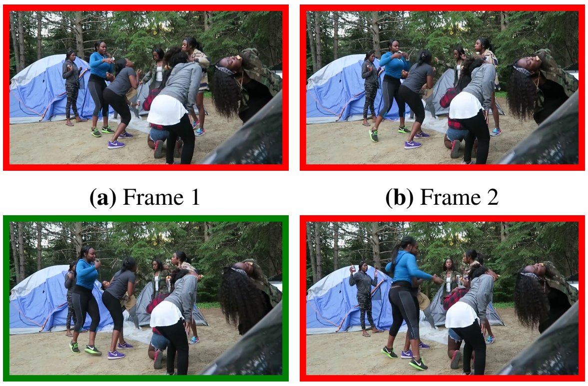
Image Retrieval from Contextual Descriptions

Anonymous ACL submission

Abstract

The ability to integrate context, including perceptual and temporal cues, plays a pivotal role in grounding the meaning of a linguistic utterance. In order to measure to what extent current vision-and-language models master this ability, we devise a new multimodal challenge, Image Retrieval from Contextual Descriptions (IMAGECODE). In particular, models are tasked with retrieving the correct image from a set of 10 minimally contrastive candidates based on a contextual description. As such, each description contains only the details that help distinguish between images. Because of this, descriptions tend to be complex in terms of syntax and discourse and require drawing pragmatic inferences. Images are sourced from both static pictures and video frames. We benchmark several state-of-the-art models, including both cross-encoders such as ViLBERT and bi-encoders such as CLIP, on IMAGECODE. Our results reveal that these models dramatically lag behind human performance: the best variant achieves an accuracy of 20.9 on video frames and 59.4 on static pictures, compared with 90.8 in humans. Furthermore, we experiment with new model variants that are better equipped to incorporate visual and temporal context into their representations, which achieve modest gains. Our hope is that IMAGECODE will foster progress in grounded language understanding by encouraging models to focus on fine-grained visual differences.



(a) Frame 1 **(b)** Frame 2

(c) Frame 3 **(d)** Frame 4

Figure 1: An example of the new challenge, Image Retrieval from Contextual Descriptions (IMAGECODE): *"The girl in blue is to the left of the girl in the middle with the purple shoes. The girl in blue is not obscured in any way."* Frames 5–10 are left out for simplicity's sake. The target image, frame 3, is in green, whereas the incorrect frames are in red.

1 Introduction

Natural languages are highly contextual (Fodor, 2001): for a listener, recovering the speaker's *intended* meaning requires integrating information from different streams, such as grounding in perception (Pecher and Zwaan, 2005), shared world knowledge, and temporal reasoning (Wilson and Sperber, 1998). These processes, more generally, fall under the umbrella term of *pragmatics* (Grice, 1957). Despite recent progress in multimodal systems, it remains unclear to which extent they can handle settings where context plays a major role, such as in real-world communication.

To this end, we present a new challenge that requires multimodal models to leverage context to retrieve images from text. In particular, given a contextual description and a set of minimally contrastive candidate images, i.e. differing only in some details, the model has to retrieve the target image. In order to discriminate between similar images, human annotators naturally produce highly nuanced and grammatically complex descriptions. An example of our new challenging dataset, Image Retrieval from Contextual Descriptions (IMAGECODE), is shown in Figure 1.

During the data collection process, sets of similar images are selected among static pictures from Open Images (Kuznetsova et al., 2020) and (a larger portion) among video frames from diverse domains. Including both types of images allows for diversifying the dataset while representing different degrees of visual similarity within each set. Next, we crowdsource a *contextual* description of a target image (presented together with the rest of the set) that contains only differences relevant for retrieval. After a filtering phase involving human retrievers, we obtain a large-scale dataset with 94,020 images and 21,202 descriptions associated with image sets of size 10.

As a result of this annotation protocol, successfully completing the task requires models to integrate several kinds of context: **i**) the image set, as the descriptions only make sense in the context of several other images and are not suitable as stand-alone captions. In fact, aspects of the image that are very salient and that therefore would normally be emphasized are not useful in our proposed task. Instead, the focus of our descriptions are fine-grained details that help discriminate between images (see Figure 1); **ii**) the speaker's intention. Due to their high degree of image similarity, contextual descriptions may be literally true for multiple images; however, once the speaker's intention is taken into account, the correct image can be determined by virtue of pragmatics (see Figure 2); **iii**) temporal sequences: for video frames temporal reasoning is also required to compare different moments of an unfolding event.

On our new dataset IMAGECODE, we benchmark a series of vision-and-language models that achieve state-of-the-art performance on other multimodal tasks, including both cross-encoders such as ViLBERT (Lu et al., 2019) and bi-encoders such as CLIP (Radford et al., 2021). We report several findings. First, accuracy on static images is vastly superior than on video frames. Therefore, the degree of similarity among the candidate images has an overwhelming impact on retrieval performance. Second, all state-of-the-art models generally struggle with image retrieval from contextual descriptions, whereas humans consistently achieve high accuracy.

Hence, we propose model variants capable of better taking context into account: **i**) once an image-description pair is encoded, we refine this representation by attending to the other images in the set; **ii**) we augment image encodings with special temporal embeddings. Based on our results, models take advantage of this additional information fruitfully but only to a limited degree.

Because of its challenging nature, due to the minimally contrastive images and complex descriptions, we believe that IMAGECODE will help make visio-linguistic models more context-aware and sensitive to fine-grained details. The dataset and models would be publicly released with the camera-ready version.

