# ERC Advanced Grant Project CALCULUS: Extending the Boundary of Machine Translation

**Jingyuan Sun, Mingxiao Li, Ruben Cartuyvels and Marie-Francine Moens**
Department of Computer Science, KU Leuven, Belgium
{jingyuan.sun, mingxiao.li, ruben.cartuyvels, sien.moens}@kuleuven.be

## 1 Fact Sheet

- **Project Acronym:** CALCULUS
- **Grant Agreement ID:** 788506
- **DOI:** https://doi.org/10.3030/788506
- **Funding Agency:** European Research Council (ERC) under the EXCELLENT SCIENCE programme
- **Duration:** Start Date: 1 September 2018 - End Date: 30 September 2024
- **Principle Investigator:** Prof. