Multimodal Language Modelling across Languages and Cultures: Grounding Strategies for Concepts and Events

Suchir Salhan , Fangyu Liu and **Nigel Collier**
Language Technology Lab, TAL, University of Cambridge, UK
{sas245,fl399,nhc30}@cam.ac.uk

Abstract

Recent advances in multimodal language modelling have seen state-of-the-art performance in downstream vision-language tasks achieved by models that employ contrastive semantic pre-training. While grounding linguistic embeddings is typically assumed to improve the quality of natural language representations, we undertake an intrinsic semantic evaluation of multimodal representations obtained in contrastive visual pretraining in CLIP (Radford et al., 2021) and its video-text equivalent VideoCLIP (Xu et al., 2021). The effects of image and video grounding on concrete and abstract nominal concepts and verbal events are compared to unimodal BERT (Devlin et al., 2019) and Mirror-BERT (Liu et al., 2021) baselines. We focus on two case studies, verbal telicity and colour, to explore the fine-grained effects of image and video pretraining on nominal and verbal representations. The typological generalisability of our monolingual results are subsequently explored by evaluating the performance of Italian CLIP (Bianchi et al., 2021) and multilingual CLIP (Carlsson et al., 2022). Our findings are interpreted in the context of psycholinguistic and semantic research on verb grounding.

1 Introduction

Grounded Language Modelling aims to associate language with the world, for instance through the embodiment of human sensory perception and motor control (Emerson 2020, Harnad 1990). Lexical categories emerge as learners harness the sensory cues and physical heuristics of a given communicative context. Dyadic, interpersonal communication allow learners to establish how acquired words are connected to multimodal categories. Communication between agents is a necessary preliminary to the emergence of grounded conceptual representations. Chandu et al (2021) suggest a *dynamic* grounding strategy is necessary to enable language models to establish a 'common ground', a shared space of multimodal representations, with humans through a series of interactions and clarifications. Language grounding is facilitated by embodied, situated action taking that allow learners to simulate actions to learn about the world that can be tested and modified through communication (Bisk et al 2020).

While computational linguists continue to develop new grounded language models that achieve state-of-the-art performance in downstream multimodal tasks, there has been criticism by cognitive scientists and philosophers about whether the grounding strategies employed in language modelling are faithful to the principles of embodied cognition. McKenzie (2022) motivates a conceptual distinction between 'grounding', where concepts are connected to those perceived as ontologically prior, and 'big-G' grounding – a more abstract and generic notion of fundamentality. This distinction motivates the idea of a multidimensional **grounding category space** that represents (1) the relative groundedness of **words in the same semantic category**, and (2) the relative groundedness of **different semantic categories** as an alternative to the linear gradient of concreteness scores ranging from concrete to abstract typically employed in intrinsic semantic evaluation. Manzotti et al (2019) further argues that embodied AI continues to face a number of ontological challenges, such as circularity. Embodiment requires representations of actions that are simulated by the body but are distinct from the physical object itself, relying on notions of self-individuation and sense making that depend on a 'recognition of self' by the agent – rather than the simulation of physics and physical trajectories. This suggests that recent efforts on attempting to induce 'verb physics' in vision-language models may be misplaced. From this perspective, embodiment may or may not improve lead to performance improvements in AI systems, but it does not constitute a necessary part of language modelling. Man-

zotti (xxxx?) suggests that the theoretical problems facing the implementation of the principles of embodied cognition in AI systems offers supports for a radical realist theory of cognition, rather than for the theory of embodied cognition. Under this alternative view, meaning representations are formed through an identity relation between an external physical object and the agent's experience. For example, Beaton (2016) argues that:

> when I see an apple, for instance, my experience is directly of that apple itself, with no intervening mental image or representation.

From this criticism, we can formulate an **anti-embodiment hypothesis for language modelling**, as follows:

The Anti-Embodiment Hypothesis for Language Modelling: The embodiment and grounding strategies utilised in state-of-the-art multimodal language models do not necessarily provide a consistent and universally beneficial improvement to the lexical semantic representation of concepts and events.

The Anti-Embodiment Hypothesis for Language Modelling raises the question about whether the grounding and embodiment strategies employed computational linguists are necessary and beneficial, and whether alternative cognitive frameworks can be utilised to greater better performing language models.

In order to empirically evaluate this hypothesis, we evaluate the performance of a state-of-the-art vision-language model— the CLIP ("Contrastive Language Image Pretraining") image classification model (Radford et al. 2021)— and its multilingual and video-language variants. This allows us to assess the the effect of different grounding strategies (both video and image pre-training) on concept and event representations, and evaluate whether grounding and embodiment strategies provide a typologically uniform modelling advantage. We compare the performance of CLIP and its associated variants to unimodal baselines BERT (Devlin et al 2019) and Mirror-BERT (Liu et al 2021), which turns MLMs into effective lexical and sentence encoders by solely relying simply on self-supervision.

In particular, we aim to establish the fine-grained lexical semantic capabilities that grounding and embodiment strategies effect. As noted by Emerson (2020):

> Grounding is hard, ... some semantic constructions [...] are much harder for grounded language models to learn than others.

Notably, intrinsic semantic evaluation of multimodal language models has established that there is a consistently worse performance on understanding verbal events compared to nominal concepts (Beiborn et al 2018). In order to establish why this is the case, we evaluate the performance of multilingual language models with respect to human benchmarks established through psycholinguistic studies that focus on human multimodal perception of concrete and abstract nominals and verbs.

We further evaluate the performance of different grounding strategies in two case studies: (1) **colour**, which is an example of a grounded conceptual space and (2) **verbal telicity**, which establishes whether the action semantics of the verbal event has or does not have a definitive end-point. We explore these two case studies using multilingual grounded language models to assess the cross-linguistic evidence for and against the anti-embodiment hypothesis for language modelling.

2 Background

2.1 Grounding Nouns and Verbs: The Effect of Abstractness/Concreteness

Previous studies evaluating the ability of multimodal language language models in grounding nouns and verbs have established that **multimodal representations of verbal events are of a lower quality than nominal concepts**. Beinborn et al (2018) undertake an initial analysis of verb representations evaluated on the imSitu dataset (Yatskar et al 2016), which consists of images depicting verbal events with annotations that describe how the verbal arguments are linked to visual referents. They compare this to representations from Glove (Pennington et al 2014) as a unimodal baseline. Their findings are summarised in *Figure 1*, where it is illustrated that more highly embodied verbal actions, such as the verb pair *fall-dive*, yield a higher correlation coefficient ρ than verb pairs with a lower embodiment, like *know-decide*. Several

intrinsic semantic evaluation datasets, like SimLex-999 (Hill et al 2014), contain concreteness scores for verb pairs.

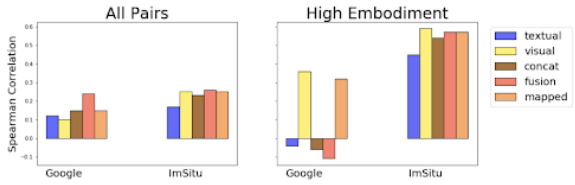

Figure 1: Spearman correlation between the cosine similarity of the embeddings representation of a verb pair and the corresponding similarity rating from the SimVerb dataset (Gerz et al 2016). Figure taken from Beiborn et al (2018)

We hope to ascertain *why* the grounding of verbal actions and events pose challenges for state-of-the-art multimodal language models, compared to the performance of these models in learning nominal concepts.

There have, however, been several psycholinguistic studies examining the effects of grounding nouns and verbs at differing levels of concreteness and abstractness. Moseley & Pulvermuller (2014:28-42) localise the neural activity of concrete and abstract nouns and verbs: concrete nouns and verbs elicit different brain signatures in the frontocentral cortex: concrete verbs activate the motor and premotor cortex more strongly, while concrete nouns activate inferior frontal areas. Abstract nouns and verbs elicit different brain signatures in the frontocentral cortex. This suggests that the brain activation patterns to words belonging to different lexical categories has a crucial dependence on the concreteness/abstractness of lexical items. Muraki et al (2020) establish neural evidence that suggests that there is a heterogeneous representation of abstract verbs in the brain. The categorisation of abstract verbs appears to be dependent on the modality of the associated experience, whether it represents a mental state (e.g *accept*), an emotional state (e.g *accuse*), or a non-bodily state (e.g *aid*). They find that there is only a clear difference in the neural activity of heavily disembodied abstract verbs, compared to concrete verbs.

There has recently been work using different grounding strategies for verbs in multimodal language models. The psycholinguistic evidence on the processing of concrete verbs has prompted Ebert et al (2022) to investigate the role of trajectories in the induction of concrete verb semantics in multimodal language models. They hypothesise that representation learning in a three-dimensional visuo-spatial world without language supervision allows vision-language models to learn more nuanced conceptual distinctions between concrete verb pairs, like *throw-toss*. Ding et al (2021) propose a unified framework for jointly learning visual concepts and infer 'verb physics' from video-language models. This raises the question about whether video and image pretraining leads to a significant difference in the quality of verb representations in multimodal language models.

2.2 Video and Image Grounding

As noted by Chen et al (2019), grounding language models in images has been popular in NLP over the past decade, and there has been a recent emergence of grounding in videos. Video grounding aims to identify the temporal boundaries `(start, end)` for a given moment of interest. Cao et al (2021) delineate two main types of video-language models: (1) top-down models that generate a set of moment proposals and select the best matching one, and (2) bottom-up models that directly regress the temporal boundaries of the referential segment from each frame. Ross et al (2018) attempt to train a semantic parser using captioned videos, which can turn sentences into logical forms using a much wider range of training data.

Yun et al (2021) evaluate whether vision-langauge pretraiing leads to an improvement in the quality of lexical semantic representations compared to unimodal language models. They find, upon comparing the performance of multimodal (VisualBERT, VideoBERT) and unimodal BERT across a range of clustering and probing tasks, that there was not a significant improvement in the quality of representations obtained through vision-language pretraining. Their intrinsic evaluation of multimodal and unimodal representations is primarily limited to nominal semantics, focusing on tasks like adjective-noun composition and semantic roles. We hope to extend the scope of their intrinsic evaluation of multimodal and unimodal language models to consider both verb and noun semantics, and focusing on a different set of state-of-the-art language models that utilise contrastive visual semantic pretraining.

2.3 Contrastive Visual Semantic Pretraining

Contrastive Language-Image Pretraining (Radford et al., 2021) learns visual representations from natural language supervision. CLIP consists of a visual

encoder V, either ResNet (He et al., 2016) or ViT (Dosovitskiy et al., 2020) with a text encoder T, like a Transformer (Vaswani et al., 2017). CLIP encodes the dot product between their outputs, which is used as an alignment score, as illustrated in *Figure 2*. CLIP is a zero-shot multimodal image classifier which adapts the GPT-2 architecture to encode image captions.

CLIP is pretrained to distinguish aligned image-text pairs from randomly combined ones by a contrastive loss. Instead of training on vision benchmarks, CLIP leverages abundant language supervisions from 400 million web-crawled image-text pairs and can conduct a variety of image classification tasks without specific optimising.

CLIP projects the encoded images and captions into a joint embedding space, where the model maximizes the cosine similarity of the correct image-caption pair while minimizing the cosine similarity of each caption with every other image in the batch (Radford et al., 2021). CLIP projects only a representation of the entire caption into the joint language-image space, and uses CWEs in order to produce this representation.

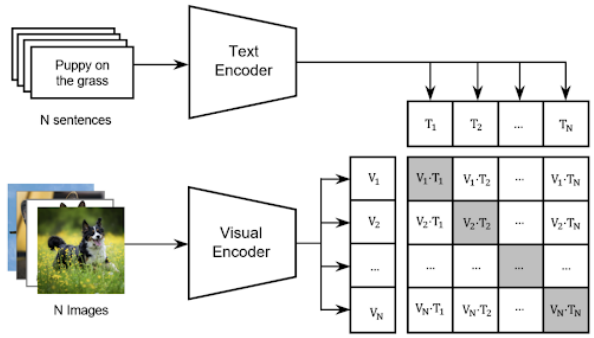

Figure 2: CLIP Architecture. Figure taken from Song (2022)

Khandewal et al (2022) investigate the effectiveness of CLIP encoders for a range of Embodied AI tasks, which involve agents that learn to navigate and interact with their environments.

Wolfe & Caliskan (2022) examine the effects of contrastive visual semantic pretraining employed in CLIP, compared to its unimodal equivalent GPT-2. They find that CLIP word embeddings outperform GPT-2 on wordlevel semantic intrinsic evaluation tasks, and achieve a new corpus-based state of the art for the RG65 evaluation, at .88, and that CLIP also forms fine-grained semantic representations of sentences, and obtains Spearman's $\rho = .73$ on the SemEval-2017 Semantic Textual Similarity Benchmark with no fine-tuning, compared to no greater than $\rho = .45$ in any layer of GPT-2.

Xu et al (2021) introduce VideoCLIP, a contrastive model for pretraining a unified video-text representation for zero-shot video and text understanding. This is captured by Transformer model parameters θ_v and θ_t for video and text. VideoCLIP contrasts temporally overlapping positive video-text pairs with negative video-text pairs obtained using nearest neighbour retrieval. The overlapping video-text pair is built by sampling a text clip, sampling a timestamp within the boundary of the textclip as the centre for video clip, and grow a video clip with a random duration from the centre timestamp. The correspondence between video and text is learned using a contrastive loss objective given in (1) below:

$$NCE(z_v, z_t) = \frac{e^(z_v \cdot \frac{z_t^+}{\tau})}{\sum_{z \in \{z_t^+, z_t^-\}} e^(z_v \cdot \frac{z}{\tau})} \quad (1)$$

where $NCE(z_v, z_t)$ is the contrastive loss on text-to-video similarity, τ is a temparature hyperparameter, z_t^+ is the positive embedded text clips that overlap with the video clip embedding z_v and z_t^- are the negative embedded text clips that are implicitly formed by other text clips in the training batch.