2 Related Work

There is a long tradition of grounding language understanding on single images, in the form of visual question answering (Goyal et al., 2017; Hudson and Manning, 2019), visual dialogue (de Vries et al., 2017; Das et al., 2017), or visual entailment (Xie et al., 2019). Recently, more and more focus has been directed to settings where the visual context consists of multiple images, either conventional static pictures (Vedantam et al., 2017; Hu et al., 2019; Suhr et al., 2019; Forbes et al., 2019; Hendricks and Nematzadeh, 2021; Yan et al., 2021; Hosseinzadeh and Wang, 2021; Bogin et al., 2021; Liu et al., 2021), or video frames (Jhamtani and Berg-Kirkpatrick, 2018a; Bansal et al., 2020). While many of these benchmarks involve just two images, COVR (Bogin et al., 2021) and ISVQA (Bansal et al., 2020) provide more images, similar to our sets of 10 images.

ISVQA and Spot-the-diff (Jhamtani and Berg-Kirkpatrick, 2018a) are most similar to our dataset, IMAGECODE. ISVQA is based on several video frames that are synthetic and cover a restricted domain, with short questions for Visual Question Answering. Spot-the-diff provides two frames from surveillance video cameras and descriptions of all their differences. IMAGECODE is unique as a) we cover a wider range of domains; b) we construct image sets that are maximally similar while being distinguishable through natural language (Section 3) and c) we limit descriptions to *relevant* differences. This results in (a) diverse, (b) complex and (c) pragmatically informative descriptions.

IMAGECODE elicits pragmatic reasoning (Andreas and Klein, 2016; Cohn-Gordon et al., 2018) as a listener has to consider the context and resolve ambiguities resulting from nuanced differences to solve the task.

3 Data Collection

Our data collection involves two steps with a human describer and retriever. The describer is given a set of 10 highly similar images $S = [I_1, I_2, ..., I_{10}]$, one of them marked as the target image I_t, and has to write a description D that clearly distinguishes I_t from the other distractor images. In the second step, the retriever is given the same 10 images and the description from the first step and has to identify the target image based on the description. S and D are only added to our dataset if the retrieval is successful.

Below, we outline the main stages of data collection: first, the collection of similar, contrastive images in Section 3.1. Then, the crowdsourcing of contextual descriptions in Section 3.2 and validation of the examples via image retrieval (Section 3.3). The final IMAGECODE dataset consists of 94,020 images (partitioned into 9,402 sets) and 21,202 contextual descriptions (16,594 in the train split, 2,302 and 2,306 in the validation and test split respectively).

3.1 Collecting Similar Images

In the first stage, we collect sets of images that are highly similar but still distinguishable from each other by a human. To quantitatively measure the pairwise similarity of two images, we compute the Euclidean distance between their encodings extracted from a pre-trained CLIP model (Radford et al., 2021).[1] To study the effect of different degrees of similarity, further variegate our dataset, and enable temporal reasoning, we source our candidate images from collections of static pictures as well as videos, as detailed below.

Static Pictures. We obtain image sets from one of the largest repositories of static pictures, the Open Images Dataset V6 (Kuznetsova et al., 2020), containing 1.74M images. For each image, we retrieve the 9 closest images from the training set based on their CLIP encodings. We then randomly sample 4,845 of these image sets.

Video Frames. As sources for our video frames, we use **i**) Video-Storytelling (Li et al., 2019), covering social events (wedding, birthday, Christmas, camping); **ii**) general-domain MSR-VTT (Xu et al., 2016); and **iii**) YouCook (Das et al., 2013), covering cooking events. We choose these datasets as they contain publicly available and general-purpose videos (not specific to downstream tasks). We retain the original splits for train, validation, and test.

To obtain disjoint sets of 10 similar frames, we first segment the videos into smaller scenes (also known as shots) via the scene detection functionality of `ffmpeg` (Tomar, 2006). Then, for each scene, we add its first frame to the set of selected images. We then iterate over every following frame and add it to the set if its pairwise Euclidean distance with each of the previously selected frames is larger than a threshold.[2] Once the set contains 10 images, we reiterate the procedure for a new set. If the scene ends and the current set contains less than 10 images, the set is discarded.

Dataset	After §3.1	After §3.3
MSR-VTT	11,643	8,045
Video-Storytelling	11,459	8,153
YouCook	894	588
Open Images	4,845	4,416

Table 1: Number of descriptions from each source of images at different stages of the annotation process.

During this process, we additionally remove frames that **i**) are too blurry, i.e. their BRISQUE score (Mittal et al., 2012) is larger than 0.65; or **ii**) contain too much text, which is detected with the OCR tool Tesseract (Smith, 2007).[3] We use all of YouCook's image sets and (due to cost constraints) randomly sample image sets from Video-Storytelling and MSR-VTT for crowdsourcing (cf. Table 1). We remark that image sets are further filtered at the final stage of annotation (Section 3.3).

3.2 Crowdsourcing Contextual Descriptions

After creating sets of highly-similar images in Section 3.1, we request annotators from Amazon Mechanical Turk (AMT) to write contextual descriptions for each target image in a set. Each round, a set of images is presented in random order for static pictures and respecting temporal order for video frames. This encourages annotators to take the dynamics of the event into account. We then (randomly) select 3 target images per set, and ask annotators to produce a description that discriminates them from the other images in the set. To encourage pragmatic reasoning, we do not ask for *all* the differences (just those sufficient for retrieval) and do not allow explicit mentions of other images (see Figure 2). We select high-quality annotators according to criteria in Appendix B and assign partly disjoint sets of annotators to train and test in order to avoid annotator bias (Geva et al., 2019).[4]

[1] We also experimented with ResNet-50 features, but we found CLIP results to be more similar to that of humans in preliminary experiments.

