024-03496-6 arXiv:2405.10004 [eess]

[45] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. [n. d.]. LabelMe: A Database and Web-Based Tool for Image Annotation. 77, 1 ([n. d.]), 157–173. https://doi.org/10.1007/s11263-007-0090-8

[46] Alžběta Šašinková, Jiří Čeněk, Pavel Ugwitz, Jie-Li Tsai, Ioannis Giannopoulos, David Lacko, Zdeněk Stachoň, Jan Fitz, and Čeněk Šašinka. 2023. Exploring Cross-Cultural Variations in Visual Attention Patterns inside and Outside National Borders Using Immersive Virtual Reality. *Scientific Reports* 13, 1 (Nov. 2023), 18852. https://doi.org/10.1038/s41598-023-46103-1

[47] Naeha Sharif, Lyndon White, Mohammed Bennamoun, Wei Liu, and Syed Afaq Ali Shah. 2020. WEmbSim: A Simple yet Effective Metric for Image Captioning. In *2020 Digital Image Computing: Techniques and Applications (DICTA)*. 1–8. https://doi.org/10.1109/DICTA51227.2020.9363392

[48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2556–2565. https://doi.org/10.18653/v1/P18-1238

[49] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From Pixels to Prose: A Large Dataset of Dense Image Captions. https://doi.org/10.48550/arXiv.2406.10328 arXiv:2406.10328 [cs]

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 2256–2265. https://proceedings.mlr.press/v37/sohl-dickstein15.html

[51] R. William Soukoreff, View Profile, I. Scott MacKenzie, and View Profile. [n. d.]. Metrics for text entry research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 113–120. https://doi.org/10.1145/642611.642632

[52] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. MedICaT: A Dataset of Medical Images, Captions, and Textual References. https://doi.org/10.48550/arXiv.2010.06000 arXiv:2010.06000 [cs]

[53] Gemma Team. 2025. Gemma 3 Technical Report. arXiv:2503.19786 (March 2025). https://doi.org/10.48550/arXiv.2503.19786 arXiv:2503.19786 [cs].

[54] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. [n. d.]. CIDEr: Consensus-based Image Description Evaluation. https://doi.org/10.48550/arXiv.1411.5726 arXiv:1411.5726 [cs]

[55] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. [n. d.]. Collecting Image Description Datasets using Crowdsourcing. https://doi.org/10.48550/arXiv.1411.3041 arXiv:1411.3041 [cs]

[56] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, A. J. Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen,

Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. EmbeddingGemma: Powerful and Lightweight Text Representations. arXiv:2509.20354 (Nov. 2025). https://doi.org/10.48550/arXiv.2509.20354 arXiv:2509.20354 [cs].

[57] Keith Vertanen, Crystal Fletcher, Dylan Gaines, Jacob Gould, and Per Ola Kristensson. [n. d.]. The Impact of Word, Multiple Word, and Sentence Input on Virtual Keyboard Decoding Performance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC Canada, 2018-04-21). ACM, 1–12. https://doi.org/10.1145/3173574.3174200

[58] Keith Vertanen and Per Ola Kristensson. [n. d.]. Complementing text entry evaluations with a composition task. 21, 2 ([n. d.]), 8:1–8:33. https://doi.org/10.1145/2555691

[59] Keith Vertanen and Per Ola Kristensson. [n. d.]. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (New York, NY, USA, 2011-05-30) *(MobileHCI '11)*. Association for Computing Machinery, 295–298. https://doi.org/10.1145/2037373.2037418

[60] DOUGLAS J. WHITE. 1991. The Maximal-Dispersion Problem. *IMA Journal of Mathematics Applied in Business and Industry* 3, 2 (Jan. 1991), 131–140. https://doi.org/10.1093/imaman/3.2.131

[61] Jacob O. Wobbrock and Brad A. Myers. [n. d.]. Analyzing the input stream for character- level errors in unconstrained text entry evaluations. 13, 4 ([n. d.]), 458–489. https://doi.org/10.1145/1188816.1188819

[62] Jeremy Wolfe and Todd Horowitz. [n. d.]. What Attributes Guide the Deployment of Visual Attention and How Do They Do It? 5 ([n. d.]), 495–501. https://doi.org/10.1038/nrn1411

[63] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. https://doi.org/10.48550/arXiv.2505.09388 arXiv:2505.09388 [cs]

[64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics* 2 (Feb. 2014), 67–78. https://doi.org/10.1162/tacl_a_00166

[65] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR2020)*. https://openreview.net/forum?id=SkeHuCVFDr

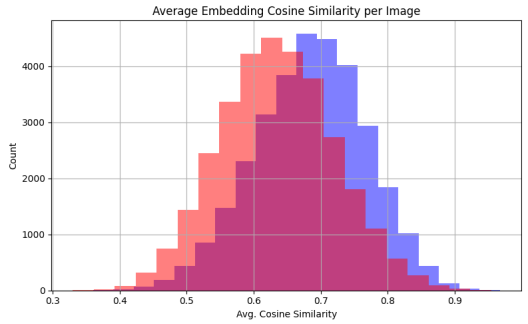


Fig. 6. Average embedding similarities per image, with unprocessed (blue) and lemmatized (orange) descriptions.

A Analysis of Flickr30k Linguistic and Visual metrics

A.1 Linguistic metrics

We begin with the analysis of text descriptions in the Flickr30k dataset. First, we perform a lemmatization step using the Python Natural Language ToolKit (NLTK) library and applying Wordnet-based Part-of-speech tagging, as well as English stopword removal. Subsequent processing is performed on both the original and lemmatized captions. The first metric we explore is TF-IDF based similarity. For each photo, we compute a TF-IDF vector matrix based on the caption unigrams (single words). Then, we compute the average of all pairwise cosine similarities for the image descriptions. Similarly, we calculate the average of all pairwise Jaccard similarities for the image descriptions. Further, for each description set of an image, we compute the entropy, Kullback-Leibler divergence (relative entropy computed against the corpus of all captions in the subset) and perplexity, as described in Paek & Hsu's paper on sampling phrase sets for text entry .