Bianchi et al (2021) introduce Contrastive Language-Image Pre-training for the Italian Langauge (CLIP-Italian), which is trained on a corpus of 1.4 million image-text pairs. CLIP-Italian uses the Italian BERT model [1] as a text encoder and uses pretrained checkpoint of `clip-vit-base-patch32` provided by OpenAI. [2].

Carlsson et al (2022) introduce the Multilingual CLIP Model, which exploits the modularisation of the CLIP architecture by using cross-lingual teacher learning to retrain the text-encoder for non-English languages. The Multilingual CLIP model rests on the assumption that the image-text pretraining in CLIP has allowed to production of similar multimodal embeddings for a matching text-image pair. This method consequently does not require any image data, relying entirely on machine translation for the CLIP text encoder. The teacher learning paradigm, introduce by Hinton et al (2015),

[1] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[2] https://huggingface.co/openai/clip-vit-base-patch32

is a domain-agnostic machine learning technique that allows the knowledge of a pretrained 'teacher' model, in this case CLIP trained on English image-text data, is transferred to another 'student' model that is pretrained in a different language. This removes the need for image training data in the target language, which allows the model to be utilised for low-resource languages. The cross-lingual teacher learning paradigm is trained using by minimising the Mean Squared Error (MSE) between the embeddings of the teacher and student model. The set of translations X are translated into the target language, creating a caption set X^* While the teacher model encodes X as embeddings E_T, the student model encodes X^* as embeddings E_S. The loss between X and X^* using Mean Squared Error (MSE), as in (2):

$$Loss = MSE(E_T, E_S) \tag{2}$$

3 Intrinsic Semantic Evaluation

In order to assess the evidence for and against The Anti-Embodiment Hypothesis, we evaluate the performance of CLIP and Video-CLIP multimodal vision-language models against intrinsic lexical semantic benchmarks. We compare the performance of these multimodal models to a BERT and Mirror-BERT baseline.

3.1 Models

In the intrinsic evaluation, we use the open-source pretrained VideoCLIP model [3] with pretrained S3D checkpoints for video feature extraction[4].

We modify the inference model code to only use the:

```
output["pooled_text]
```

This avoids outputting the score for the pooled video:

```
output= {"score": torch.bmm
(output["pooled_video"][:, None, :]
output["pooled_text"][:, :, None]
).squeeze(-1).squeeze(-1)}
```

We use the `clip-vit-base-patch32` architecture for evaluating the CLIP model.

We use two unimodal models as baselines to compare the effects of video and image grounding: BERT (Devlin et al 2019) and Mirror-BERT (Liu et al 2021).

Mirror-BERT (Liu et al 2021): Mirror-BERT is a fast and effective contrastive learning technique that converts masked language models, like BERT, into encoders in 30 seconds without access to additional external knoweldge. Given a set of non-duplicated strings $\mathcal{X}$, individual labels y_i are assigned to each string. The resultant dataset $\mathcal{D} = \{(x_i, y_i) \| x_i \in \mathcal{X}, y_i \in \{1, \ldots, \|\mathcal{X}\|\}\}$ is turned into a self-reduplicated dataset $\mathcal{D}'$, where every term (x_i, y_i) is reduplicated as $(\bar{x_i}, \bar{y_i})$. The Mirror-BERT encoder is finetuned on data where random span masking is applied. This randomly replaces a random string of length k with `[MASK]` in either x_i or $\bar{x_i}$. Given a self-reduplicated dataset $\mathcal{D}'$, positive pairs are clustered together and negative pairs are pushed away using InfoNCE loss (Oord et al 2018).

3.2 Word-Level Intrinsic Semantic Evaluation Tasks

We evaluate BERT, Mirror-BERT and CLIP models on the SimLex-999, SimVerb-3500, CARD-600, RG65, MEN-3000 lexical semantic datasets.

SimLex-999 (Hill et al., 2015) has 666 noun pairs, 222 verb pairs and 111 adjective pairs. SimLex-999 assesses lexical similarity across nouns, verbs and adjectives. SimLex-999 assigns a concreteness quartile, ConcQ, to every word pair in the dataset, ranging from 1 (most abstract) to 4 (most concrete). For example, the word pair *sofa-chair* have concreteness quartile score of 4 (i.e very concrete)

MEN-3000 (Bruni et al., 2014) has 3,000 noun pairs. In unsupervised setting, we use all 3,000 pairs for testing. In the original configuration, 2,000 pairs are in the development part of the data set, 1,000 pairs are in the test part. In unsupervised setting, we use all 3,000 pairs for testing. The concepts contained were drawn from a visual data set (ESP-Game8). As a result, visual embeddings are expected to have a high coverage on it.

RG-65 (Rubenstein and Goodenough, 1965) has only 65 word pairs. Similarity of each pair is scored according to a scale from 0 to 4. It is one of the oldest word similarity data set but it still widely used in the community. It was constructed for testing synonym similarity (e.g. gem-jewel scored 3.94/4.00 and magician-oracle scored 1.82/4.00).

SimVerb-3500 (Gerz 2016): SimVerb-3500 cov-

[3] https://github.com/facebookresearch/fairseq/tree/main/examples/MMPT

[4] https://github.com/antoine77340/S3D_HowTo100M

ers all normed verb types from the USF free-association database, providing at least three examples for every VerbNet class. This broad coverage facilitates detailed analyses of how syntactic and semantic phenomena together influence human understanding of verb meaning.

Card-660 (Pilehvar et al., 2018) has 660 word pairs. It stresses subword and rare expression evaluations (e.g. *spontaneusness-rerurnability* and *retweeting-RTing*) , which can exhibit bit a different aspect of characteristics comparing to the data sets above.

3.3 Method

Each dataset contains a list of word pairs. Each word in the word pair is used as an input in the text encoder of the unimodal and multimodal models. The similarity between the embedding vector for each item in the word pair is calculated using cosine similarity, as in (3):

$$sim(x, y) = \left\langle \frac{x}{x_2}, \frac{y}{y_2} \right\rangle : R^d \times R^d \rightarrow R \quad (3)$$

We can compare the model similarity score to the gold standard similarity score reported in the dataset using Spearman's rank correlation coefficient ρ.

3.4 Results

Table 1 contains the results of the intrinsic evaluation of CLIP, Video-CLIP, BERT and Mirror-BERT on lexical semantics datasets listed above. **The Mirror-BERT enhanced baseline outperforms both CLIP and Video-CLIP on all datasets**, including those that assess verbal semantics (SimLex-999 Verb, SimVerb-3500), noun semantics (SimLex-999 Noun, MEN-3000, RG65) and on rare words dataset (Card-660).