[2] The distance threshold was manually chosen as 0.35 based on qualitative results.

[3] The rationale of the second criterion is to prevent workers from focusing on the overlaid text rather than image content.

[4] For further details on crowdsourcing instructions, analysis of annotator bias and the AMT interface, please refer to

Phenomenon	all %	videos %	static %	Example from IMAGECODE	Definition
Context	47.3	**57.3**	6.6	Figure 2	Visual context or pragmatic inference required.
Temporal	15.0	**18.5**	4.1	*A smiling boy just **begins to** look towards the dog.*	Temporal markers (e.g., *after*) and verbs (e.g., *starts*)
Quantities	48.5	47.7	**51.0**	*There is an **equal amount** of yellow and white between **both** hands.*	—
Spatial Relations	70.5	**72.2**	65.3	*The cloud on **top left side** of box only has **half of** it showing.*	—
Negation	17.9	**20.7**	6.1	*The spoon is at the top right corner, it is **not** moving any of the food.*	—
Visibility / Occlusion	45.5	**54.5**	8.6	*The flowers the woman in the teal strapless dress is carrying are **completely obscured** by the man in the black shirt's head.*	An entity is covered or partially outside of the image.
Nuances	26.3	**31.6**	5.1	*There is the **slightest of openings** to see the end of the bridge through the obstruction.*	Description grounded on small patch of pixels or very non-salient aspects.
Co-reference	41.5	**42.4**	38.8	*The cloud on top left side of box only has half of **it** showing.*	—
Meta Properties	12.0	**13.9**	6.1	***Bright shot** of a girl and boy standing up straight. Her eyes are closed.*	Blurriness, brightness, overlays, and transitions of frames.

Table 2: Distribution of challenging phenomena in IMAGECODE based on 200 (or 1000 if underlined) manually annotated examples.

Metric	val	test
Human Accuracy	90.9	90.8
Krippendorff's α (nominal)	.797	.795
Krippendorff's α (interval)	.872	.869

Table 3: Human performance (accuracy) and inter-annotator agreement (Krippendorff's α) on the validation and test splits of IMAGECODE.

	ours	NLVR2	Spot-the-diff
Average length	23.3	15.3	10.6
Word types	6,916	6,602	2,282
Average tree depth	5.1	4.8	4.3
Average sentences	1.6	1.0	1.0

Table 4: Comparison of the text statistics of IMAGECODE with other vision-and-language datasets.

3.3 Human Validation via Image Retrieval

Finally, we validate the annotation crowdsourced in Section 3.2 by asking AMT workers to retrieve the correct target image from a set given its contextual description. For the final dataset, we retained only the examples that i) were retrieved successfully in the training set by a single worker or ii) were retrieved successfully by *at least* 2 out of 3 workers in the validation and test sets. As a consequence, we filtered out 26.5% of the contextual descriptions generated in Section 3.2. Table 1 compares the number of examples retained at each stage throughout the dataset creation.[5]

4 Data Analysis

4.1 Human Accuracy and Agreement

To quantify the reliability of the process outlined in Section 3, we report the inter-annotator agreement on our final dataset in Table 3. We use Krippendorff's α as a metric (the higher the better), which accounts for incomplete data, since the number of annotators per example is not fixed. We treat the index of the target image either as a nominal variable for static images or as an ordinal variable for video frames. In both cases, we find a high degree of agreement. Moreover, in Table 3, we also report human accuracy– the percentage of times an annotator retrieved the correct target image from a contextual description (as described in Section 3.3). This provides an upper ceiling for the model performances (see Section 6).

4.2 Language Statistics

In Table 4, we measure a series of statistics of the descriptions collected for IMAGECODE and compare them with other vision-and-language datasets with multiple naturalistic images (cf. Section 2), such as NLVR2 (Suhr et al., 2019) and Spot-the-diff (Jhamtani and Berg-Kirkpatrick, 2018b).[6] In particular, we count the average description length,

Appendix C and Appendix D.

[5]Again, the set of workers validating train and test sets were partly disjoint to avoid annotator bias.

[6]For comparability, we measured the statistics for all the datasets with the same tools.

(a) Frame 1 & Frame 2 **(b)** Frame 3 **(c)** Frame 4 & Frame 5

Figure 2: An example with description: "*No bridesmaid visible at all.*". Visual context is necessary to identify the correct target image, by cross-referencing the portions of images with bridesmaids (red boxes).

the number of distinct word types, the average dependency tree depth of each sentence,[7] and the average number of sentences per description. Based on these metrics, we find evidence that IMAGECODE's descriptions are longer and more syntactically complex than in the other datasets. Moreover, they include multiple sentences (11.8% of examples have 3 or more).

4.3 Vision Statistics

By calculating the average pairwise Euclidean distance between CLIP-based encodings of images in the same set, we find that video frames are more similar than static pictures – as expected – by a factor of 1.13. Moreover, we find that descriptions of video frames mention human body parts (72.1%) more often than static pictures (30.2%). On the other hand, names of colors appear in descriptions of static pictures (61.4%) more frequently than video frames (33.6%).[8] Thus, annotators resort to different strategies to discriminate between different types of image sets, focusing on the aspects that vary the most.