These metrics capture linguistic diversity in a statistical manner, however are unable to capture semantic similarity, which may account for individual differences in explaining the same concept. For example, when shown an image of a girl holding a flower, one participant might describe it as *"A young girl holding a flower"*, while another might write *"A child with a flower in hand"*, or *"A little girl holding a daisy"* and so forth. As is obvious, all of these variations are semantically identical, yet statistical measures would show them to be very different. As is evident from literature, it is unrealistic to expect that participants will use exactly the same words to describe stimuli shown to them during an image-based evaluation task. Instead, we can aspire to guide their attention so that they can provide a limited range of probable words, which we might anticipate (e.g. *[girl, child], [hold, in-hand], [flower, daisy]*). To account for semantic similarity despite lexical variation, we also use a state-of-the-art (SOTA) language embedding model (EmbeddingGemma [56]) to obtain description embeddings in a 768-dimensional space. We further compute the average pairwise embedding cosine similarities for each image and add this to our metrics. Figure 6 shows the distribution of the embedding similarities for all images in the sample subset. We note that the lemmatized descriptions show a slightly lower mean embedding similarity ($\bar{x} = 0.635, \sigma = 0.085$) compared to the unprocessed descriptions ($\bar{x} = 0.684, \sigma = 0.083$).

Figure 7 shows our computed metrics plotted against each other. We note in Table 6 that unprocessed description embedding similarity positively correlates strongly with statistical significance with the TF-IDF based similarity and Jaccard similarity, while it negatively correlates moderately with entropy. A statistically significant but very weak correlation also exists with perplexity. Similarly, lemmatized description embedding similarity positively correlates strongly with statistical

	Embeddings	Jaccard	TF-IDF	Entropy	KL-Div.
Jaccard	**0.729**				
TF/IDF	**0.664**	**0.87**			
Entropy	**-0.534**	**-0.717**	**-0.653**		
KL. Div.	-0.02	-0.179	-0.333	0.052	
Perplexity	-0.088	-0.115	0.176	**0.454**	-0.33

Table 6. Pearson correlation matrix for *unprocessed* description linguistic metrics. Bold highlights moderate or stronger statistically significant correlations ($\rho \geq 0.4$). All correlations statistically significant at the $p < 0.001$ level.

	Embeddings	Jaccard	TF-IDF	Entropy	KL-Div.
Jaccard	**0.807**				
TF/IDF	**0.628**	**0.717**			
Entropy	**-0.648**	**-0.747**	**-0.531**		
KL. Div.	-0.196	-0.269	-0.168	0.213	
Perplexity	-0.251	-0.248	-0.159	**0.752**	0.069

Table 7. Pearson correlation matrix for *lemmatized* description linguistic metrics. Bold highlights moderate or stronger statistically significant correlations ($\rho \geq 0.4$). All correlations statistically significant at the $p < 0.001$ level.

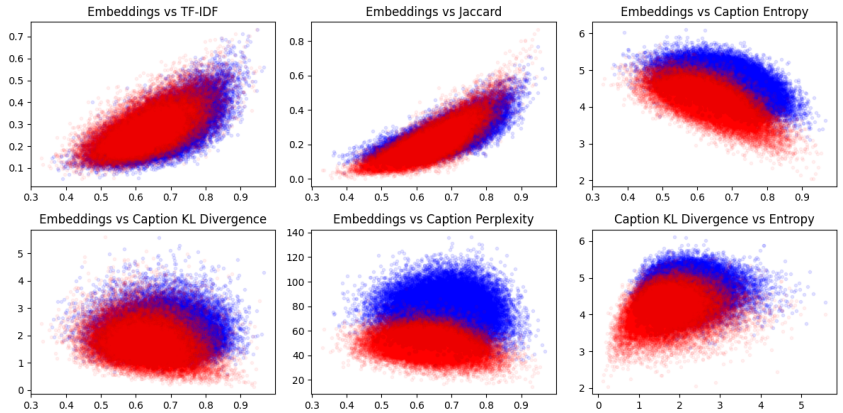


Fig. 7. Pairwise scatterplot comparisons of linguistic metrics.

significance with the TF-IDF based similarity and Jaccard similarity, while it negatively correlates strongly with entropy. Weak statistically significant correlations also exist with perplexity and relative entropy (see Table 7). A higher TF-IDF or Jaccard similarity score for an image indicates use of the same words across all of its five descriptions. It is reasonable thus that images with higher such scores would also exhibit a higher embedding similarity score. The phenomenon is also captured comparably well by the entropy metric, although embeddings demonstrate a better correlation with both TF-IDF and Jaccard scores, particularly for the lemmatized descriptions.

	Low	Mid	High
YOLO Objects	8.816 (6.152)	6.552 (5.638)	4.959 (4.637)
YOLO Saliency	39.694 (25.286)	35.756 (24.406)	33.563 (23.931)
Img. Saliency	26.189 (12.733)	22.722 (11.644)	20.443 (10.13)
Salient Regions	4.385 (1.494)	4.015 (1.512)	3.709 (1.45)

Table 8. Means and standard deviations of all visual metrics per category.

	Kruskal-Wallis ($\chi^2_{(2)}$)	**Mann-Whitney U (Z)**		
	-	Low-Mid	Low-High	Mid-High
YOLO Objects	2256.607	39501380.5	45103873.5	37437413
YOLO Saliency	285.027	34368545	36210941	33138731
Img. Saliency	1083.197	37329086	41013146	35233229.5
Salient Regions	837.089	36029457	39753676.5	35264069

Table 9. Statistical test results for visual metrics across congruence categories (all statistically significant at the $p < 0.001$ level).

B Selection of final 10 images

We plot the mean values of each of the visual metrics per congruence category, as shown in Fig. 8. We observe distinct differences across all metrics and categories (Table 8). Due to the non-normal distribution of all data series, as examined with Kolmogorov-Smirnov tests, we perform non-parametric tests to examine differences for statistical significance. Post-hoc Bonferroni correction is applied to all pairwise tests as required. We find that all differences are statistically significant as shown in Table 9, showing that visual metric means may be adequate to representatively characterise each category (once of course these categories have been defined based on linguistic metrics).

Therefore, to select representative images from each of the categories, we select, from each congruence category, those images that meet the following criteria: a) a number of salient regions equal to the rounded mean of the category; b) YOLO objects equal to the rounded mean of the category; c) image mean saliency within the 95%c.i. of the category mean, and; d) YOLO object mean saliency within the 95%c.i. of the category mean. Applying these criteria, we observe that none of the images actually match, for any of the categories. Therefore we relax criteria c) and d) to one standard deviation from the mean. The process yields 90 images for the low congruence category, 171 images for the medium category and 89 for high.