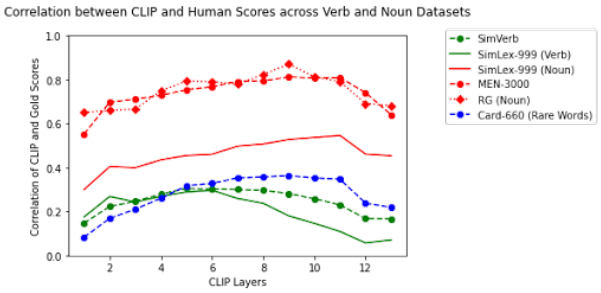

Figure 3: Figure showing the Spearman correlation coefficients ρ between human gold standard similarity scores and the cosine similarity between CLIP (Radford et al 2021) embeddings across intrinsic semantic evaluation datasets. The correlation scores are reported across the 13 layers of the CLIP model. Datasets focusing on nominal semantics are shown in red: MEN-3000, RG, SimLex-999 (Noun). Verb datasets are shown in green: SimVerb3500, SimLex-999 (Verb). The Card-660 dataset for rare words is illustrated in blue.

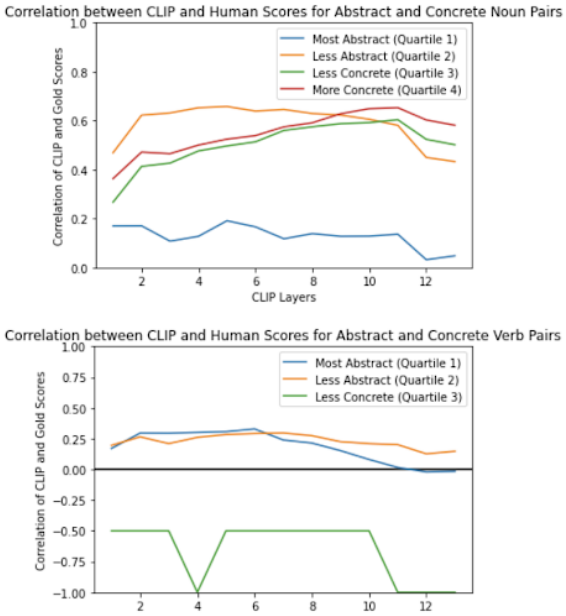

Figure 4: Figure showing the Spearman correlation coefficients ρ between human gold standard similarity scores and the cosine similarity between CLIP (Radford et al 2021) embeddings across the concreteness and abstractness quartiles of the SimLex-999 dataset. Figure 3(a) illustrates the effects of noun concreteness. Figure 3(b) illustrates the effects of verb concreteness

Table 1: Intrinsic Semantic Evaluation of CLIP and Video-CLIP against Noun and Verb Datasets

		CLIP	Video-CLIP	BERT Baseline	Mirror-BERT
SimLex-999	Overall	0.423	0.135	0.072	**0.515**
	SimLex-999 Noun	0.460	0.124	0.075	**0.535**
	SimLex-999 Verb	0.199	0.193	-0.027	**0.412**
MEN-3000 (Noun)		0.739	0.583	0.190	**0.802**
SimVerb-3500 (Verb)		0.246	0.107	0.013	**0.389**
RG65		0.751	-0.486	0.186	**0.829**
Card-660		0.277	-0.012	0.091	**0.333**

Figure 3 shows that there is a much higher correlation between gold standard and CLIP similarity scores in noun datasets than verb and rare word datasets across all 13 layers of the model. *Figure 4* shows the effect of concreteness and abstractness on the similarity between the gold standard and CLIP similarity scores in the SimLex-999 Noun and SimLex-999 Verb datasets. It illustrates that there is a much lower correlation coefficient for abstract nouns ($\rho \approx 0.2$ across all CLIP layers) compared to more concrete nouns, where $\rho \approx 0.7$ across the CLIP layers. It also illustrates that the correlation coefficient for concrete verbs, where $\rho \approx 0.25$ across CLIP layers, is much lower than concrete nouns. Note that the negative correlation coefficient for abstract verbs is due to a paucity of verb pairs annotated as abstract with `concQ = 3`. Similar findings for VideoCLIP are reported in the Appendix, where the VideoCLIP cosine similarity scores have a greater accuracy for more concrete terms.

The video-text pretraining in Video-CLIP raises two exceptions to the general trend that multimodal language models struggle with nouns more than verbs: (1) the correlation coefficient for SimLex-999 Verb 0.193 is greater than SimLex-999 Noun 0.124, and (2) there is poor performance on the noun dataset RG65. This may be indicative that video pre-training leads to improved representations for verbs compared to image pre-training. However, this may not necessarily be the case, as there is a much stronger correlation in the MEN-3000 dataset than the SimVerb-3500.

Subsequently, CLIP is evaluated on the Multi-SpA-Verb semantic similarity evaluation resource for verbs in English and other languages (Majewska et al 2020). This dataset is comprised of 17 verb classes that are clustered together. The similarity between word pairs in the class of verbs are calculated, and can be used as a human gold standard that we can compare the performance of CLIP. Some verb classes are inherently more abstract than others: for example, Class 1 consists of concrete verbs like *beat, punch, smash, slap*, while Class 8 consists of more abstract verbs like *ask, confess, discuss, inquire*. The correlation coefficients between the CLIP cosine similarities and the gold standard scores across the more abstract verb clusters are illustrated in *Table 2*

Abstract Verb	**CLIP**
accuse, condemn, forbid, blame	-0.091
achieve, aim, tackle, accomplish	-0.155
acquire, have, keep, borrow	0.227
dismay, frustrate, upset, irritate	-0.250
ask, confess, discuss, inquire	-0.227
approve, desire, prefer, respect	-0.179
calculate, analyze, predict, guess	-0.064

Table 2: **Abstract Verb Classes:** Spearman correlation coefficient ρ between gold standard verbs similarity scores in the English Multi-SpA-Verb dataset (Majewska et al 2020) across the five senses and predicted scores obtained using CLIP (Radford et al 2021) embeddings

Figure 5 shows the correlation scores between CLIP cosine similarity and the gold standard scores for the abstract verb clusters across the 13 model layers. We can see that the model performance of CLIP has a low, near-negative correlation with the gold standard scores across all abstract verb classes. There is a slight dip in the correlation coefficients in the middle layers of the model.

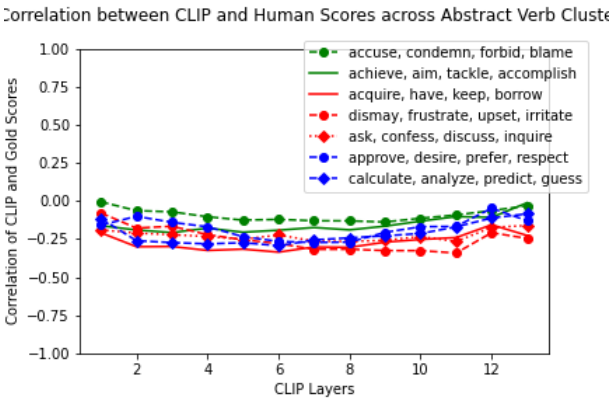

Figure 5: Figure showing the Spearman correlation coefficients ρ between human gold standard similarity scores and the cosine similarity between CLIP (Radford et al 2021) embeddings across abstract verb clusters in Multi-SpA-Verb (Majewska 2020). The correlation scores are reported across the 13 layers of the CLIP model.