4.4 Challenging Phenomena

Finally, we identify 9 interesting and challenging phenomena in IMAGECODE and annotate whether they are present in 200 examples from the validation set. We provide the definition of each phenomenon, its frequency, and an illustrative example in Table 2. More information is given in Appendix F. For 4 of these phenomena unique to IMAGECODE, we further annotated 800 examples for the purpose of error analysis in Section 6. Inspecting these examples, we find a high number of cases where the visual context (47.0%) is required to complete the task. For instance, consider Figure 2: the description *"No bridesmaid visible at all."* requires a retriever to resolve the co-references of the entities in 5 frames. In particular, the body parts of the bridesmaids (red boxes) visible in frames 2 and 4 would not be identifiable as such without frame 1 and 5, respectively (where they appear with matching dresses and flowers in their hands).

Another group of phenomena characteristic for IMAGECODE originates from its minimally contrastive setup: annotators might focus on how an event unfolds over time (*temporal context*), on what is missing in a specific frame but visible in the others (*negation*), on what moved out of frame (*visibility / occlusion*), or on small regions and patches of pixels (*nuances*). Importantly, these phenomena are less prominent in static pictures than in video frames (cf. Table 2).

5 Methods

5.1 Baselines

In order to assess whether vision-and-language models can retrieve the correct image from a contextual description on a par with humans, we benchmark two state-of-the-art models that represent two main families of multimodal architectures (Bugliarello et al., 2021; Miech et al., 2021): **i**) ViLBERT, a cross-encoder where language and vision streams can interact via cross-attention at intermediate layers (Lu et al., 2019); **ii**) CLIP, a bi-encoder where language and vision streams are independent (Radford et al., 2021). It is worth noting that ViL-

[7] We use spaCy (Honnibal and Montani, 2017) as a parser.

[8] We calculated these percentages based on a list of 171 body parts in English collected by Tjuka (2021) and a list of colors in English from games4esl.com.

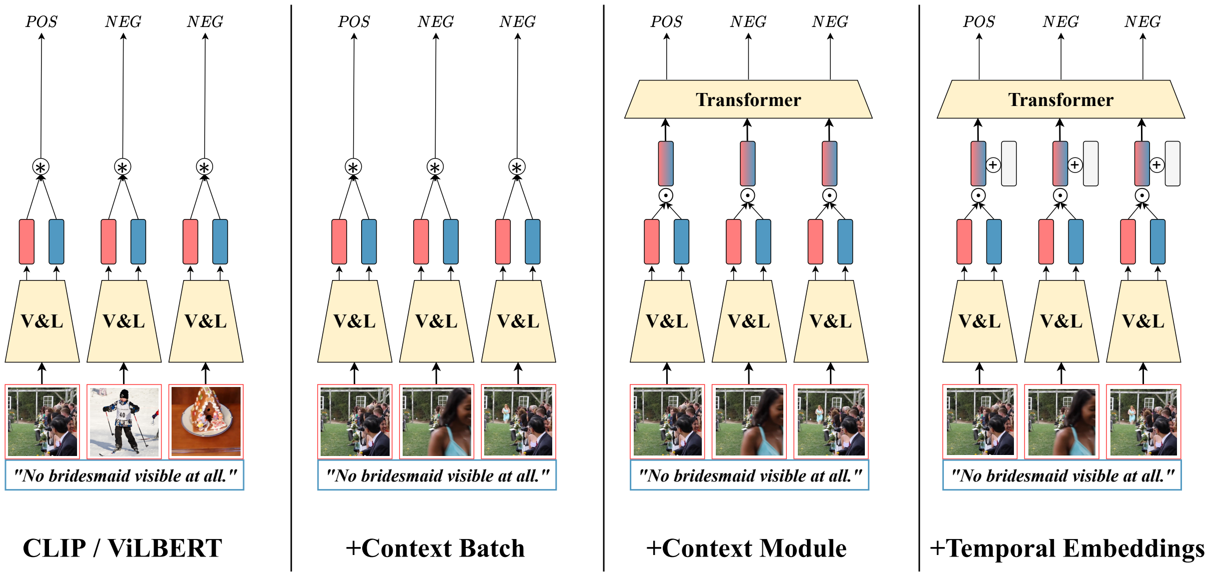


Figure 3: Models with increasing levels of context integration: see Section 5 for more details. In the figure, we colour visual embeddings in red, text embeddings in blue, and positional embeddings in grey. *POS* is the score for the target image and *NEG* for the other candidates. ⊛ represents dot product for CLIP and element-wise multiplication followed by a linear layer for ViLBERT. ⊙ represents element-wise multiplication. For ease of exposition, we show 3 images instead of 10.

BERT is more expressive due to its architecture, whereas CLIP boasts a higher parameter count and is pre-trained on a larger dataset.

We evaluate these models under two different regimes: i) *zero-shot* inference, where pre-trained models are deployed on the IMAGECODE test set directly; and ii) *fine-tuning*, where the models are refined on the full training set before evaluation. We cast the training objective as binary classification for ViLBERT and as 10-class classification for CLIP.[9] Crucially, in both cases, positive and negative examples during training are sampled at random independently from the image set they belong to (see the first column of Figure 3). Thus, the visual context of the other images in a set is only indirectly accessible at inference time, where the image with the highest probability is predicted.

5.2 Integrating Context into Vision-and-Language Models

For the fine-tuning regime, we further investigate some modifications in the training setup and model architecture that facilitate the integration of visual and temporal context into the model. First, we use an alternative objective where both CLIP and ViLBERT are trained on 10-class classification, but the 1 positive and 9 negatives are sourced from the same image set. The consequence of including positive and negative examples from the same image set in the same mini-batch is providing a wider visual context. We refer to this variant as +CONTEXTBATCH (second column of Figure 3).