To extract the 10 suitable images from each congruence category, at first, we attempted random sampling, but noticed this resulted in multiple images from the same category, relating to the same type of scene (e.g. in a park, with dogs, etc.). In order to increase the diversity of the selected images from each congruence category, we perform a two-step process. First, we use a vision-enabled LLM to generate a description of each image. We use QWEN3-VL 8B Instruct (4-bit quantized model), a SOTA open vision-language model [63]. The instruction prompt we used, was largely based on the instructions given to the actual human workers in [19]. This was complemented by a custom system prompt to instruct the model to pay closer attention to the salient objects, persons or animals in the scene, since past research shows that Westerners (like the workers used to caption the Flickr30k set) have a different cognitive visual processing style compared to

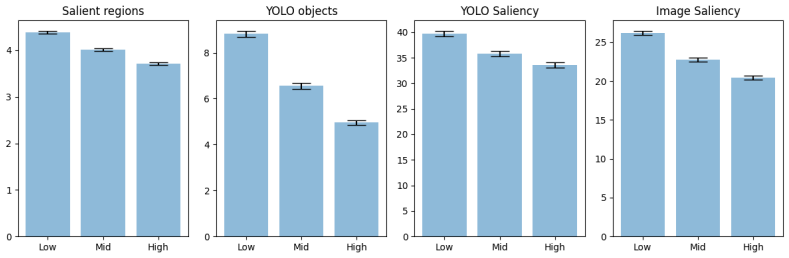


Fig. 8. Visual metric scores across description congruence categories. Error bars at 95%c.i.

Eastern cultures [7, 37, 46]. Table 10 demonstrates output of the model for a given image, ran five times. We observed empirically over multiple iterations and images, that the output is largely semantically similar, with small linguistic variations, thus one example was deemed sufficient to capture the image scene description. Subsequently, we compute the embeddings of derived descriptions using EmbeddingGemma. Finally, based on these embeddings, we perform a greedy farthest point sampling (GFPS) algorithm to obtain a set of 10 images, so that they are maximally dispersed in the 768-dimension embedding space [60]. Exact FPS is NP-hard, but in this case we are not interested in an exact solution. The approximation offered by greedy FPS is very fast: $O(K \cdot N \cdot D)$, where: N = Total number of points in the dataset, K = number of points to select (i.e. 10) and D = dimensionality of the points (i.e. 768). Still, since it depends on a random initial vector as a starting point, we employed a multi-start optimisation process with 1,000 random initializations. We selected the candidate set that maximized the objective function (mean pairwise cosine distance across the 10 selected images in each category), resulting in a 7.04% improvement over a single-pass execution. A manual visual inspection of the selected images confirmed that they were sufficiently distinct in terms of the scene they presented, therefore we proceeded with this selection.

C Distractor stimulus generation process

In the original paper by Friedman [15], the researchers replaced highly items in a given scene (e.g. office) with highly improbable items (distractors) from one of the other five scenes in their experiment (e.g. replace a car in a city scene, with a vegetable cart from a farm scene). As the types of scenes in our pipeline are not pre-determined, we took a different approach, leveraging LLM ability to imagine possible such unlikely onjects. To derive appropriate distractors, we took a two-pronged approach. First, we fed the image descriptions derived previously from the vision-enabled QWEN3-VL 8B model, to a text-to-text LLM (QWEN3-14b [63]) and asked it to first comprehend the scene, recommend a distractor, and provide an instruction to apply the modification to the original image. Next, we provided the original image as input to a multimodal LLM (Gemma-3-12b [53]) and asked it to provide the same outputs as QWEN3-14b. Finally, we provided the derived modification instructions from both processes to the diffusion model used for the previous transformations (FLUX.2-klein-9B), in order to obtain the distractor-modified stimuli. We then manually selected the distractor stimulus derived from either QWEN3 or Gemma-3 generated instructions that seemed more coherent, as some recommendations and the resulting stimuli from either model were not ideal.

Before proceeding to the online study, we manually inspected all original images to ensure they did not contain inappropriate or offensive content. We also manually inspected distractor images to ensure that the in-painted distractors were not inappropriate or offensive, and further that they did

Image	LLM-generated caption
	(1) Two young boys in light blue shirts and white caps play baseball on a grassy field while an adult in a light blue shirt kneels to coach them, and two men relax in chairs in the background. (2) Two young boys in baseball gear play on a grassy field while an adult in a blue shirt crouches nearby, and two men relax in chairs in the background. (3) Two young boys in baseball gear play on a grassy field while adults relax in chairs nearby. (4) Two young boys in light blue shirts and white caps play baseball on a grassy field while an adult in a blue shirt kneels to field the ball, and two men relax in chairs in the background. (5) A woman kneels on a grassy field, pointing toward two young boys playing baseball, while two men relax in chairs in the background.

Table 10

not occupy excessively large parts of the image or obscure important salient features. No changes to the original image selection, or distractor rendering were deemed necessary.

D Online Study

D.1 Study Description

The data collection process for this study consisted of a fully online procedure, performed on the participants' own smartphones. This introduces some variability in the results, but is an inherent concern with all crowdsourced tasks, which we believe is mitigated by the high participant number ($N = 100$, 50 female) and thorough randomisation of stimulus assignment and presentation, as will be explained later in this section. We developed an interactive web-based application to show participants the selected stimuli and collect their descriptions (Fig. 9). Images were shown for 10 seconds each and participants had to memorise the contents, forming an internal mental schema of what they saw, and then provide a description based on their recollection of that schema (Fig. 9a).

At the start of the experiment, participants' informed consent was obtained and a brief description of the experiment was presented to them. We then presented a series of five images as practice, each representing one of the stylistic variations. These images were selected randomly from the 79 unused images from the high congruence subset (see Section 3.3). For the first two images, we provided an example of descriptions to familiarise participants with the type of input expected from them (Fig. 9b). Then, for the remaining three, we did not provide any such hints, leaving participants to experiment with input as they wished.

For the main part of the experiment, each participant was presented with one of the five stimuli sets for each of the three congruence categories, in a balanced manner that ensured no participant saw the same image with a different transformation. As such, each participant provided a description for 3×10 stimuli and was not exposed to all five variations (see Table 11). Images were presented in congruence category blocks, for 10 seconds each, giving ample time for participants to move visual

content from gist memory into conceptual short-term memory [10, 40] and to form a structured visual representation [9]. We selected to hide the image during typing, in order to capture the internal schema formed by participants [8] without risking artificial inflation of the text due to prolonged exposure (i.e. submitting secondary non-salient details).

