KOSMOS 1 & 2

A Preprint

Overview

The KOSMOS-1 and KOSMOS-2 papers present a clear line of work aimed at extending large language models beyond text, first toward general multimodal perception (KOSMOS-1) and then toward explicit visual grounding (KOSMOS-2). Both papers share the same core philosophy: treat a causal language model as a universal interface, and align other modalities to it using token-level representations and next-token prediction. Overall, the work is conceptually clean, ambitious in scale, and influential, but it also comes with several practical and methodological limitations

KOSMOS-1: Core Idea and Motivation

KOSMOS-1 is built around a simple but powerful idea: a large language model can serve as a universal interface for intelligence if other modalities, especially vision, are properly aligned to it. Instead of designing specialized multimodal architectures or task-specific heads, the paper treats perception as another form of input that can be embedded into the language model's token stream. The underlying assumption is that the reasoning, generalization, and instruction-following abilities of large language models will naturally extend to multimodal tasks once perceptual signals are injected in a compatible way.

This idea reflects a strongly language-centric view of intelligence(I am not sure that language and tokens is enough to describe scenes in some kind of hidden state however authors show it works). Rather than asking how vision models can learn reasoning, the paper asks how reasoning models can be extended to see. As a result, KOSMOS-1 prioritizes generality and simplicity over precision or task specialization.

Overall Architecture

Architecturally, KOSMOS-1 is a decoder-only, autoregressive Transformer model. It predicts the next token conditioned on all previous tokens, regardless of whether those tokens originate from text or from other modalities. From the model's perspective, everything is ultimately represented as embeddings in a single sequence.

Visual inputs are handled through a separate vision encoder. Each image is first processed by a pretrained CLIP ViT-L/14 model, which converts the image into a grid of high-level visual features. These features are then passed through a resampler module(idea from DeepMind's Flamingo paper), which performs attentive pooling to reduce the number of visual tokens. This reduction step is essential for keeping the computational cost manageable, especially when images are interleaved with long text contexts.

The resulting visual embeddings are inserted into the language token stream using special delimiter tokens such as `<image>` and `</image>`. After this point, the Transformer does not distinguish between language and vision. Both are treated uniformly as input embeddings.

MAGNETO Backbone

Instead of a standard Transformer, KOSMOS-1 uses MAGNETO as its backbone architecture. MAGNETO is a Transformer variant designed to improve training stability and scalability, especially in large models trained from scratch.

The key architectural change introduced by MAGNETO is the use of additional LayerNorm operations within each Transformer block. These LayerNorms are applied in a way that stabilizes gradient flow through both the self-attention and feed-forward sublayers. Along with this, MAGNETO uses a carefully derived initialization scheme that helps prevent optimization issues when scaling to billions of parameters.

Additionally KOSMOS-1 employs XPOS relative positional encoding. XPOS improves generalization to sequence lengths that differ from those seen during training, which is important for multimodal prompts that can be much longer than typical text-only inputs.

Training Data and Datasets

A major strength of KOSMOS-1 lies in its training data strategy. The model is trained from scratch on web-scale multimodal corpora that fall into three main categories.

The first category is large-scale text data. This includes filtered versions of The Pile and Common Crawl, along with curated news and story datasets. These data are used to establish strong language modeling, reasoning, and instruction-following capabilities. Care is taken to remove benchmark contamination and duplicates.

The second category is image–caption pairs. These are drawn from large public datasets such as LAION-2B, etc. These datasets are noisy but extremely diverse, exposing the model to a wide range of visual concepts and their linguistic descriptions. Scale is prioritized over annotation quality.

The third category is interleaved image–text data extracted from web pages. These documents contain images embedded within paragraphs of text, similar to real-world web content.

All three data types are unified under the same input representation and training objective, which reinforces the idea that multimodal learning should not require modality-specific losses.

Training Objective and Algorithmic Perspective

From an algorithmic standpoint, KOSMOS-1 is deliberately simple. The model is trained using standard next-token prediction with a cross-entropy loss. The loss is applied only to discrete text tokens. Visual embeddings are never predicted; they only serve as conditioning context.

Formally, the model learns a conditional distribution of the form:

$$P(\text{next token} \mid \text{previous tokens}, \text{image embeddings})$$

This formulation avoids explicit alignment losses or contrastive objectives. Instead, alignment between vision and language emerges implicitly through large-scale co-occurrence in the data.

After multimodal pretraining, the model undergoes an additional phase of language-only instruction tuning using datasets such as Unnatural Instructions and FLAN-style instruction data. Notably, this tuning does not include images, yet the improved instruction-following behavior transfers to multimodal tasks. This further supports the paper's claim that language acts as the main control and reasoning channel.