This setup only conveys the visual context as a weak signal, since the model has no chance to directly compare the images in the same set. Hence, we experiment with enhancing the architecture of vision-and-language models with a mechanism inspired by Bogin et al. (2021). In particular, given an encoder (CLIP or ViLBERT), we obtain the representations of a contextual description $\mathbf{x}_L \in \mathbb{R}^e$ (where e is the model hidden size) and of the images in a set $(\mathbf{x}_V^{(1)}, \dots, \mathbf{x}_V^{(10)}), \mathbf{x}_V^{(i)} \in \mathbb{R}^e$ from their final layer.[10] Then, we create a series of multimodal embeddings via element-wise multiplication: $\mathbf{m} = (\mathbf{x}_L \odot \mathbf{x}_V^{(1)}, \dots, \mathbf{x}_L \odot \mathbf{x}_V^{(10)})$. Finally, we feed these to a l-layer Transformer $\mathrm{Tf} : \mathbb{R}^{10\times e} \to \mathbb{R}^{10\times e}$ to obtain context-aware multimodal embeddings $(\mathrm{Tf}(\mathbf{m})_1, \dots, \mathrm{Tf}(\mathbf{m})_{10})$. Since each description–image pair can now attend on the others in a set, the model can fully exploit the visual context. We obtain the score for the i-th pair through a linear classifier head $W \in \mathbb{R}^{1\times e}$. The target image is predicted as

$$\arg\max_i \ \mathrm{softmax}\left[W\left(\mathrm{Tf}(\mathbf{m})_i + \mathbf{m}^{(i)}\right)\right] \quad (1)$$

[9] We found this solution to work better for each model in practice, which is justified by their different pre-training objectives.

[10] We use the CLS tokens for ViLBERT.

Note that we add a highway layer from the input to the output of the Transformer. We label this model variant +CONTEXTMODULE.

Finally, in addition to visual context, we make models aware of the temporal context too, as shown in the fourth column of Figure 3. For video-based examples only, the multimodal embeddings of each description-image pair are summed with a learnable positional embedding $\mathbf{t} \in \mathbb{R}^e$ that reflects the temporal order of the frames.[11] Thus, $\mathbf{m} = (\mathbf{x}_L \odot \mathbf{x}_V^{(1)} \oplus \mathbf{t}^{(1)}, \ldots, \mathbf{x}_L \odot \mathbf{x}_V^{(10)} \oplus \mathbf{t}^{(10)})$. Multimodal embeddings are then fed to a Transformer as above. We label this variant encapsulating both visual and temporal context +TEMPORALEMBEDDINGS.

5.3 Experimental Setup

For all CLIP experiments, we use a pre-trained model with the vision backbone VIT-B/16.[12] We train the full models with a batch size of 360 examples (i.e., 36 image sets) for CLIP and 150 examples for ViLBERT. We perform early stopping based on the validation accuracy with a maximum of 30 epochs. In the variants that adopt the base version of a model, we select a learning rate of $4 \cdot 10^{-6}$ for CLIP, $5 \cdot 10^{-6}$ for ViLBERT, and $4 \cdot 10^{-5}$ for ViLBERT +CONTEXTBATCH. We find these values via hyper-parameter search on the range $[10^{-4}, 10^{-7}]$.

For all the model variants that modify the model architecture, we adopt the following setup: first, we fine-tune the full model in the +CONTEXTBATCH regime as detailed above. Afterwards, we freeze the encoder parameters and train the components responsible for processing the multimodal embeddings, described in Equation (1). More details are provided in Appendix E.

All descriptions in IMAGECODE exceeding the maximum length of CLIP and ViLBERT are truncated. Due to their negligible amount, this does not affect performance significantly.

6 Results

In Table 5, we report the performance of the models from Section 5 for all the examples in IMAGECODE as well as for the subsets containing only video frames or static pictures. Note that the random chance baseline has an accuracy of 10%. In what follows, we compare the results across several dimensions.

	all	**video**	**static**
ZERO-SHOT			
CLIP	22.4	15.6	47.8
FINE-TUNING			
CLIP	23.6	17.1	48.2
+CONTEXTBATCH	28.0	19.7	59.2
+CONTEXTMODULE	28.2	19.9	**59.4**
+TEMPORALEMBEDDINGS	**28.9**	**20.9**	58.8
ZERO-SHOT			
ViLBERT	19.3	13.5	40.8
FINE-TUNING			
ViLBERT	20.9	13.1	**49.9**
+CONTEXTBATCH	20.9	15.0	42.7
+CONTEXTMODULE	22.3	16.1	45.6
+TEMPORALEMBEDDINGS	**24.5**	**18.0**	49.3

Table 5: Performance (accuracy) on IMAGECODE across two training regimes (zero-shot and fine-tuning), two models (CLIP and ViLBERT) and 4 model variants. We report separate figures for all the examples and two disjoint subsets: video frames and static pictures.

Zero-shot vs. fine-tuning. In the zero-shot setting, we observe that CLIP representations are surprisingly superior to ViLBERT even though CLIP has separate streams to encode an image and its description. In the simplest fine-tuning setting (i.e., if negatives are randomly sampled independent of the image set), we find that there is only a small increase in performance for both CLIP and ViLBERT (+5.4% and +8.3%, respectively) compared to zero-shot inference. This demonstrates that in the regime where images in the same set do not appear in the same batch during training, models cannot extrapolate how to leverage the visual context at inference time.

