

# Report: ViLMA

Jesus Miguel Adrian Matos

July 19, 2024

## Abstract

This report is based on the seminar entitled *Visual grounding of verbs and nominalisations in multimodal models*, taken by Albert Gatt on May 29, 2024. In particular, I focus my attention on explaining the **VILMA** functionalities: I will not mention what datasets the **VILMA** tests were made on, but rather I will explain how it selects the foil-caption to carry out the tests. The report is structured as follows:

**Previous concepts:** concepts that are used in the report.

**Video Language Model Assessment:** I list models that will be evaluated by VILMA and I explain the VILMA tests.

**Pretrained models:** I list models evaluated by VILMA.

**Conclusions:** I observe several models on which the VILMA tests were applied, and I give a conclusion on VILMA benchmarks.

## 1 Previous concepts:

### 1.1 Visio-linguistic:

Visio-linguistic refers to the intersection and integration of visual and linguistic (language-based) information. It involves understanding and interpreting meaning that is conveyed through a combination of visual elements (images, diagrams, and visual symbols) and linguistic elements (text, spoken words, and written language). This term is particularly relevant in fields such as:

- **Cognitive Science and Psychology:** Studying how the human brain processes and integrates visual and linguistic information.
- **Computer Science and Artificial Intelligence:** Developing algorithms and systems that can understand and generate multimodal content, such as captioning images, generating descriptive text for videos, or creating systems that can engage in visual question answering (VQA).
- **Education:** Using visual aids alongside text to enhance learning and comprehension.
- **Communication and Media:** Designing effective ways to convey information using a combination of visual and textual elements, such as infographics, advertisements, and multimedia content.

In essence, visio-linguistic approaches consider how visual and linguistic inputs complement each other to enhance understanding and communication.

### 1.2 Image-Language models(ILMs):

An image-language model is an artificial intelligence system designed to process and understand both visual and textual data, integrating these two modalities to perform various tasks. These models can generate text descriptions from images, produce images from text descriptions, and understand the relationships between visual and textual elements. Key capabilities and applications of image-language models include:

- **Image Captioning:** Generating descriptive text for a given image.
- **Visual Question Answering (VQA):** Answering questions related to the content of an image.
- **Image Generation from Text:** Creating images based on textual descriptions.
- **Cross-modal Retrieval:** Finding images that match a given text or vice versa.
- **Image-Text Matching:** Evaluating the relevance or similarity between an image and a text description.

These models typically combine techniques from computer vision (for processing videos) and natural language processing (for handling text). They often use deep learning architectures such as:

- **Convolutional Neural Networks (CNNs):** For extracting features from images.
- **Transformers:** For handling and generating text, and sometimes for processing image features.
- **Multimodal Models:** Like CLIP (Contrastive Language-Image Pretraining) and DALL-E, developed by OpenAI, which are specifically designed to understand and generate both images and text.

### 1.3 Video-Language models (VidLMs):

Video-language models are advanced machine learning systems designed to understand and generate both video content and associated natural language descriptions. These models can interpret and generate content involving the complex interplay between visual data (videos) and textual data (language). Key capabilities and applications of video-language models:

- **Video Captioning:** Automatically generating descriptive text for video content.
- **Video Question Answering (Video QA):** Answering questions based on the content of a video.
- **Text-to-Video Generation:** Creating video sequences from textual descriptions.
- **Video Retrieval:** Finding relevant videos based on textual queries or descriptions.
- **Action Recognition:** Identifying and classifying actions depicted in video clips.

These models typically combine techniques from computer vision (for processing videos) and natural language processing (for handling text). Here, there are some Architecture and Models:

- **Encoder-Decoder Frameworks:** Commonly used where the encoder processes video frames to create a rich representation and the decoder generates the corresponding textual description.
- **Transformer-based Models:** These models leverage attention mechanisms to handle the complexity of video and language data, such as the Vision Transformer (ViT) and variants like VideoBERT.
- **Fusion Techniques:** Methods to effectively combine and align visual and textual data, such as cross-modal attention and joint embedding spaces.