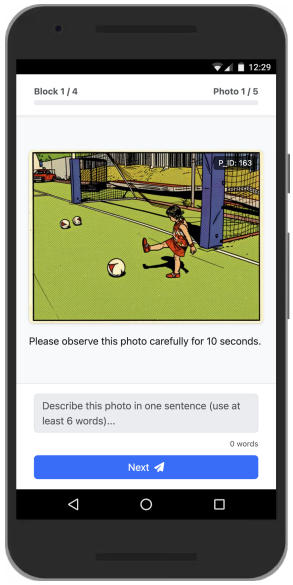
After giving participants time to internalise the scene, the image was hidden and participants could begin typing their description, based on a reconstructive recollection of the formed schema (Fig. 9c). Participants were allowed to take a break between blocks for as long as they wished (progress was automatically saved). At the end of each block, data was submitted to our server. Importantly, block order for each participant was randomised upon registration, and so was image order within each block, in order to mitigate fatigue effects. Pilot testing with 3 convenience-sampling sourced participants from our university demonstrated the experiment took approximately 20 minutes to complete, when carried out on a mobile device (smartphone).

Code ensured that the test could only be carried out with a mobile device. Further, ensure reasonable description lengths, we enforced a lower limit of six words per description and an upper limit of 50 words. To ensure description quality, upon entering, participant descriptions were sent to our server for consistency checking. There, a small Gemma-3-4b LLM (8-bit quantized) checked the participant caption against the actual image, and if a very low consistency or relevance was detected (e.g. entering gibberish, or a description that did not match the image at all) participants were asked to provide a more suitable description (Fig. 9d). The mechanism for this was to prompt the LLM to provide a numeric score (out of 5), rating how well the provided description matched the image, with a threshold for rejection $t \leq 2/5$ (see provided code in Section 7 for exact prompting instructions). If a participant failed to submit a suitable caption after three repeated failures (not enough words, or AI-detected low consistency), the application allowed the participant to repeat the trial showing the image again for a further 10 seconds. The experiment application environment collected submitted descriptions, their AI-detected consistency and timestamped keystroke events from the `onInput()` event handler of the input area. The captured data was saved locally and then sent to a remote MySQL database server at the end of every block.

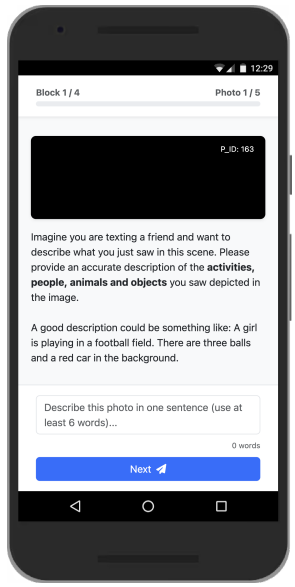
A significant dilemma faced during the experiment design phase was whether to restrict advanced text entry support (e.g. autocorrect, glide typing, voice typing). Vertanen & Kristensson [58] have argued that in crowdsourcing tasks, the added cognitive effort, associated with forcing the typing of text that does not necessarily match participants' own style or preference, is actually a threat to internal validity (referring here to transcription tasks). Consequently, we believe that forcing participants to forego the input support they are accustomed to, constitutes an additional cognitive burden that could artificially impact the core of our task (visual memory schema recall while typing). Taking into account that our goal is not to evaluate *how* people type with a particular input method, but rather *what they might* type given an image stimulus, we decided not to restrict input support. Further, since participants were randomly assigned to image conditions, the noise introduced by different input methods and advanced support use, should be distributed across all experimental conditions. Therefore, the differences we observe should not be attributable to the input methods, but rather to the stimulus itself.

D.2 Participant Recruitment

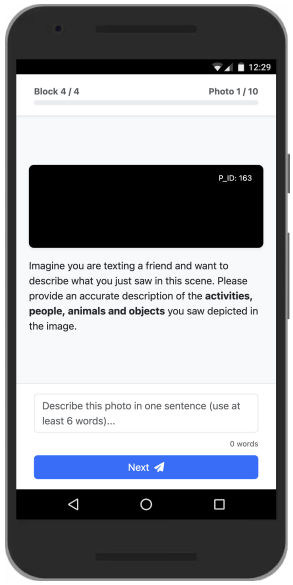
To ensure obtaining 20 descriptions for each image, a total of 100 participants were recruited through an online platform (Prolific) with the following criteria: Having English as their first and primary language, spending most of their time before turning 18 in an English-speaking country (Australia, Canada, Ireland, New Zealand, United Kingdom, United States), age 18-45, approval rate $> 90\%$, no language related or vision disorders and no colourblindness. The sample was stratified by gender (50% female).



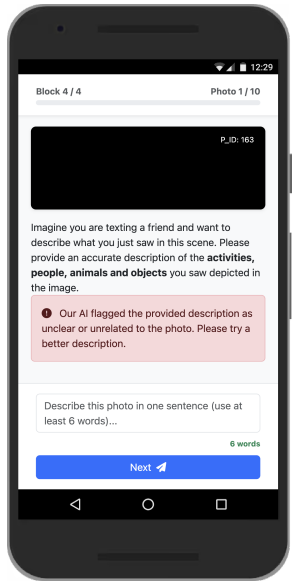
(a) Sample stimulus presentation



(b) Practice stimulus: image hidden, and instructions to participants, including hints on suitable descriptions.



(c) Standard stimulus: image hidden, and instructions to participants.



(d) Standard stimulus: feedback after AI-detected incoherent input.

Fig. 9. Online application experiment interface.

Congruence	ImageID	Image version 0	1	2	3	4
High	H1 ... H10	P1	P2	P3	P4	P5
Medium	M1 ... M10	P5	P1	P2	P3	P4
Low	L1 ... L10	P4	P5	P1	P2	P3

Table 11. Design of image assignment balancing for participants. Each participant is assigned 3 blocks of 10 images, ensuring the same image is not presented to them twice, in the form of any of its variations. With this assignment design, five participants are needed to obtain one caption for every image version. Block order and image order within blocks was randomised for every participant.

E Online Study Analysis

E.1 Participant demographics and study duration

Based on the recruitment platform data, the majority of our participants was born in the UK (68), followed by the USA (18), Canada (6), Australia (4), Hong-Kong, Nigeria and Ireland (1). Gender split was 50% female. Age mean was 34.394 ($\sigma = 7.010$). Mean completion time for the study was 2105.394 seconds (≈35 minutes) with a standard deviation of 3864.614 seconds (≈64 minutes). This is expected as we allowed participants to take breaks between blocks. The median time to complete the study was 1539 seconds (≈25 minutes). The majority of the 322,267 input events captured by the experiment application related to single keystrokes (66.66%). Character insertions accounted for 60.35% of the entire dataset, while deletions accounted for 6.17%. The remaining 33.4% consist of single or multi-character insertion events that suggest use of advanced input support methods (`insertCompositionText` 30.23%; `insertReplacementText` 1.04%; `insertText` 2.08%). Upon examination, these suggest some participants used voice typing or glide-typing, with word completions/suggestions (`insertCompositionText` and `insertText` events), while `insertReplacementText` refers to autocorrection events. In the following analyses, figures show statistics for the practice, as well as actual trial data. The latter are presented for reference only, and are excluded from statistical analyses.