Adding context. For the fine-tuning regime, we observe instead a different trend once the visual context of the other images in a set is provided during training (+CONTEXTBATCH): CLIP receives a significant boost in performance (+18.7%), which is particularly accentuated for static pictures. On the other hand, ViLBERT's performance remains the same, as this variant is beneficial for video frames but detrimental for static pictures. Stacking a special module for contextualizing multimodal representations on top of the encoders (+CONTEXTMODULE), instead, yields gains for ViLBERT compared to +CON-

[11] In the examples with static pictures, no temporal embedding is added.
[12] https://github.com/openai/CLIP

TEXTBATCH, whereas CLIP is almost unaffected. This shows that all models can exploit visual context, but different strategies (contrastive training or dedicated modules) may be necessary.

Finally, both CLIP and ViLBERT achieve the highest performance when fine-tuned with both visual and temporal context. Adding temporal positional embeddings on top of the contextual module (+TEMPORALEMBEDDINGS) yields an accuracy of 28.9 for CLIP and 24.5 for ViLBERT. Crucially, even the best-performing models lag significantly behind the (micro-averaged) human accuracy of 90.8 (cf. Table 3). Hence, despite some limited ability to integrate context, models are currently incapable of the fine-grained reasoning and pragmatic inferences needed to solve IMAGECODE.

Pre-trained model. Across all model variants and training regimes, CLIP consistently achieves higher accuracy than ViLBERT. This implies that a larger amount of parameters and pre-training examples are more beneficial than ViLBERT's more expressive model architecture. Thus, these results violate the expectations that ViLBERT's cross-attention would be more suitable to jointly encode highly nuanced visual details and descriptions (Miech et al., 2021).

Video frames vs. static pictures. The highest accuracy on the subset of the data with video frames (20.9) is far lower than that for static pictures (59.4). This confirms that videos represent the main challenge in IMAGECODE, both because of the higher similarity of images in a set and of the particular factors of variation that help differentiate among them (cf. Section 4.3 and examples in Appendix F). Additionally, model performance on video frames seems to increase more consistently as more context (both visual and temporal) is provided, whereas there is no clear trend in the case of static pictures.

Error Analysis. On a broad level, we have seen that video frames are much more challenging for models. Next, to identify more fine-grained causes for the overall low performance of the vision-and-language models on IMAGECODE, we compute the Pearson's correlation between accuracy and a series of possible explanatory variables. In particular, we find a weak negative correlation with the number of tokens in the description ($\rho = -0.11$) and a weak positive correlation with the average pairwise Euclidean distance between CLIP encodings of the images in a set ($\rho = 0.22$), which represents visual similarity.

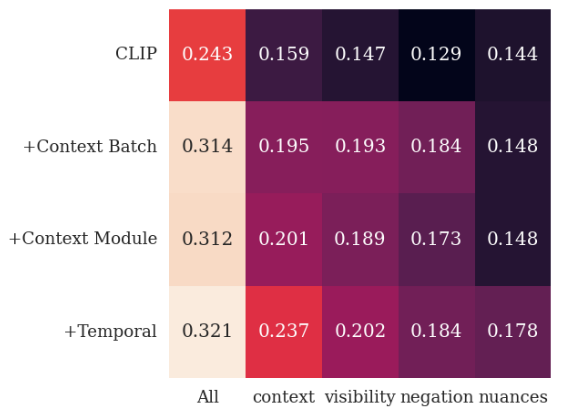


Figure 4: Performance of different CLIP variants (rows) on subsets of examples containing phenomena of interest (columns) in 1000 annotated validation examples. The hue of each cell indicates accuracy.

By focusing on the 1000 annotated examples in Table 2 we observe a stark drop from overall performance on the subset of examples containing nuances, visibility/occlusion, and negation (Figure 4). This confirms insights from Kassner and Schütze (2020) and Hosseini et al. (2021) on the difficulty of modeling negation in text-only models.

7 Conclusions and Future Work

We created a new challenge, Image Retrieval from Contextual Descriptions (IMAGECODE), which is designed to evaluate the ability of vision-and-language models to integrate visual, pragmatic, and temporal context into their predictions. In particular, given a complex and nuanced *contextual* description, a model is required to retrieve the corresponding image from a set of highly similar candidates. We benchmarked state-of-the-art bi-encoder and cross-encoder models, such as CLIP and ViLBERT. Moreover, we proposed new variants of these models that are more suitable to solve this task, by augmenting them with a module to attend on the other images in a set and temporal embeddings. We found that IMAGECODE is highly challenging for all variants: even the best model (28.9) lags behind human performance (90.8) dramatically. Images sourced from video frames display the largest gap in performance. The most challenging phenomena in IMAGECODE include pragmatics, negation, fine-grained distinctions between images, and occlusion among others.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about Pragmatics with Neural Listeners and Speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Ankan Bansal, Yuting Zhang, and Rama Chellappa. 2020. Visual question answering on image sets. In *ECCV*.

Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. 2021. COVR: A test-bed for visually grounded compositional generalization with real images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9824–9846, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jerry Fodor. 2001. Language, thought and compositionality. *Royal Institute of Philosophy Supplements*, 48:227–242.

Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. 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)*, pages 708–717, Hong Kong, China. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

H Paul Grice. 1957. Meaning." the philosophical review 66: 377-88. 1969. *Utterer's Meaning and Intentions." The Philosophical Review*, 78:147–77.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing Image-Language Transformers for Verb Understanding. *arXiv:2106.09141 [cs]*. ArXiv: 2106.09141.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Mehrdad Hosseinzadeh and Yang Wang. 2021. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2725–2734.