## 1.4 Video-language datasets:

Video-language datasets are collections of video clips paired with corresponding textual descriptions, annotations, or other language-based data. These datasets are crucial for training and evaluating models in various tasks, including video understanding, captioning, question answering, and more. Here are some notable video-language datasets:

### 1. MSR-VTT (Microsoft Research Video to Text):

- **Description:** A large-scale video dataset with 10,000 video clips and 200,000 sentences. Each video has 20 sentences describing its content.
- **Use Cases:** Video captioning, retrieval, and understanding.

### 2. YouCook2:

- **Description:** Contains 2,000 long, unedited videos of cooking activities from YouTube. Each video is annotated with temporal segmentations and textual descriptions of each cooking step.
- **Use Cases:** Video segmentation, action recognition, and video captioning.

### 3. Charades:

- **Description:** Features 9,848 videos of daily indoor activities, each annotated with multiple textual descriptions and action labels.
- **Use Cases:** Activity recognition, video understanding, and human action detection.

### 4. ActivityNet Captions:

- **Description:** A subset of the ActivityNet dataset, with 20,000 videos from diverse activities. Each video is paired with temporally localized sentences describing the video content.
- **Use Cases:** Dense video captioning, temporal localization, and video summarization.

### 5. AVA (Atomic Visual Actions):

- **Description:** Contains 430 15-minute video clips with 80 atomic visual actions annotated per frame.
- **Use Cases:** Action detection, video understanding, and temporal action localization.

## 1.5 Foil:

Descriptions of the image that are highly similar to the original ones, but contain one single mistake ('foil word') [?].

## 1.6 MCQ model:

The MCQ (Multiple-Choice Question) model is a widely used format for assessing knowledge, skills, and abilities in educational settings and beyond. Here's an explanation of its key aspects:

### • Components of MCQ

1. **Stem:** This is the question or problem statement. It provides the context and sets up the scenario for which the examinee must select the correct answer.
2. **Options:** These are the possible answers provided to the examinee. Typically, there are several options (usually 4-5), and only one of them is correct.
3. **Key:** This is the correct answer among the options.

4. **Distractors:** These are the incorrect options designed to distract or mislead examinees who do not know the correct answer. Good distractors are plausible and often based on common misconceptions or errors.

- **Types of MCQs**

1. **Single Best Answer:** The most common type, where there is only one correct answer.
2. **True/False:** Simplified MCQs where the examinee must decide whether a statement is true or false.
3. **Multiple True/False:** Each option must be judged as true or false independently.
4. **Matching:** Examinees match a set of stems with a set of options.
5. **Assertion-Reason:** Two statements are given, and the examinee must determine if each is true and if one statement correctly explains the other.

## 2 Video Language Model Assessment (ViLMA):

It is a task-independent benchmark that details evaluation of the capabilities of the models (VidLMs) on a firm basis.

Through carefully selected counterfactuals, VILMA offers a set of controlled evaluations that shed light on the true potential of these models (VidLMs).

While such assessments shed light on task performance and support cooperative analysis, they are limited in their abilities to reveal the visio-linguistic capabilities that models exhibit across tasks.

Therefore, VILMA is focused on measuring the temporal understanding capabilities of a **VidML**. VILMA has 6 test, the first is the preliminary and the rest are the main ones. This works as follows, each of the main tests has a **specific foiling functions** and also a specific objective for the preliminary test (**Proficiency test**). An example of how these tests are applied can be seen in Figure 1.






Test (#exs.)	Video Caption (blue) / Foil (orange)	Foil Generation	Sample Frames
<b>Action Counting</b> (1432)	Someone lifts weights exactly <b>two</b> / <b>five</b> times.	Number replacement	
<b>Situation Awareness</b> (911)	A <b>policeman</b> / <b>blond man</b> holds a <b>blond man</b> / <b>policeman</b> against a wall.	Actor swapping	
	A man in blue <b>holds</b> / <b>chops</b> up a man in green.	Action replacement	
<b>Change of State</b> (998)	Someone <b>folds</b> / <b>unfolds</b> the paper.	Action replacement	
	Initially, the paper is <b>unfolded</b> / <b>folded</b> .	Pre-state replacement	
	At the end, the paper is <b>folded</b> / <b>unfolded</b> .	Post-state replacement	
	Initially, the paper is <b>unfolded</b> / <b>folded</b> . Then, someone <b>folds</b> / <b>unfolds</b> the paper. At the end, the paper is <b>folded</b> / <b>unfolded</b> .	Swap-and-replacement	
<b>Rare Actions</b> (1443)	<b>Drilling into</b> / <b>Calling on</b> a phone.	Action replacement	
	Drilling into a <b>phone</b> / <b>wall</b> .	Object replacement	
<b>Spatial Relations</b> (393)	Moving steel glass <b>towards</b> / <b>from</b> the camera.	Relation replacement	