E.2 Description lengths

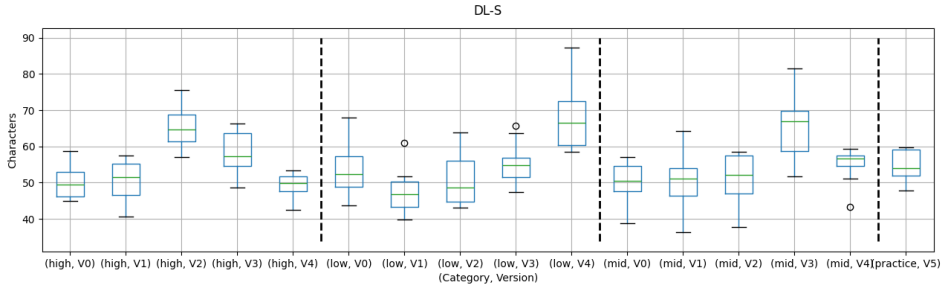
First we examine the length of descriptions submitted by participants, while applying the following modifications to the raw input: a) removal of all punctuation characters, and b) stopword removal (using the NLTK library list). We observe that mean length of captions submitted for all images was 80.395 characters ($\sigma = 15.607$), while excluding stopwords, this metric drops to 55.892 characters ($\sigma = 9.971$). The mean number of words per description is 16.054 ($\sigma = 3.230$), which again drops to 8.779 ($\sigma = 1.585$) with stopword removal. In line with previous research using composition tasks, this demonstrates that the image-based tasks are successful in eliciting longer texts from participants. Further, this text contains a significant percentage of non-stopword content (>50%). Treating these four metrics across the three congruency categories as sets of independent samples, we find no statistically significant differences (Kruskal-Wallis tests: Description Length (DL) $H_{(2)} = 0.485$, $p =$

0.785; Description Word Count (DWC) $H_{(2)} = 0.812, p = 0.666$; DL-stopwords removed (DL-S) $H_{(2)} = 0.022, p = 0.989$; DWC-stopwords removed (DWC-S) $H_{(2)} = 0.867, p = 0.6482$). Overall thus, participants entered comparable amounts of text across all image congruence categories.

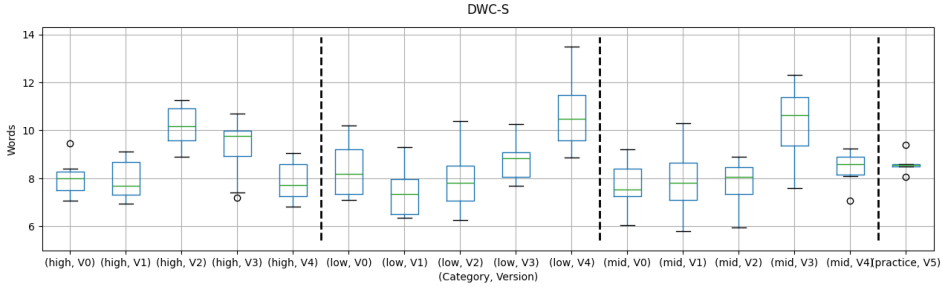
Given this result, it appears sensible to examine the effect of image variation within each of the congruence categories. In the rest of this section, we proceed with analyses using Linear Mixed-effects Models (LMMs) using descriptions with stopwords removed (i.e. only the *interesting* data). We use LMMs since our data is hierarchical (congruence category and image version) and also since participants did not see all five image versions, nor did they see multiple versions of the same image, therefore their contributions can be treated as independent samples. To account for the unbalanced experimental design (where participants viewed a subset of image variations) and to control for random variation inherent to specific images and participants, we fitted Linear Mixed-Effects Models (LMM) using the Restricted Maximum Likelihood (REML) method. Fixed effects were specified for Congruence Category (High, Medium, Low), Image Version (V0–V4), and their interaction. Random intercepts were included for Base Image ID to account for stimulus-specific variability. The LMM employs the High congruence and original image version (V4) as the reference category (baseline). We present the main findings in the section text, while full data pertaining to the LMM fitting can be found in the Appendix.

Starting with description length (DL-S, Fig. 10a, Table 12), the baseline condition (High Congruence, Original Photo V4) elicited an average non-stopword text length of 49.39 characters ($SE = 2.14$). Low Congruence images (V4) elicited significantly more text ($\beta = 19.11, z = 6.33, p < .001$) compared to the High Congruence baseline, but we did not find a statistically significant difference in Medium category images. We observed a distinct crossover interaction between image complexity and stylistic transformation. In High Congruence images, the Plakatstil (V0) and Pencil (V1) styles showed no significant difference from the original photo ($p > .05$). However, the Cartoon (V2) and Distractor (V3) variations significantly increased the amount of text entered by 15.83 ($p < .001$) and 9.03 ($p < .001$) characters, respectively. For Medium congruence images, only V2 (cartoon) affected text length with statistical significance ($\beta = -19.940, p < 0.001$). In Low Congruence images, this effect was sharply reversed. We observed significant negative interaction effects for all transformations, including Plakatstil (V0 $\beta = -15.550, p < .001$), Pencil (V1 $\beta = -22.185, p < .001$), Cartoon (V2 $\beta = -33.31, p < .001$) and Distractor (V3 $\beta = -22.20, p < .001$).

Next, we examine word count (DWC-S, Fig. 10b, Table 12) after stopword removal. The baseline condition (High,V4) elicited an average of 7.885 words per description. Low congruency images elicited more words ($\beta = 2.855, p < 0.001$), and there was no statistically significant difference for Medium congruency images. As with description length in characters, in High congruence images, the Plakatstil (V0) and Pencil (V1) styles showed no significant difference from the original photo ($p > .05$). However, the Cartoon (V2) and Distractor (V3) variations significantly increased the amount of text entered by 2.315 ($p < .001$) and 1.475 ($p < .001$) words respectively. For Medium congruence images, only V2 (cartoon) affected word count with statistical significance ($\beta = -3.020, p < 0.001$). In Low Congruence images, we observed significant negative interaction effects for all transformations, including Plakatstil (V0 $\beta = -2.475, p < .001$), Pencil (V1 $\beta = -3.380, p < .001$), Cartoon (V2 $\beta = -5.160, p < .001$) and Distractor (V3 $\beta = -3.505, p < .001$).