Hexiang Hu, Ishan Misra, and Laurens van der Maaten. 2019. Binary Image Selection (BISON): Interpretable Evaluation of Visual Grounding. *arXiv preprint arXiv:1901.06595*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018a. Learning to Describe Differences Between Pairs of Similar Images. *arXiv:1808.10584 [cs]*. ArXiv: 1808.10584.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018b. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, Brussels, Belgium. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural Information Processing Systems*, 32.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking Fast and Slow: Efficient Text-to-Visual Retrieval with Transformers. *arXiv:2103.16553 [cs]*. ArXiv: 2103.16553.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708.

Diane Pecher and Rolf A Zwaan. 2005. *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Annika Tjuka. 2021. A list of color, emotion, and human body part concepts.

Suramya Tomar. 2006. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-Aware Captions from Context-Agnostic Supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–1079, Honolulu, HI. IEEE.

Deirdre Wilson and Dan Sperber. 1998. Pragmatics and time. *Pragmatics and Beyond New Series*, pages 1–22.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

An Yan, Xin Wang, Tsu-Jui Fu, and William Yang Wang. 2021. L2C: Describing visual differences needs semantic understanding of individuals. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2315–2320, Online. Association for Computational Linguistics.

A Length Distribution of the Image Descriptions

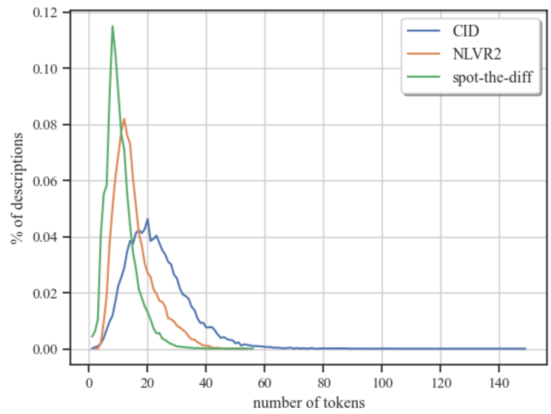


Figure 5: Distribution of the number of tokens across contextual descriptions in ImageCoDe.

B Criteria for Selecting Annotators

We keep data quality high through entry requirements (English speaking country, over 98% approval rate, etc.), qualification test, whitelisting workers and manually inspecting data. Most importantly our two-stage setup also allowed us to automate monitoring data quality as we could measure the description and retrieval accuracy of workers and only whitelisted those with high accuracy. We paid 0.25$ per description and 0.1$ per retrieval.

C Annotator Bias

The majority of descriptions in our test and validation split come from workers who did not work on the training set in order to avoid annotation bias. Our validation set contains 502 descriptions from workers "seen" from the training set and 1,800 description from "unseen" workers. In Table 6 we can see that models perform slightly better on seen workers across our CLIP model variants.

D Crowdsourcing Interface

Our AMT interface for the description task can be seen in Figure 6. The retriever interface looks conceptually similar, with a select-button for each image. Note that workers see images almost in almost half of full-screen (opposed to the shown examples in this PDF) and can quickly go back and forth between consecutive frames with arrow-keys, making it significantly easier to spot and compare nuanced changes.

	seen workers	**unseen workers**
FINE-TUNING		
CLIP	**23.9**	23.8
+ContextBatch	**34.5**	29.0
+ContextModule	**33.3**	29.2
+TemporalEmbeddings	**32.1**	30.8

Table 6: Performance (accuracy) on two subsets of the distinct validation split: seen workers (workers who also produced description on the train split) and unseen workers (who only worked on the test and validation data).

E Additional Hyper-parameters

The Transformer consists of 2 layers in CLIP variants and 4 layers in the ViLBERT variants, both employing gelu activation. The learning rate for the fine-tuning of the Transformer and linear heads is $2 \cdot 10^{-6}$ for the CLIP +ContextModule, 10^{-4} for CLIP +TemporalEmbeddings, and $2 \cdot 10^{-5}$ for both ViLBERT variants. We use the Volta-framework (Bugliarello et al., 2021) for the standardized ViLBERT model.

F Examples from ImageCoDe for all phenomena

For each phenomenon we provide 1 example and a definition we used for annotation purposes. Since most examples contain more than one phenomenon, some phenomena will be effectively showcased several times. Note that we picked examples that are relatively easy to understand and spot differences in.