What role does contrastive visual semantic pre-training have on lexical semantic capabilities of LMs?

- **Noun Semantics:** Strong performance in CLIP, particularly in higher model layers, where it is roughly equivalent to Mirror-BERT. Both Video-CLIP and CLIP outperform the BERT baseline, with VideoCLIP's performance on RG65 being an exception.
- **Verb Semantics:** Worse performance overall in all modalities, with Video-CLIP as a possible exception. CLIP and Video-CLIP outperform BERT baseline, but performs worse than Mirror-BERT.
- **Handling Rare Words:** Expected reduced performance across all models, with Video-CLIP having a particularly poor performance.
- **Effect of Image and Video Pre-training:** Overall, the scores for CLIP outperform Video-CLIP across all metrics.

Overall, while there is some evidence supporting the hypothesis that image and video grounding strategies can lead to improvements in the quality of lexical semantic representations, the strong performance of the unimodal language model Mirror-BERT illustrates that grounding in multimodal language models does not provide a consistent or uniform advantage over unimodal text-only representations. Consequently, this provides support for the Anti-Embodiment Hypothesis for Language Modelling. We further probe this hypothesis by focusing on two case studies on grounding nouns and verbs.

4 Case Studies in Grounding Events and Concepts: Colour and Telicity

In order to further probe the results of the intrinsic semantic evaluation of the multimodal language models, we evaluate the performance of CLIP and Video-CLIP in colour perception and verb telicity.

Why Telicity (and what is it)? Thrush et al (2020) perform a fine-grained syntactic analysis, evaluating the few-shot learning capabilities of the unimodal BERT model. They focused on the ability for unimodal language models to learn selectional preferences and different verbal alternations. To our knowledge, an equivalent case study evaluating the grammatical capabilities of multimodal language models has not been performed. We focus on the ability of multimodal language model to learn about **aspect**. This is the property that describes how an action, event or state of a verb phrase (VP) is situated in time. Telicity is one type of aspectual feature that marks whether a verbal action is **telic** and has an endpoint, or is otherwise **atelic** and lacks an endpoint. The **aspect hypothesis** claims that children typically associate past tense and perfective aspect with telic verbs, as these are actions that denote semantics of activity, accomplishment and achievement (Shirai & Anderson 1995; Todorova et al 2000).

In line with recent research that indicates video-language models have a better ability to encode verb trajectories (Ebert et al 2022), we hypothesise that a video-language model like VideoCLIP should have a better ability to learn about verbal telicity.

Why Colour? Colour is an example of a grounded conceptual space. Chamorro-Martinez et al (2020) model color categories as a fuzzy granular Conceptual Space, in the sense of Gardenfors (2000). The semantic category of colour allow human to partition the visual world. The linguistic category of colour has pragmatic implications, with conventional association between certain colours and emotions (e.g *red* and a state of anger in English). Colour is also a semantic category with significant typological variation— speakers across different languages and cultures choose to categorise different colours with different categories (Berlin & Kay 1991, McCarthy et al 2019).

4.1 Effect of Video and Image Grounding on Verbal Telicity

Metheniti et al (2022) assess the capability of unimodal Transformer language models, like BERT and RoBERTa, to learn about telicity. They utilise a telicity dataset produced by Friedrich & Gateva (2017). [5] An extract of the dataset is reproduced in *Figure 7*, where sentences containing a telic verb (e.g *John built a house in a year*, where *'building'* is a completed action) are annotated with 1.0 and sentences containing an atelic verb are annotated with 0.0 (e.g *John watched TV*, where *'watching'* does not have a definitive endpoint). As telicity is a lexical semantic property (typically) attributed to verbs, the dataset indicates the position of the verb in the sentence.

verb	verb_idx	label	Sentence	CLIP Embeddings
Seeing	0	1.0	Seeing a fire truck out in this weather is not...	[[-0.009166452, 0.0117951585, 0.008015579, -0....
double	17	0.0	At the beginning of March , headquarters sent ...	[[-0.0009636504, -0.0058122175, 0.0005942853, ...
punching	10	0.0	The motive was not so different from Clousarr ...	[[-0.010044061, 0.013732125, 0.0063730273, 0.0...
explain	15	0.0	Heston maganimously holds up a hand to read th...	[[-0.016657915, 0.0037022957, -0.0024827886, 0...
escaped	9	0.0	If there was any reason for this then it escap...	[[-0.0025535163, -0.017992515, -0.0005750102, ...

Figure 6: Extract of Telicity Training dataset produced by Friedrich Gateva (2017), used by Metheniti et al (2022).

Following the method outlined by Metheniti et al (2022), we extract the contextualised word embeddings from CLIP and VideoCLIP, which can then be utilised in task-specific models for classification. We conduct an experiment without finetuning where a logistic regression model from `scikit-learn` is applied to the contextualised embedding of the annotated verb in the dataset of each layer of the multimodal model. We can use the results from the telicity classification to examine how much information about verb telicity has been learnt by the model.

Table 3 reports the accuracy, precision and recall of the telicity classification experiment for CLIP and Video-CLIP, compared to the BERT scores obtained by Metheniti et al (2022). The best performing model was the unimodal BERT model `bert-large-cased`, which outperformed both multimodal models with image and video pretraining.

[5] https://github.com/lenakmeth/telicity_classification/

Table 3: Results of Telicity Classification by Multimodal models CLIP and VideoCLIP, compared to the unimodal BERT baseline obtained by Metheniti et al (2022).

	CLIP	**Video-CLIP**	**BERT**
Accuracy	**0.644**	**0.531**	0.88
Precision	**0.637**	**0.9**	0.87
Recall	**0.606**	**0.031**	0.87

Overall, this suggests that the hypothesis that video-language models offer a uniform advantage to the lexical semantics of verbs does not appear to hold, despite the apparent advantage that video pre-training offers to the classification of verbal events as telic or atelic.

4.2 Effect of Video and Image Grounding on Colour Perception

Abdou et al (2019) probes the colour representations of unimodal language models, observing that warmer colours show higher correlation coefficient with gold standard scores reported in the Color Lexicon of American English, compared to cooler colours. They suggest that this this may linked to an information theoretic communication bias associated with warmer colours observed by Gibson et al (2017), who claim that warmer colours are communicated more efficiently cross-linguistically.

We assess whether this claim generalises to multimodal language models, and whether higher correlation with the gold standard can be obtained in multimodal language models.

Paik et al (2021) introduce a dataset of human perceived distributions for 521 common objects called CoDa, which can be used to analyse and compare the colour distribution found in human perception and language models.

We generate CLIP and VideoCLIP for sentences in the CoDa dataset and train a classifier using the colour distributions in CoDA. *Figure 6* illustrates the correlation of the probability distribution of the CoDa test set with the predicted distribution.

We can see that the 'cooler colour hypothesis' generalises to multimodal language modelling, as there is a lower Kendall's tau correlation coefficient for colours like green and blue. Overall, the correlation scores for VideoCLIP are between $0.5 - 0.7$ and are higher than those for CLIP. This suggests that video pre-training, in particular, may be a useful grounding strategy for the colour.

5 Typological Perspective on Grounding Events and Concepts

Finally, we offer an empirical assessment of the final component of the Anti-Embodiment Hypothesis by performing an intrinsic semantic analysis of the capabilities of CLIP Italian and Multilingual CLIP: the benefits of grounding concepts and events should extend cross-linguistically.

5.1 Intrinsic Evaluation of CLIP Italian

First, we evaluate the CLIP Italian model using the Multilingual SimLex-999 evalation metric, and obtain a correlation coefficient $\rho = 0.351$. We confirm the finding of Bianchi et al (2021) that CLIP Italian outperforms Multilingual CLIP in this metric.

Subsequently, CLIP Italian is evaluated on the Multi-SpA-Verb metric, a multilingual semantic similarity evaluation resource for verbs in English, Finnish, Italian, Japanese, Polish, and Mandarin Chinese (Majewska et al 2020). This dataset is comprised of 17 verb classes that are clustered together. The similarity between word pairs in the class of verbs are calculated, and can be used as a human gold standard that we can compare the performance of CLIP Italian. Some verb classes are inherently more abstract than others: for example, Class 1 consists of concrete verbs like *beat, punch, smash, slap*, while Class 8 consists of more abstract verbs like *ask, confess, discuss, inquire*.

Abstract Verb	**CLIP Italian**
accuse, condemn, forbid, blame	0.134
achieve, aim, tackle, accomplish	-0.025
acquire, have, keep, borrow	-0.071
dismay, frustrate, upset, irritate	0.082
ask, confess, discuss, inquire	0.000
approve, desire, prefer, respect	-0.064
calculate, analyze, predict, guess	-0.064

Table 4: **Abstract Verb Classes:** Spearman correlation coefficient ρ between gold standard verbs similarity scores in the Italian Multi-SpA-Verb dataset (Majewska et al 2020) across the five senses and predicted scores obtained using CLIP Italian (Bianchi et al 2021) embeddings

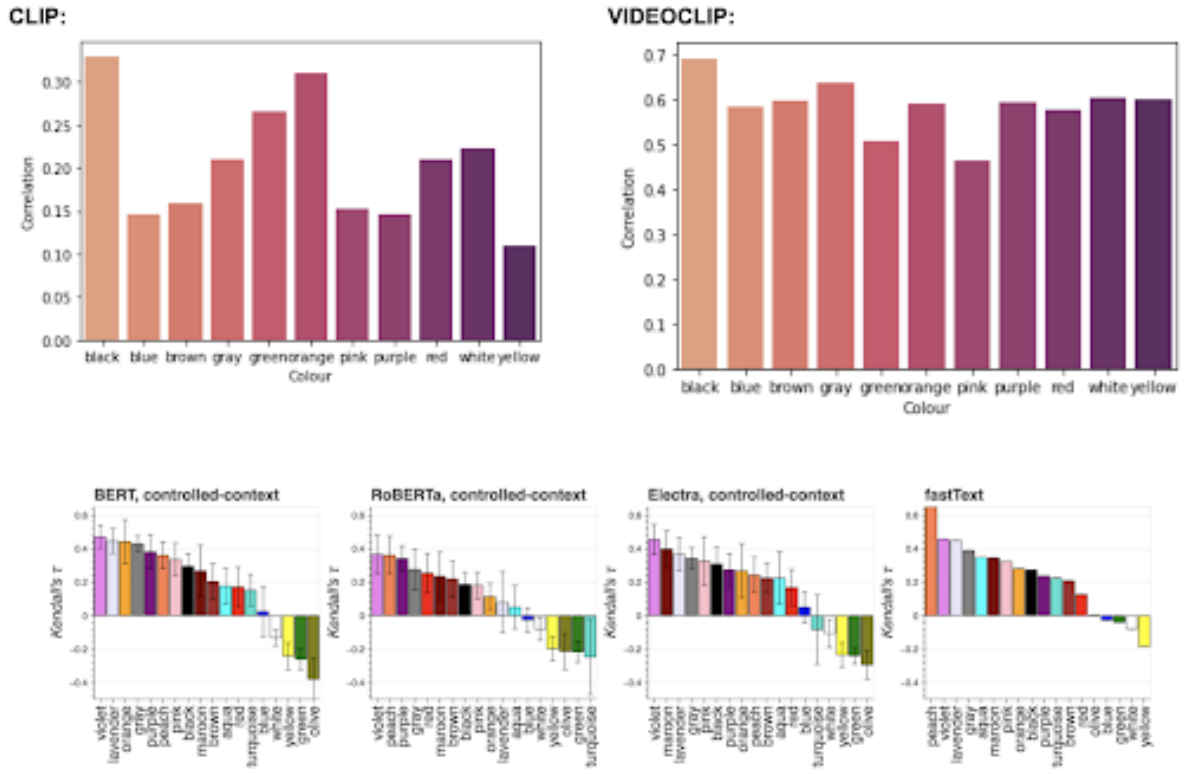

Figure 7: Kendall's tau correlation coefficient for CLIP and Video-CLIP colour probabilities, compared to unimodal LMs. Results for unimodal models replicate experiments performed by Abdou et al (2019)

Table 4 illustrates the results for verb classes that contain prototypically abstract verbs: generally, these classes lead to a weak and occassionally negative correlation coefficient, highlighting the poor performance of the CLIP Italian model on abstract verb semantics.

The correlation coefficient between the CLIP similarity scores and the Multi-SpA-Verb scores for verb classes containing prototypically concrete verbs (e.g *beat, punch*) are higher than for the abstract classes. Image-text pre-training appears to yield a better performance for some concrete classes, like *glance, observe, perceive, look*, than other concrete classes like *demolish, erode, wreck, disintegrate*. This suggests that the advantages posed by image-text pre-training in CLIP Italian do not lead to a uniform improvement in verb semantics.