(a) Description length in characters (stopwords removed).



(b) Description word count (stopwords removed).

Fig. 10. Description length and word count metrics per congruence category and image version (V0=plakatstil, V1=pencil, V2=cartoon, V3=distractor, V4=original, V5=practice images).

Table 12. Mixed Linear Model Regressions fitting results: Description length (characters).

Model:	MixedLM	**Dependent Variable:**	Desc. Length
No. Observations:	150	**Method:**	REML
No. Groups:	30	**Scale:**	24.3417
Min. group size:	5	**Log-Likelihood:**	-446.9934
Max. group size:	5	**Converged:**	Yes
Mean group size:	5.0		

	Coef.	**Std.Err.**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	49.390	2.136	23.123	0.000	45.204	53.576
C(category)[T.low]	19.110	3.021	6.326	0.000	13.190	25.030
C(category)[T.mid]	5.535	3.021	1.832	0.067	-0.385	11.455
C(version)[T.V0]	0.820	2.206	0.372	0.710	-3.505	5.145
C(version)[T.V1]	1.110	2.206	0.503	0.615	-3.215	5.435
C(version)[T.V2]	15.830	2.206	7.174	0.000	11.505	20.155
C(version)[T.V3]	9.030	2.206	4.093	0.000	4.705	13.355
C(cat)[T.low]:C(ver)[T.V0]	-15.550	3.120	-4.983	0.000	-21.666	-9.434
C(cat)[T.mid]:C(ver)[T.V0]	-5.760	3.120	-1.846	0.065	-11.876	0.356
C(cat)[T.low]:C(ver)[T.V1]	-22.185	3.120	-7.110	0.000	-28.301	-16.069
C(cat)[T.mid]:C(ver)[T.V1]	-5.610	3.120	-1.798	0.072	-11.726	0.506
C(cat)[T.low]:C(ver)[T.V2]	-33.305	3.120	-10.673	0.000	-39.421	-27.189
C(cat)[T.mid]:C(ver)[T.V2]	-19.940	3.120	-6.390	0.000	-26.056	-13.824
C(cat)[T.low]:C(ver)[T.V3]	-22.200	3.120	-7.115	0.000	-28.316	-16.084
C(cat)[T.mid]:C(ver)[T.V3]	1.965	3.120	0.630	0.529	-4.151	8.081
Group Var	21.281	1.613				

Table 13. Mixed Linear Model Regressions fitting results: Description Length (Word Count)

Model:	MixedLM	**Dependent Variable:**	Word Count
No. Observations:	150	**Method:**	REML
No. Groups:	30	**Scale:**	0.6554
Min. group size:	5	**Log-Likelihood:**	-201.0574
Max. group size:	5	**Converged:**	Yes
Mean group size:	5.0		

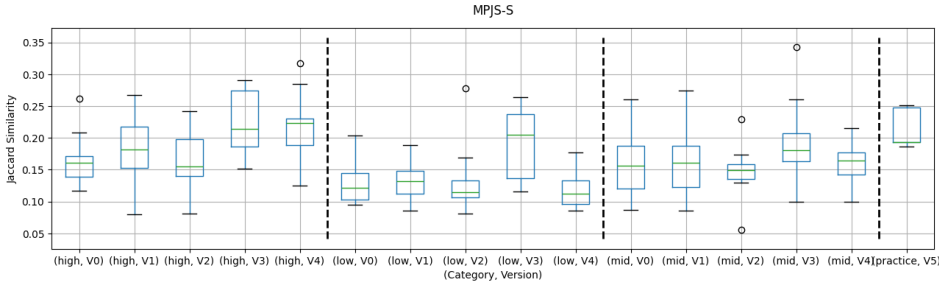
	Coef.	**Std.Err.**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	7.885	0.337	23.414	0.000	7.225	8.545
C(category)[T.low]	2.855	0.476	5.995	0.000	1.922	3.788
C(category)[T.mid]	0.610	0.476	1.281	0.200	-0.323	1.543
C(version)[T.V0]	0.090	0.362	0.249	0.804	-0.620	0.800
C(version)[T.V1]	0.040	0.362	0.110	0.912	-0.670	0.750
C(version)[T.V2]	2.315	0.362	6.394	0.000	1.605	3.025
C(version)[T.V3]	1.475	0.362	4.074	0.000	0.765	2.185
C(cat)[T.low]:C(ver)[T.V0]	-2.475	0.512	-4.834	0.000	-3.478	-1.472
C(cat)[T.mid]:C(ver)[T.V0]	-0.870	0.512	-1.699	0.089	-1.873	0.133
C(cat)[T.low]:C(ver)[T.V1]	-3.380	0.512	-6.602	0.000	-4.383	-2.377
C(cat)[T.mid]:C(ver)[T.V1]	-0.655	0.512	-1.279	0.201	-1.658	0.348
C(cat)[T.low]:C(ver)[T.V2]	-5.160	0.512	-10.078	0.000	-6.163	-4.157
C(cat)[T.mid]:C(ver)[T.V2]	-3.020	0.512	-5.898	0.000	-4.023	-2.017
C(cat)[T.low]:C(ver)[T.V3]	-3.505	0.512	-6.846	0.000	-4.508	-2.502
C(cat)[T.mid]:C(ver)[T.V3]	0.335	0.512	0.654	0.513	-0.668	1.338
Group Var	0.479	0.229				