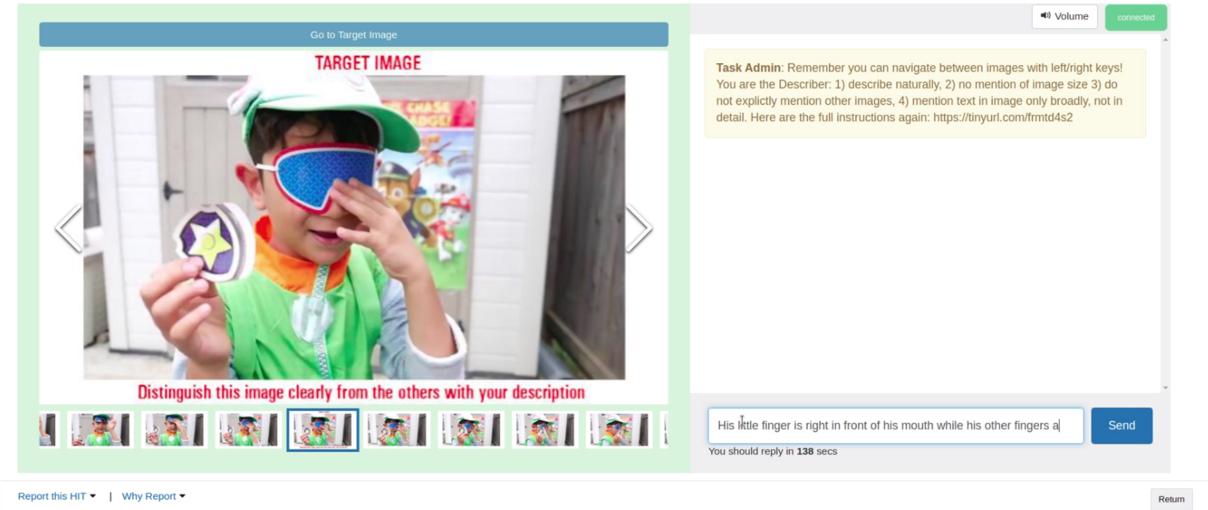
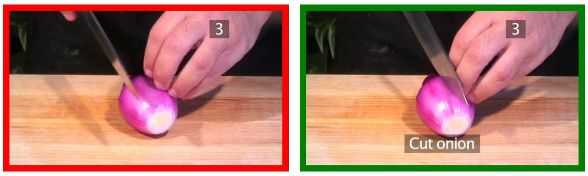


Figure 6: AMT interface for the describer task.

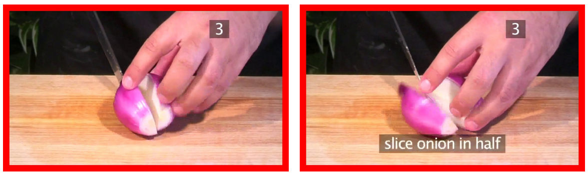
(a) Frame 6 **(b)** Frame 7

(c) Frame 8 **(d)** Frame 9

Figure 7: Example of **Context**: *"Both hands are on the piece of bread closest to the person."* Note: This is contextual since since without any context of other images, the description is also literally true for Frame 9. A model might even score it higher since the direct visual appearance is closer to typical bread. Definition: To understand the description, a listener has to consider other images and/or the speakers intention of describing only one of the images. In line with Grice's maxim of quality, a description is contextual if it is literally true for several images but we know it was intended for only one image. A description is also contextual if an objects cannot clearly be identified in the target image directly but only through cross-referencing other images.



(a) Frame 3 **(b)** Frame 4



(c) Frame 5 **(d)** Frame 6

Figure 8: Example of **negation**: *"The knife is most centrally placed to insert into the onion **without** having fully cut deeply into it yet."* Definition: Explicit linguistic negation ("not", "unseen", "non-") or negation quantifiers ("no person").

(a) Image 1

(b) Image 3

(c) Image 8

(d) Image 9

Figure 9: Example of **quantifiers/quantities**: *"A yellow **3 way** traffic light with a green arrow on the side facing closest to the camera"* Definition: We annotate for quantifiers (most, every, no, several,...) and absolute quantities ("five") as well as relative quantities (ratios like " a third of his hand").

(a) Frame 4

(b) Frame 5

(c) Frame 6

(d) Frame 7

Figure 10: Example of **spatial relations/reasoning**: *"The small girl **in front** is looking **directly to the right** with her **right hand on the side** of her face."* Definition: Any relations or adjectives regarding space. Examples: "in the top left corner", "left to the chair", but also camera perspective, or body orientation ("turned towards...")

(a) Frame 6

(b) Frame 7

(c) Frame 8

(d) Frame 9

Figure 11: Example of **temporality**: *" A smiling boy **just begins** to look towards the dog."* Definition: While most examples based on video frames implicitly require some temporal knowledge, we focus on explicit textual mentions of 1) temporal markers ("after", "during", "about to", etc) and 2) temporal verbs ("beginning to", "end to").

(a) Frame 7

(b) Frame 8

(c) Frame 9

(d) Frame 10

Figure 12: Example of **visibility/occlusion**: *" The tire is directly **on top of** the person's right shoe and you can **just barely see** fingers at the top. "* Definition: A description that mentions objects/people being occluded, (partially) out of frame, or in the process of leaving the frame.

(a) Frame 4

(b) Frame 5

(c) Frame 6

(d) Frame 7

Figure 13: Example of **nuances** (we marked small details with red/green rectangles): *" The person's palm is towards us and touching the left bottom corner of the cake. There is a* ***small amount of dark space between the right bottom corner of the photo and the edge of the cake****. "* Definition: Minor details, that are either a) not salient at all and would usually be left unmentioned and/or b) language reference is grounded on a small patch of pixels. Note that this phenomena is often linked with very minimally contrastive images.

(a) Image 3

(b) Image 5

(c) Image 6

(d) Image 10

Figure 14: Example of **coreference**: *" A woman with a white background smiles at the camera. Most of* ***her*** *body is visible.* ***She*** *is wearing a black outfit. "* Definition: Linguistic coreference.

(a) Frame 7

(b) Frame 8

(c) Frame 9

(d) Frame 10

Figure 15: Example of **meta properties**: *" The cucumber is just to be cut into, you can see a* ***transparent image covering the image****."* Definition: Descriptions that mention aspects that stem from the way the photo/video was taken: two overlayed images (when a video transitions), black-and-white, blurriness, brightness.