In cognitive science, embodied representations are intended to be grounded in a multi-sensory fashion. The Lancaster Sensorimotor Norms (Lynott et al 2019) is a psycholinguistic dataset that presents ensorimotor strength for 39,707 concepts across six perceptual modalities (touch, hearing, smell, taste, vision, and interoception) and five action effectors (mouth/throat, hand/arm, foot/leg, head excluding

Concrete Verb	**CLIP Italian**
beat, punch, smash, slap	0.086
accelerate, decrease, shrink, increase	-0.018
climb, jump, roam, slide	0.093
bake, grate, slice, broil	-0.004
cough, gulp, inhale, sniff	0.002
chirp, hoot, roar, whistle	0.120
build, fasten, mend, restore	-0.008
drag, fling, haul, toss	0.081
demolish, erode, wreck, disintegrate	-0.060
glance, observe, perceive, look	0.206

Table 5: **Concrete Verb Classes:** Spearman correlation coefficient ρ between gold standard verbs similarity scores in the Italian Multi-SpA-Verb dataset (Majewska et al 2020) across the five senses and predicted scores obtained using CLIP Italian (Bianchi et al 2021) embeddings

mouth/throat, and torso), gathered from a total of 3,500 individual participants using Amazon's Mechanical Turk platform. [6] Recently, a sensorimotor dataset has been developed for Italian, providing scores for nouns and verbs (Repetto et al 2022). [7]

In order to ascertain how multimodal image-text pretraining differentially improves the multisensory quality of lexical semantic representations, we extract the CLIP Italian embeddings for the words in the Italian Sensorimotor norm dataset. We then train a logistic regression classifier using the CLIP embeddings to predict the sensory vector, where `Sense = [taste, smell, touch, audition, vision]`. A sample of the training set is illustrated in *Figure 7*.

Sense	CLIP Embeddings
[0.0, 2.86, 3.81, 10.48, 92.38]	[0.037453, -0.06964, -0.033154, -0.084774, 0.0...
[20.95, 10.0, 33.33, 17.14, 65.71]	[-0.049369, -0.098791, -0.0089411, -0.045771, ...
[1.0, 37.14, 97.0, 20.0, 60.95]	[-0.0044422, -0.14901, -0.12051, -0.13286, 0.0...
[5.71, 59.05, 68.57, 20.0, 83.81]	[-0.0884, 0.0363, -0.1301, 0.0405, 0.017, -0.0...
[9.52, 11.43, 17.14, 25.71, 56.19]	[-0.049161, 0.0012345, -0.13357, -0.042416, -0...

Figure 8: Training set of multi-sensory logistic regression classifier (Repetto et al 2022)

The results of the classification experiment are summarised in *Table 5*, where the correlation coefficient between the predicted grounding score across the five senses computed by the logistic regression classifier on the CLIP Italian embeddings is compared with the human gold standard. We can see that overall there is a weak multisensory grounding effect from CLIP Italian. CLIP Italian leads to a higher correlation in nouns across the five senses than verbs, where ther is a much weaker correlation coefficient. This further suggests that the grounding effects of concrete verbs using image-pretraining is much more limited than for nouns in the CLIP Italian model.

A similar analysis is performed for Italian adjectives on a dataset provided by Morucci et al (2019). Overall, this indicates that the image-text pretraining in CLIP Italian leads to an non-uniform improvement to the lexical semantic quality of Italian words, with a weak effect on verb semantics and particularly on abstract verbs.

[6]The Lancaster Sensorimotor Norm dataset is available from: https://osf.io/7emr6/

[7]Dataset is available from: https://osf.io/rcsnm/

Sense	Overall	Noun	Verb
Taste	-0.147392	0.328682	-0.143063
Smell	0.082427	0.312573	0.125839
Touch	0.039453	0.396762	-0.034844
Audition	0.000262	0.226903	-0.063218
Vision	-0.102450	0.315619	-0.066314

Table 6: Spearman correlation coefficient ρ between gold standard grounding scores in the Italian Sensorimotor Dataset (Repetto et al 2022) across the five senses and predicted scores obtained using CLIP Italian (Bianchi et al 2021) embeddings

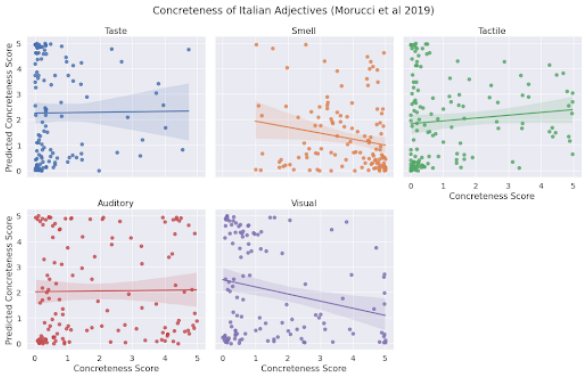

Figure 9: Distribution of predicted concreteness scores generated by CLIP Italian (Bianchi et al 2021) for adjectives that are principally grounded in one sense, compared to the gold standard concreteness scores reported by Morrcui et al (2019)

	Overall Score	Noun Score	Verb Score
English	0.306	0.496	0.114
Arabic	0.206	0.404	0.123
Chinese	0.369	0.530	0.250
Finnish	0.240	0.327	0.084
French	0.217	0.395	0.122
Hebrew	0.244	0.396	0.161
Polish	0.314	0.453	0.259
Russian	0.273	0.469	0.193
Swahili	0.142	0.298	0.136

Table 7: Spearman correlation coefficient ρ between gold standard similarity scores between word pairs in the Multi-SimLex-999 dataset (Vulić et al 2020) and predicted scores obtained using Multilingual CLIP embeddings across a typologically diverse set of languages

5.2 Intrinsic Evaluation of Multilingual CLIP

We assess the typological generalisability of our findings by evaluating the Multilingual CLIP model [8] on the Multi-SimLex-999[9] dataset. We use `LABSE Vit-L/14` with `LaBSE` as the text encoder and `OpenAI ViT-L/14` as the image encoder. The findings are summarised in *Table 6*. The correlation coefficient for nouns is higher than the corresponding coefficient for verbs in all languages, indicating that this is a typologically generalisable tendency in image-text pre-training.

6 Discussion and Conclusion

Evidence supporting the Anti-Embodiment Hypothesis:

- **Nouns vs Verbs:** There has been a consistent advantage for processing nouns over verbs in image-text and video-text pre-training.
- This tendency has observed in the improved performance of VideoCLIP in colour perception compared to a unimodal baseline.
- This tendency is clear throughout the model layers of CLIP, and is also typologically generalisable in CLIP Italian and Multilingual CLIP
- **Lower Performance on Abstract Nouns:** CLIP has worse performance in abstract nouns than concrete nouns
- **Lower performance on Abstract Verbs:** Poorer performance on more abstract verb clusters in Multi-SpA-Verb in CLIP Italian
- **Uneven performance in Sensorimotor Norms:** CLIP Italian does not provide a uniform advantage across all five senses when evaluated against the Sensorimotor norms dataset

Evidence against the Anti-Embodiment Hypothesis:

- **Findings are consistent with psycholinguistic studies on concreteness of nouns and verbs:** Evidence for heterogeneity in abstract verbs in CLIP Italian, consistent with the assessment of Muraki et al (2020). Grounding results in CLIP and Video-CLIP are consistent with
- **Potential for improved verb performance:** The performance of Video-CLIP in telicity classification suggests that video-text pre-training can lead to a higher precision than using image-text pre-training. Video-CLIP performs better in SimLex-999 verb than SimLex-999 noun, suggesting that video-text pretraining may concur an advantage for inducing verb trajectory representations

Recommendations for Multimodal Language Modelling Strategies:

- **Effect of Mirror-BERT Representations:** The strong intrinsic semantic performance of Mirror-BERT magnifies the quality of lexical semantic representations.
- **Enhancing Video-Text Semantics:** Could the contrastive learning technique employed Mirror-BERT be extended to image-text and video-text pre-training?
- **Improving Verb Semantics:** Verb semantic representations, particularly for abstract verbs, can be enriched using syntactic representations and truth-conditional representations (Emerson 2018 *et seq*)
- **Typological Generalisability:** MARVL (Liu et al 2021) sets out a protocol that collects conceptual data driven the the experiences of native speakers. This paradigm could be extended to focus on more fine-grained aspects of verb semantics and the typological variation surrounding them.