E.3 Description Similarity

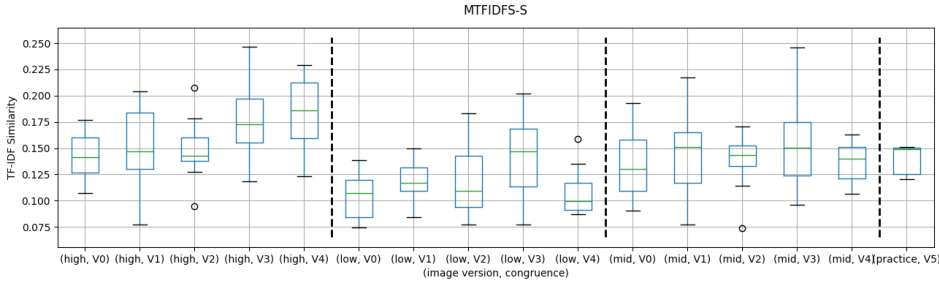
For description similarity, we employ the previously used Jaccard similarity, TF-IDF similarity and Embedding similarity measures, reporting again for descriptions after stopword removal. Starting with Jaccard similarity as the mean pairwise Jaccard similarity across the 20 captions for each image (MPJS-S, Fig. 11a, Table 14), the baseline condition (High,V4) elicited an average score of 0.218, which can be considered relatively low and indicates a diversity of words used to describe each image. Both Medium and Low congruency original images (V4) show even greater diversity ($\beta = -0.059, -0.097$, $p < .01$ respectively). In the High congruence images, all abstractive stylistic variations V0-V2 showed small but statistically significant decrease in similarity scores (V0 $\beta = -0.053$, $p < 0.01$, V1 $\beta = -0.034$, $p < 0.05$, V2 $\beta = -0.053$, $p < 0.01$), while the distractor variation had no effect. For the Medium category, we observe a small statistically significant effect only for V0 (Plakatstil, $\beta = 0.054$, $p < 0.05$). In the Low category, however, all stylistic variations had statistically significant effects on increasing the description similarity (V0 $\beta = 0.060$, $p < 0.05$, V1 $\beta = 0.047$, $p < 0.05$, V2 $\beta = 0.064$, $p < 0.01$, V3 $\beta = 0.064$, $p < 0.01$). However, the effect, as can be seen, is quite small.

Moving to the TF-IDF similarity measure (MTFIDFS-S, Fig. 11b, Table 15), the baseline condition (High,V4) elicited an average score of 0.184, which considered low and, just like the Jaccard metric, indicates a diversity of words used to describe each image. Both Medium and Low congruency original images (V4) show greater diversity compared to the baseline ($\beta = -0.047, -0.076$, $p < .01$ respectively). In the High congruence images, all abstractive stylistic variations V0-V2 showed small but statistically significant decrease in similarity scores (V0 $\beta = -0.043$, $p < 0.01$, V1 $\beta = -0.032$, $p < 0.01$, V2 $\beta = -0.036$, $p < 0.01$), while the distractor variation had no effect. For the Medium category, we observe a small statistically significant effect for all abstractive stylistic variations, this time increasing similarity (V0 $\beta = -0.041$, $p < 0.05$, V1 $\beta = -0.037$, $p < 0.05$, V2 $\beta = -0.037$, $p < 0.05$), again the distractor variation had no effect. In the Low category, however, all stylistic variations had statistically significant effects on increasing the description similarity (V0 $\beta = 0.039$, $p < 0.05$, V1 $\beta = 0.042$, $p < 0.05$, V2 $\beta = 0.046$, $p < 0.01$, V3 $\beta = 0.043$, $p < 0.05$). The effect, is again quite small.

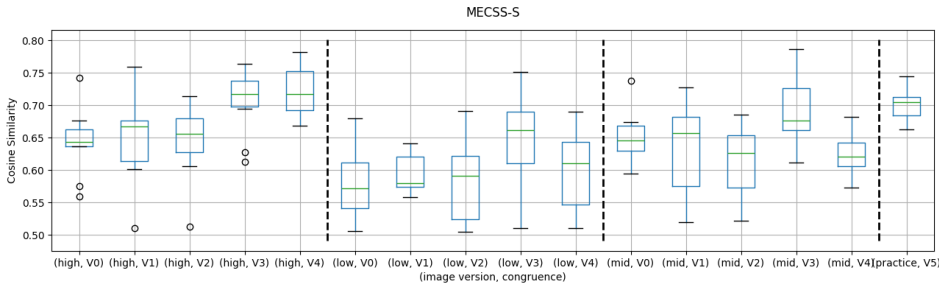
Finally, for the embedding cosine similarity measure (MECSS-S, Fig. 11c, Table 16), the baseline condition (High,V4) elicited an average score of 0.722, which can be considered quite high, matching the High category embedding similarities from the original Flickr30k dataset. Medium and Low congruency original images (V4) show a lower similarity compared to the baseline ($\beta = -0.095, -0.123$, $p < .01$ respectively). In the High congruence images, all abstractive stylistic variations V0-V2 showed small but statistically significant decrease in similarity scores (V0 $\beta = -0.080$, $p < 0.01$, V1 $\beta = -0.074$, $p < 0.01$, V2 $\beta = -0.075$, $p < 0.01$), while the distractor variation had no effect. For the Medium category, we observe statistically significant effects for all stylistic variations, this time increasing similarity (V0 $\beta = 0.102$, $p < 0.05$, V1 $\beta = 0.082$, $p < 0.01$, V2 $\beta = 0.062$, $p < 0.05$, V3 $\beta = 0.080$, $p < 0.01$), noting the large effect of variation V0 (plakatstil) that stands out from the rest. In the Low category, all stylistic variations also had statistically significant effects on increasing the description similarity (V0 $\beta = 0.058$, $p < 0.05$, V1 $\beta = 0.068$, $p < 0.01$, V2 $\beta = 0.057$, $p < 0.05$, V3 $\beta = 0.063$, $p < 0.05$). As with previous metrics, the effect is quite small.



(a) Mean pairwise Jaccard similarity (stopwords removed).



(b) Mean pairwise TF-IDF similarity (stopwords removed).



(c) Mean pairwise embedding cosine similarity (stopwords removed).

Fig. 11. Lexical and semantic similarity metrics per congruence category and image version (V0=plakatstil, V1=pencil, V2=cartoon, V3=distractor, V4=original, V5=practice images).

Table 14. Mixed Linear Model Regressions fitting results: Mean Pairwise Jaccard Similarity

Model:	MixedLM	**Dependent Variable:**	MPJS-S
No. Observations:	150	**Method:**	REML
No. Groups:	30	**Scale:**	0.0014
Min. group size:	5	**Log-Likelihood:**	213.1101
Max. group size:	5	**Converged:**	Yes
Mean group size:	5.0		