Figure 1: The VILMA tests on VidLMs [?, Table 1].

Structure of each main test [?]:

- Structure 1.**
1. In step 1, be harvested high-quality examples from existing **video-language datasets**.
  2. In step 2, be created counterfactual examples or '**foils**', so that a test requires distinguishing correct from counterfactual video+text pairs.
  3. In step 3, be created a **Proficiency test** to gauge if a model learns the capabilities we deem necessary to solve the main test.
  4. In step 4, be applied automatic and manual validation of the examples and their counterfactuals to control for biases and to ensure a high-quality evaluation benchmark.
  5. Finally step, be tested whether existing VidLMs can solve the **Proficiency tests** and distinguish correct from counterfactual examples in the main tests.

Explanation of the Tests of VILMA:

Here, I explain how the tests work. In step 1 of the structure 2, the choice of dataset is mentioned. I will not mention these since in this report I am focusing on explaining the functionalities and objectives of these tests. That said, each test uses a different dataset with specialized annotations for the purpose of the test (for more information review [?]).

### 2.0.1 PROFICIENCY TESTS:

**Brief definition:** It is a preliminary test for the main tests. This test will be called with different objectives from the main tests.

This test computes the capacity of the model (VidML) to solve simple visuo-linguistic tasks, that do not have a demanding temporal model for this test.

This test is useful to rule out if there is bias in a model, for example, when a VidML passes the main test, but not the **Proficiency test**, then there should be bias in the VidML.

Thus, as this test will be called by all main tests, it is important to mention that not all **Proficiency test** objectives are useful for all main tests, below are the objectives for each main test [1]:

Objectives for **Spatial Relation**, **Change State** and **Situation Awareness** test: The objective is to identify objects mentioned in a caption.

Objectives for **Action Counting** and **Rare Action** test: The objective is to recognize actions in a caption and validate existence of objects (objects on which some action falls and subjects who perform it) in the frame respectively.

**Procedure of Proficiency test is:** The SpaCy dependency parser is used to identify *target words* (verbs for actions and nouns for objects) and mask them (the mask is a type of tokenization). Then the *target words* are replaced by *foil words* generated with **Masked Language Modeling (MLM)** using RoBERTa or T5, generate 3 *words foil*, which gives us 3 *foil captions*.

To validate the *words foils* applying an ALBERT 10 model finetuned on **Natural Language Inference (NLI)**. ALBERT 10 model is used to discard *words foils* that are entailed (E) by the *original caption*, that mean, these *foil words* do not change the meaning of the *caption*, because the *foil caption* should not coincide with the frame that represents the *original caption*, *foil words* that are neutral (N) or contradictory (C) be accepted as *foil words candidates*.

The second validation for *foil words candidates* is the **GRUEN** score. The **GRUEN** score is calculated through **BERT** model finetuned on the **Corpus of Linguistic Acceptability (CoLA)**. **GRUEN** rejects samples smaller than a pre-established threshold.

If none of the *foil words* pass both steps (NLI and CoLA), that *caption* is discarded.