Strengths and Limitations of the Approach

The main strength of KOSMOS-1 is its conceptual simplicity and generality. By reducing multimodal learning to language modeling, the paper presents a unified framework that is flexible, scalable, and compatible with existing LLM advances. The use of large-scale interleaved data is particularly effective for encouraging emergent multimodal behavior.

However, this design also introduces clear limitations. Visual understanding is implicit and coarse. The model has no explicit notion of objects, regions, or spatial structure, which limits its ability to perform precise visual

reasoning or grounding. Much of the visual capability is inherited from the frozen CLIP encoder rather than learned end-to-end.

In addition, while the model supports a wide range of tasks, its performance is often shallow compared to specialized systems. The emphasis is on breadth rather than peak accuracy.

Overall, KOSMOS-1 should be viewed as a foundational model. It demonstrates that aligning perception with language models is feasible and effective at scale, while also revealing the need for explicit grounding and spatial reasoning mechanisms, which are later addressed in KOSMOS-2.

KOSMOS-2: Motivation and High-Level Idea

KOSMOS-2 is a direct continuation of KOSMOS-1 and is explicitly motivated by its main weakness: the lack of visual grounding. While KOSMOS-1 can describe and reason about images, it cannot precisely refer to image regions, localize objects, or produce spatially grounded outputs. Everything remains implicit and language-only.

The core idea of KOSMOS-2 is to add grounding without abandoning the language-model-first philosophy. Instead of introducing detectors, region proposal networks, or task-specific heads, the paper encodes spatial information as discrete tokens and integrates it into the same autoregressive language modeling framework. In short, grounding is treated as another kind of language modeling problem.

Architecture Extension over KOSMOS-1

Architecturally, KOSMOS-2 stays very close to KOSMOS-1. It uses the same decoder-only MAGNETO Transformer backbone and the same vision encoder and resampler pipeline. The key difference is the introduction of location tokens and a new grounding-aware input format.

Bounding boxes are discretized into a fixed grid over the image. Each bounding box is represented by a small sequence of location tokens corresponding to its top-left and bottom-right corners. These location tokens are added directly to the model's vocabulary, meaning the model can both read and generate spatial information in token form.

To link language and vision, the paper introduces a Markdown-like representation(as previously with everything). Text spans (such as noun phrases or referring expressions) are wrapped in special tokens and followed by their associated bounding box tokens. This format explicitly teaches the model that certain pieces of text correspond to specific regions in the image.

A special `<grounding>` token is used to signal when the model should produce grounded outputs. Importantly, this mechanism does not change the core architecture; it only changes how inputs and outputs are formatted.

GRIT Dataset and Data Construction

A major contribution of KOSMOS-2 is the GRIT dataset, which enables grounding at scale. GRIT is constructed automatically from large image–caption datasets.

The construction pipeline works in two main stages. First, noun phrases are extracted from captions using a dependency parser. These noun phrases are then aligned to image regions using a pretrained grounding or detection model. Low-confidence alignments are filtered out to reduce noise.

Second, simple noun phrases are expanded into longer referring expressions using syntactic dependency relations. Redundant or nested expressions are removed, leaving a set of text spans that are linked to specific bounding boxes. The result is a massive collection of grounded image–text pairs, far larger than any manually annotated grounding dataset.

Training Objective and Algorithmic View

From an algorithmic perspective, KOSMOS-2 is trained in exactly the same way as KOSMOS-1: standard next-token prediction. The loss is applied only to discrete tokens, which now include both text tokens and location tokens.

Formally, the model learns:

$$P(\text{next token} \mid \text{previous tokens}, \text{image embeddings})$$

where the token space includes words as well as spatial location tokens.

There is no explicit grounding loss, no IoU-based supervision, and no region-level classification objective. Grounding emerges because the model is repeatedly exposed to patterns where certain text spans co-occur with certain spatial tokens.

After pretraining, KOSMOS-2 is instruction-tuned using a mixture of language-only instructions, vision–language instructions, and grounding-specific instructions. This step improves both textual responses and spatial outputs.

Problems and Limitations of KOSMOS-2

Despite being a clear improvement, KOSMOS-2 still has important limitations. Spatial precision is limited by the discretization of bounding boxes into a fixed grid. Fine-grained localization is therefore weaker than in specialized detection or grounding models.

Performance also depends heavily on language. The model performs better when referring expressions are long and descriptive, and worse when expressions are short, ambiguous, or purely spatial. This suggests that grounding is still driven more by linguistic cues than by strong visual reasoning.

Another concern is data noise. GRIT is automatically constructed and inevitably contains misalignments between text and image regions. While scale helps mitigate this, noisy supervision likely limits the upper bound of performance.

Finally, generating bounding boxes autoregressively can be inefficient and unstable, especially when multiple regions must be predicted. This reflects a broader limitation of using language modeling for structured visual outputs.