	Coef.	**Std.Err.**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.218	0.016	13.819	0.000	0.187	0.249
C(category)[T.low]	-0.097	0.022	-4.340	0.000	-0.140	-0.053
C(category)[T.mid]	-0.059	0.022	-2.640	0.008	-0.103	-0.015
C(version)[T.V0]	-0.053	0.017	-3.170	0.002	-0.086	-0.020
C(version)[T.V1]	-0.034	0.017	-2.008	0.045	-0.067	-0.001
C(version)[T.V2]	-0.053	0.017	-3.134	0.002	-0.085	-0.020
C(version)[T.V3]	0.006	0.017	0.367	0.713	-0.027	0.039
C(cat)[T.low]:C(ver)[T.V0]	0.060	0.024	2.532	0.011	0.014	0.107
C(cat)[T.mid]:C(ver)[T.V0]	0.054	0.024	2.278	0.023	0.008	0.101
C(cat)[T.low]:C(ver)[T.V1]	0.047	0.024	1.986	0.047	0.001	0.094
C(cat)[T.mid]:C(ver)[T.V1]	0.038	0.024	1.584	0.113	-0.009	0.084
C(cat)[T.low]:C(ver)[T.V2]	0.064	0.024	2.682	0.007	0.017	0.110
C(cat)[T.mid]:C(ver)[T.V2]	0.041	0.024	1.734	0.083	-0.005	0.088
C(cat)[T.low]:C(ver)[T.V3]	0.064	0.024	2.678	0.007	0.017	0.110
C(cat)[T.mid]:C(ver)[T.V3]	0.026	0.024	1.097	0.272	-0.020	0.073
Group Var	0.001	0.011				

Table 15. Mixed Linear Model Regressions fitting results: Mean Pairwise TF-IDF Similarity

Model:	MixedLM	**Dependent Variable:**	MTFIDFS-S
No. Observations:	150	**Method:**	REML
No. Groups:	30	**Scale:**	0.0007
Min. group size:	5	**Log-Likelihood:**	262.9864
Max. group size:	5	**Converged:**	Yes
Mean group size:	5.0		

	Coef.	**Std.Err.**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.184	0.010	17.945	0.000	0.164	0.204
C(category)[T.low]	-0.076	0.015	-5.234	0.000	-0.104	-0.048
C(category)[T.mid]	-0.047	0.015	-3.220	0.001	-0.075	-0.018
C(version)[T.V0]	-0.043	0.012	-3.544	0.000	-0.066	-0.019
C(version)[T.V1]	-0.032	0.012	-2.661	0.008	-0.056	-0.008
C(version)[T.V2]	-0.036	0.012	-2.988	0.003	-0.060	-0.012
C(version)[T.V3]	-0.009	0.012	-0.720	0.472	-0.032	0.015
C(cat)[T.low]:C(ver)[T.V0]	0.039	0.017	2.292	0.022	0.006	0.073
C(cat)[T.mid]:C(ver)[T.V0]	0.041	0.017	2.391	0.017	0.007	0.074
C(cat)[T.low]:C(ver)[T.V1]	0.042	0.017	2.462	0.014	0.009	0.076
C(cat)[T.mid]:C(ver)[T.V1]	0.037	0.017	2.139	0.032	0.003	0.070
C(cat)[T.low]:C(ver)[T.V2]	0.046	0.017	2.698	0.007	0.013	0.080
C(cat)[T.mid]:C(ver)[T.V2]	0.037	0.017	2.143	0.032	0.003	0.070
C(cat)[T.low]:C(ver)[T.V3]	0.043	0.017	2.529	0.011	0.010	0.077
C(cat)[T.mid]:C(ver)[T.V3]	0.024	0.017	1.415	0.157	-0.009	0.058
Group Var	0.000	0.005				

Table 16. Mixed Linear Model Regressions fitting results: Mean Pairwise Embedding Cosine Similarity

Model:	MixedLM	**Dependent Variable:**	MECSS-S
No. Observations:	150	**Method:**	REML
No. Groups:	30	**Scale:**	0.0017
Min. group size:	5	**Log-Likelihood:**	201.8065
Max. group size:	5	**Converged:**	Yes
Mean group size:	5.0		

	Coef.	**Std.Err.**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.722	0.017	41.958	0.000	0.689	0.756
C(category)[T.low]	-0.123	0.024	-5.056	0.000	-0.171	-0.075
C(category)[T.mid]	-0.095	0.024	-3.891	0.000	-0.142	-0.047
C(version)[T.V0]	-0.080	0.018	-4.369	0.000	-0.115	-0.044
C(version)[T.V1]	-0.074	0.018	-4.044	0.000	-0.109	-0.038
C(version)[T.V2]	-0.075	0.018	-4.098	0.000	-0.110	-0.039
C(version)[T.V3]	-0.017	0.018	-0.917	0.359	-0.052	0.019
C(cat)[T.low]:C(ver)[T.V0]	0.058	0.026	2.265	0.023	0.008	0.109
C(cat)[T.mid]:C(ver)[T.V0]	0.102	0.026	3.951	0.000	0.051	0.152
C(cat)[T.low]:C(ver)[T.V1]	0.068	0.026	2.652	0.008	0.018	0.119
C(cat)[T.mid]:C(ver)[T.V1]	0.082	0.026	3.187	0.001	0.032	0.132
C(cat)[T.low]:C(ver)[T.V2]	0.057	0.026	2.227	0.026	0.007	0.108
C(cat)[T.mid]:C(ver)[T.V2]	0.062	0.026	2.403	0.016	0.011	0.112
C(cat)[T.low]:C(ver)[T.V3]	0.063	0.026	2.439	0.015	0.012	0.113
C(cat)[T.mid]:C(ver)[T.V3]	0.080	0.026	3.108	0.002	0.030	0.130
Group Var	0.001	0.012				

E.4 Language representativeness

An important issue to reflect upon is the representativeness of the provided descriptions with regard to the English language. The classic transcription task phrase set by Mackenzie and Soukoreff [32] employed English character-level bigrams as a metric of language representativeness. Paek and Hsu [38] and Vertanen & Kristensson [59], use Kullback–Leibler divergence (KLD, or relative entropy) against a corpus in order to pick word-level bigrams in order to representatively sample a corpus for phrases. We employ here a hybrid method to assess the representativeness of text supplied by participants, using KLD on character-level bigrams, to compare the distributions of these bigrams between a reference corpus (Google Books nGrams Data, version 20120701), our dataset, and also the phrase sets by Mackenzie and Vertanen. We performed a non-parametric bootstrap (B=500, sample size=50) to estimate the 95% confidence intervals of the KL Divergence for each dataset relative to the Google Books corpus. Our dataset demonstrated divergence from the corpus distribution, which is comparable to that of the MacKenzie set and the Vertanen set, confirming adequate linguistic representativeness with little variation across image category and version (see Fig. 5). Interestingly, our datasets sits between the Mackenzie set, which is shown to closely adhere to formal written English norms, and the Vertanen set, which is oriented to reflect the high divergence typical of informal mobile communication. This intermediate score indicates that our dataset maintains a high degree of lexical diversity and grammatical structure (outperforming the Vertanen set in standard language coverage) while capturing more naturalistic variation than the strictly formal MacKenzie set. This suggests that the image tasks are successful in providing a balanced and adequate challenge for text entry evaluation.