### 2.0.2 ACTION COUNTING:

**Brief definition:** Measures the ability to accurately count the actions that occur in the video.

**Structure 2:**

- Step 1: specialized dataset
- Step 2:
  - First, the actions are noted at the end of the action, along with the frame number in which it ended. A structure is created to note the number of times the action is repeated, such as “someone performs exactly <number> push-ups” and the <number> value is called *placeholder*.
  - Second, the **placeholder** of *caption* be annotated with the correct value of repetitions of the action.
  - Third, *Foil-captions* are generated by changing the *placeholder number*, the *foil-caption* with placeholder that exceed the maximum value of all the actions perform in the video are discarded. In other words, among all the actions performed in the video, it counts how many times the most frequent action is repeated. This is done in two subtests.
    - \* In the first subtest called **easy**, an incorrect number of **placeholder number** of caption should be put, the incorrect values would be the values of the most frequent actions (a frequent action is the times the same action is repeated).
    - \* In the second subset called **difficult**, the *caption* no longer used, now the **foil-captions** generated by the **Proficiency test** are used (step 3), the *foil captions* be added the *placeholders* with the values of the most frequent actions.
- Step 3: The objective of the **Proficiency test** for this test is to find actions, but to use **Proficiency test** the caption should change its structure, for example if it is like this “a man performs exactly <number> push-ups” it should be replaced by “a man performs push-ups”. Having the *captions*, now the *foil-captions* are generated:
  - First the **Proficiency test** is used for the verbs, but the *captions* that have personal pronouns (like I, they, etc) and conjunctions (like and, but, etc.) are discarded.
  - Second, the **Proficiency test** is used for nouns.
  - Third, *caption* that are not valid in the video only should selected for that, the following steps should done:
    1. *Captions with similar actions* in other videos in the dataset be search.
    2. The verbs and nouns of the *caption* are replaced with those of the *captions with similar actions*.
    3. A *perplexity evaluation* is applied to the *foil-captions* to discard meaningless captions.
    4. Finally, *foil-captions* are taken at random from all *foil-captions* generate.
- Step 4: A manual review of the *foil-captions* and *captions* is carried out.
- Step 5: The *foil-captions* are entered into VidML, and their results are reviewed.

With this test we can see if the VidML is not biased by the most frequent number of actions.

### 2.0.3 SITUATION AWARENESS:

**Brief definition:** This test measures if the VidLM is aware of the actors and objects involved in an action. For example, if someone writes with a pen, then that pen should exist in the video. This test is divided in 2 tests:

- **Action Replacement:** This test evaluates the efficiency of the VidLM in distinguishing actions in a video sequence. For this, a *caption* should be taken and a *copy* is created of it, but the action (verb) of *copy* is changed.

- **Actor Swapping:** This test aims to recognize actors (someone performing an action) and objects on which an action is performed inside a video sequence. For this, a *caption* should be taken and a *copy* is created of it, but replace its actors and objects.

Following **Structure 2:**

- Step 1: Specialized dataset
- Step 2: *foil-captions* are obtained by Subtests **Action Replacement** and **Actor Swapping**.
- Step 3: For this test, the **Proficiency test** focuses on finding the identified *objects* and *actors* and so in step 4 be able to compare these *object* with the *foil-captions objects* created by the subtests of step 2.
- Step 4: A manual review of the *foil-captions* and *captions* is carried out.
- Step 5: The *foil-captions* are entered into VidML, and their results are reviewed.

#### 2.0.4 CHANGE OF STATE:

**Brief definition:** This test measures whether the model is aware of the implications of an action, such as:

- The action of walking: implies that someone is standing up, nobody can walk while sitting down.
- The action closing a door: implies that the door is open.

The test needs a set of verbs with their grammatical implications and preconditions to perform an action, these verbs are called *change of state (CoS) verbs*. These *CoS verbs* are stored in a tuple of dimension 4 ([pre-state, verb, pos-state, reverse verb]), such as [open, to close, closed, to open].

Then, the *captions* of the VidLM dataset that contain the *CoS verbs* are searched, these will be named *candidate captions*.

Then, the *candidate captions* are rebuilt the following **VILMA** structure, which differentiates the transitive verbs (verbs that also affect objects) from the intransitive verbs (which only affect the subject):

**Structure 2.** • **Action caption template:** “Someone <change-of-state-verb> the <object>.” for transitive change-of-state verbs, “The <subject> <change-state-verb>.” for intransitive ones.