Acknowledgements

The first author has been supported, in part, by Gonville & Caius College, University of Cambridge. The first author would like to thank Theresa Biberauer for her advice and support on the project.

[8] https://github.com/FreddeFrallan/Multilingual-CLIP

[9] https://multisimlex.com/

References

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, page 6848–6854.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

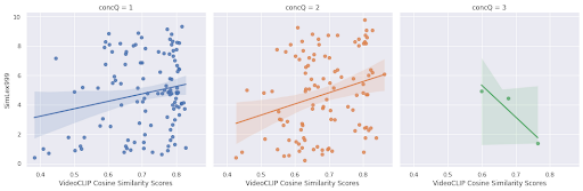

Figure 10: Correlation between cosine similarity scores obtained by VideoCLIP (Xu et al 2021) for SimLex-999 (Verb) and the gold standard scores ('SimLex999') across the quartiles of concreteness

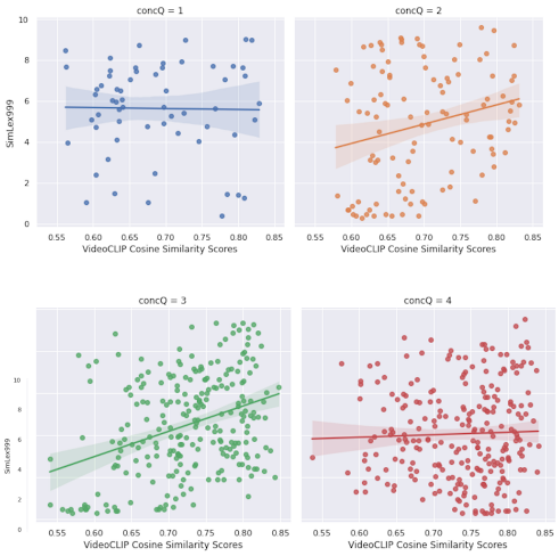

Figure 11: Correlation between cosine similarity scores obtained by VideoCLIP (Xu et al 2021) for SimLex-999 (Noun) and the gold standard scores ('SimLex999') across the quartiles of concreteness

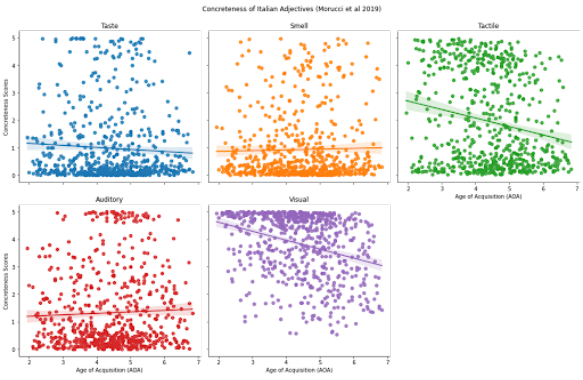

Figure 13: The developmental pathways of learning Italian adjectives ranging in abstractness/concreteness that are principally grounded in one sense.

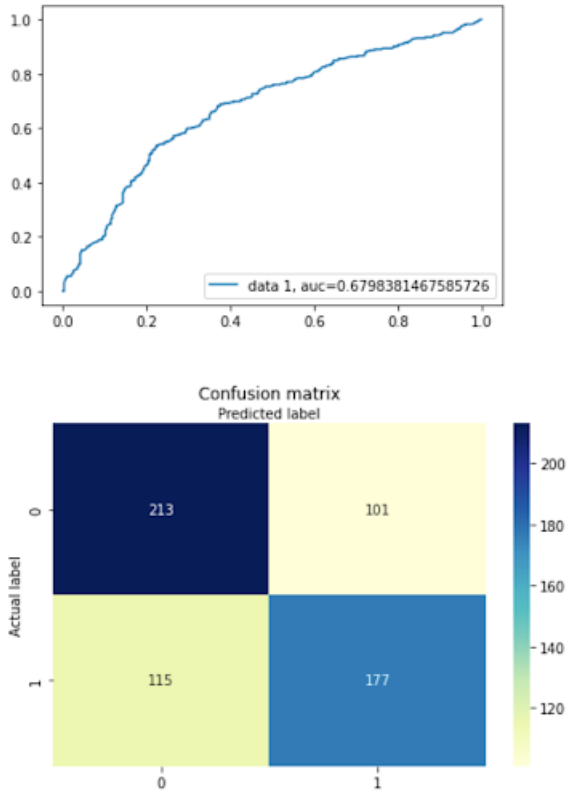

Figure 14: AOC and Confusion Matrix for CLIP Telicity Classification

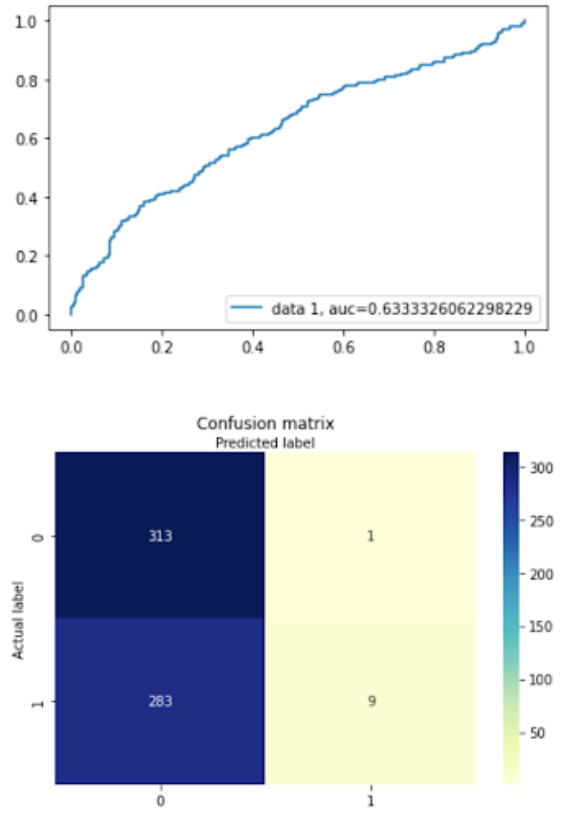

Figure 15: AOC and Confusion Matrix for VideoCLIP Telicity Classification