- **Pre-State caption template:** “Initially, the <subject/object> is <pre-state>.”;
- **Post-State caption template:** “At the end, the <subject/object> is <post-state>.”;
- **Reverse caption template:** “Initially, the <object> is <pre-state>. Then someone <change- state-verb> the <object>. At the end, the <object> is <post-state>.” for transitive change- of-state-verbs. “Initially, the <subject> is <pre-state>. Then the <subject> <change-state- verb>. At the end, the <subject> is <post-state>.” for intransitive ones.

Therefore, the **action caption template** and the **reverse caption template** are the *foil-captions*.

Following **Structure 2:**

- Step 1: Specialized dataset
- Step 2: The *foil-captions* are obtained, with the structure 2.0.4.
- Step 3: For this test, **Proficiency test** focus only on the detection of objects or actors, depending on whether it is a transitive verb or not. If the verb is transitive then **Proficiency test** detects objects and actors, and if the verb is not transitive then **Proficiency test** detects only actors.
- Step 4: A manual review of the *foil-captions* and *captions* is carried out.
- Step 5: The *foil-captions* are entered into VidML, and their results are reviewed.

### 2.0.5 RARE ACTIONS:

**Brief definition:** The objective of this test is measuring VidLM’s ability to identify novel compositions or unusual compositions, for example ”cutting a computer keyboard with a chainsaw”.

This test requires that the unusual events should be described using a *verb-noun pair*, for example “cutting a keyboard.” These descriptions should not have an actor.

The test is divided in two subtests:

- **Action Replacement:** The verbs within caption should be replaced by another that can accommodate this situation such as “typing on the keyboard”.
- **Object Replacement:** The objects within caption should be replaced by other objects found in the video, to be more specific, the other objects found should be within a range of 8 frames from where the frame of the original caption.

These two subtests **Action Replacement** and **Object Replacement** return the foil-captions.

Again, following **Structure 2:**

- Step 1: Specialized dataset.
- Step 2: The *foil-captions* are obtained by **Action Replacement** and **Object Replacement** subtests.
- Step 3: For the **Rare Action test**, **Proficiency test** only focus on finding the objects, and the object within foil-captions no longer replaced by any object, but it will be some objects found in 8 frames around the original caption frame.
- Step 4: A manual review of the *foil-captions* and *captions* is carried out.
- Step 5: The *foil-captions* are entered into VidML, and their results are reviewed.

### 2.0.6 SPATIAL RELATIONS:

The objective of this test is to measure the ability of a VidLM to understand spatial relations as spatio-temporal relations (prepositions) in a video, such as “moving something towards or from something”.

For this test, *caption* that contain a spatial or spatio-temporal relations should be looked for within the dataset of the **VidML** (I will call it a set of captions).

Then, the *foil-captions* should be created, a *foil-caption* is created as follows: A *copy of a caption* is created, then replace its prepositions in the *copy* with the prepositions from the set of *captions*. Then, a *perplexity test* is performed on the copy and if the *copy* passes the test it would be a valid *foil-caption*.

Then, the 10 *foil-captions* are taken and pass them over an **NLI** model, the neutral and contradictory *foil-caption* are kept, the rest is discarded. Then, the **GRUEN** score is calculated to discard some more. Then the remaining *foil-captions* will be the *foil-captions candidates*.

Following **Structure 2:**

- Step 1: Specialized dataset
- Step 2: The *foil-captions* are obtained as indicated above..
- Step 3: For this test, the **Proficiency test** only focuses on identifying the objects in the *caption*, similar to the **Proficiency test** carried out for the **Change of Status test**.
- Step 4: A manual review of the *foil-captions* and *captions* is carried out.
- Step 5: The *foil-captions* are entered into VidML, and their results are reviewed.



## 2.1 Pretrained VidLMs:

Models used in this benchmark of VILMA. **Unimodel Models:** Only-text based models (i.e. text-only MLs). I identified two kinds of models: the first one is the decoder-only models, focusing on generating and predicting aspects of the language tasks, such as language modelling, text completion and text generation. For testing VILMA, it has been used

- GPT2-2

Secondly, the encoder-decoder models, that are the same as the sequence-to-sequence models. They input a sequence which will be assigned another sequence as an answer, for example text translation, summarize a text and answer questions. For testing VILMA, it has been used

- T5
- BART
- BOTH

and for both, single encoder and encoder-decoder, namely:

- OPT

Similarly to VALSE, the perplexity values are computed for both captions and foil-captions, and the text input with lower perplexity scores are taken. Used parameters:

- GPT-2 124M
- OPT 6.7B

**Image Language Models:** The definition of an Image Language model has been previously introduced in 1.2. The models used for VILMA are:

- CLIP
- BLIP

For encoder-text, both uses OPT model.

**Video Language Models:** The definition of Video Language Model has reported in subsection 1.3. Hereinafter, the list of models used for VILMA:

- ClipBERT:
  - Text-encoder: BERT
  - Video-encoder: Resnet-50
  - Features:
    1. This pretrained model uses only images.
    2. No temporal order: the video-text similarity score is the average frame-text similarity score.
- UniVL:
  - Text-encoder: BERT
  - Video-encoder: S3D
  - Features:
    1. Dual-stream Architecture: UniVL employs a dual-stream architecture consisting of a video encoder and a text encoder. The video encoder processes the visual information, while the text encoder processes the language information. Both encoders are built using Transformer layers.

2. Cross-modal Transformer: After encoding the video and text separately, UniVL employs a cross-modal Transformer to fuse the information from both modalities. This cross-modal Transformer is crucial for understanding and generating coherent video-language representations.
  3. UniVL is pretrained on HowTo100M.
- VideoCLIP:
    - Text-encoder: BERT
    - Video-encoder: S3D
    - Features:
      1. It uses mean pooling to fuse modalities, this is similar to the Video-text similarity score for ClipBERT.
      2. VideoCLIP is pretrained on HowTo100M.
  - FiT:
    - Text-encoder: BERT
    - Video-encoder: TimeSFormer
    - Features:
      1. It be able to be pretrained on images(CC3M) and videos(W2).
      2. It creates a shared video-text space through contrastive learning.
  - CLIP4Clip:
    - Text-encoder: BERT
    - Video-encoder: Resnet-50
    - Features:
      1. Its primary goal is to enable effective retrieval of videos based on textual descriptions and vice versa.
      2. Video Representation: It uses the Clip video encoder, processing each frame as if it were an image.
      3. Text Representation: It use the Clip text encoder. Textual queries or descriptions are encoded into embeddings that reside in the same space as the video frame embeddings.
      4. Strategy for modeling space-time:
        - \* Simple aggregation methods like mean pooling over frame embeddings.
        - \* Sophisticated techniques such as attention mechanisms to capture the temporal dependencies between frames.
  - VIOLET:
    - Text-encoder: BERT
    - Video-encoder: Video Swin Transformer
    - Features:
      1. It could be pre-trained on images (CC3M) and videos (YT-Temporal, WebVid).
      2. Spatial and temporal dimensions of the video inputs are modelled by positional embeddings considering both spatial and temporal ordering.

- X-Clip:
  - Text-encoder: BERT
  - Video-encoder: Resnet-50, or ViT
  - Features:
    1. Contrastive Loss: During training, the model uses a contrastive loss function, typically a variant of the InfoNCE (Information Noise Contrastive Estimation) loss. This loss encourages the similarity scores of matching video-text pairs to be higher than those of non-matching pairs.
    2. It introduces the Attention Over Similarity Matrix (AOSM) module, enabling it to focus on essential frames and words while reducing the impact of irrelevant ones during retrieval.

MCQ (Definition in section1.8): A pretext task using Multiple Choice Questions (MCQ) was introduced for video-text pretraining, utilizing a dual-encoder approach. This involved a parametric module named BridgeFormer, which links local features from both VideoFormer and TextFormer to address multiple-choice questions through a contrastive learning objective. This method enhances the semantic connections between video and text representations, improving detailed semantic associations between the two modalities. Furthermore, it ensures high efficiency for retrieval, and the BridgeFormer can be omitted for downstream tasks.

- Singularity [?]:
  - Text-encoder: BERT
  - Video-encoder: ViT
  - Features:
    1. In the training phase, a single frame is randomly selected as input, and a video-level prediction is made using the information from this frame and its corresponding text input. During inference, multiple frames are uniformly sampled, and their encoded image-level representations are combined early on as input to the multi-modal encoder, see Figure ??.

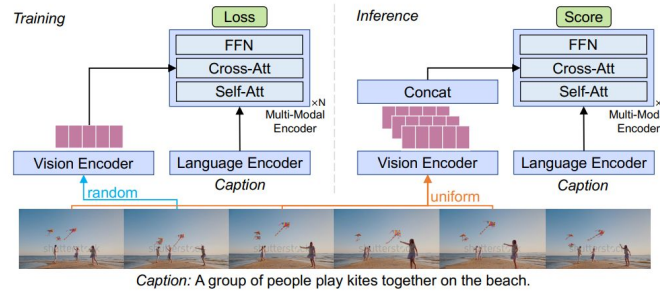


Figure 2: Process of the Singularity model [?, Figure 1].

UniPerceiver: It is used for perception tasks in videos. The model is designed to handle zero-shot and few-shot learning situations. UniPerceiver was inspired by the Transformer architecture but uses CNN for a variety of modalities such as text, images and videos [?], see Figure ??.

Merlot Reserve: It has a better understanding of temporal space for videos by combining audio, subtitles, and video frames. The model learns by substituting bits of text and audio with a MASK token and selecting the one that fits best describes the image.

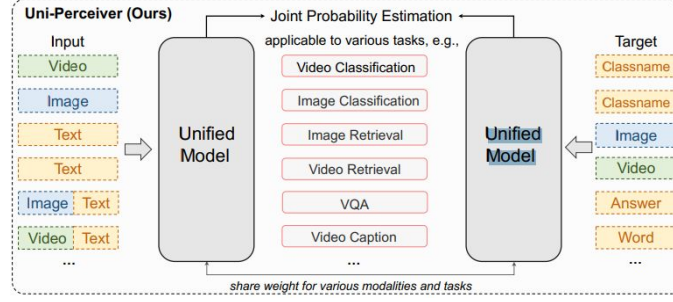


Figure 3: Process of the UniPerceiver model [?, Figure 1].

- VindLU:
  - Text-encoder: BERT
  - Video-encoder: BEiT (BERT Pre-Training of image transformer)
  - Features:
    1. Use a visual-text contrastive objective.
    2. These models add the following steps to improve their result: the inclusion of temporal attention, integration of a multimodal fusion encoder, adoption of masked modeling pretraining objectives, joint training on images and videos, utilization of additional frames both in fine-tuning and inference stages, model-parameter and data scaling.

### 3 Conclusion

In Figure 2 the scores of the **VILMA** tests are shown, for various models, where the letters **P** represent the main test and the **T** represents the **Proficiency test**.

These are ordered as follows, the first 2 are the **Unimodel Model**, the next 2 are the **Image Language Model (ILM)**, and the rest are the **VidLM**.

#### Observations from Figure 2:

So, it is relevant to notice that the best scores are for the ILM, the VidLMs have score acceptable, and as it was waiting, the Unimodel model are the ones with the lowest score, because they do not have images to validate some tests.

Also, one can also note that simple Unimodel models have good scores in **Proficiency tests**, since these do not imply a strong spatio-temporal relations.

Therefore, I can conclude that VILMA is more focused on VidLM that has spatio-temporal relations for videos, and is also very good for ILM, since the frames of a video are images and VILMA uses a constant frequency of frames to carry out their tests, which allows us to have images that represent periods of frames, and with this method a video could be considered as a set of images, but with a space-temporal relations.

Model	Action Counting			Situ. Awareness			Change of State			Rare Actions			Spatial Relations			Avg. P+T
	P	T	P+T	P	T	P+T	P	T	P+T	P	T	P+T	P	T	P+T	
Random	50.0	50.0	25.0	50.0	37.9	18.9	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	23.8
GPT-2 <sup>†</sup>	50.3	53.3	27.6	44.5	66.6	31.7	18.0	52.4	10.8	58.4	25.9	17.7	49.1	72.8	43.0	26.2
OPT <sup>†</sup>	56.2	54.6	31.0	51.7	<u>71.3</u>	<u>38.7</u>	23.1	48.0	12.9	59.0	23.9	14.9	59.0	<u>84.7</u>	<u>55.7</u>	30.6
CLIP <sup>‡</sup>	<u>90.5</u>	50.9	46.2	71.0	45.5	33.6	93.0	<b>55.2</b>	<b>52.2</b>	<u>92.7</u>	<u>93.9</u>	<u>87.8</u>	78.6	58.3	44.8	<u>52.9</u>
BLIP2 <sup>‡</sup>	80.9	54.5	43.7	<u>73.4</u>	<b>75.4</b>	<b>55.7</b>	74.5	52.1	38.1	93.8	74.5	70.5	<b>91.1</b>	<b>86.0</b>	<b>79.4</b>	<b>57.5</b>
ClipBERT	56.4	49.6	28.0	54.1	56.9	31.9	63.7	50.0	33.5	43.5	40.7	17.7	39.7	39.8	14.1	25.0
UniVL	73.4	43.6	32.2	51.6	46.6	24.1	81.3	54.3	43.0	77.5	78.0	59.9	62.5	51.7	33.2	38.5
VideoCLIP	79.1	46.4	36.5	61.6	40.3	24.9	49.8	50.8	25.9	84.0	77.8	67.5	67.9	54.7	39.7	38.9
FiT	83.9	52.4	44.6	69.8	40.0	29.1	93.0	52.1	47.8	89.7	89.4	80.7	70.5	51.9	38.7	48.2
CLIP4Clip	<b>91.2</b>	52.3	<b>48.0</b>	<b>73.8</b>	49.0	37.6	<b>94.8</b>	54.1	<u>52.1</u>	83.0	<b>94.1</b>	78.7	79.8	56.7	44.2	52.1
VIOLET	79.6	50.6	36.5	70.2	41.6	32.4	88.2	<u>54.6</u>	49.1	87.1	86.6	74.6	73.3	50.4	38.7	46.3
X-CLIP	84.1	<u>55.1</u>	46.4	63.5	44.8	31.0	85.7	52.7	46.0	83.9	85.7	72.3	74.8	56.2	43.5	47.8
MCQ	81.4	50.4	41.5	67.0	37.1	26.3	90.3	50.3	45.3	91.3	88.7	82.3	79.4	48.9	39.4	47.0
Singularity	79.6	51.1	41.5	68.8	40.9	30.1	92.8	<u>54.6</u>	50.3	<u>92.7</u>	88.4	83.1	80.7	46.8	38.9	48.8
UniPerceiver	50.6	46.4	23.0	51.4	42.1	21.1	67.5	46.1	29.1	58.2	58.8	34.7	45.5	48.0	20.1	25.6
Merlot Reserve	84.2	<b>56.0</b>	<u>46.9</u>	70.5	35.6	25.3	<u>93.4</u>	53.6	50.4	83.8	90.6	77.6	63.1	41.9	29.2	45.9
VindLU	84.5	51.2	43.5	70.5	41.6	31.2	85.4	52.6	45.6	<b>94.2</b>	93.1	<b>88.0</b>	<u>83.2</u>	45.6	39.4	49.5

Figure 4: Score of the models where the VILMA test was carried out [?, Table 2].

## References

- [1] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, and R. Bernardi. FOIL it! find one mismatch between image and language caption. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 255–265, Vancouver, Canada, July 2017b. Association for Computational Linguistics.
- [2] I. Kesen, A. Pedrotti, M. Dogan, M. Cafagna, E. Can Acikgoz, L. Parcalabescu, I. Calixto, A. Frank, A. Gatt, A. Erdem, E. Erdem. ViLMA: A Zero-Shot Benchmark for Linguistic and Temporal Grounding in Video-Language Models. arXiv:2311.07022.
- [3] J. Lei et al., "Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 7327-7337, doi: 10.1109/CVPR46437.2021.00725.
- [4] J. Lei, T. L. Berg, and M. Bansal. Revealing single frame bias for video-and-language learning. arXiv:2206.03428, 2022.
- [5] X. Zhu, J. Zhu, H. Li, X. Wu, H. Li, X. Wang, and J. Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16804–16815, June 2022.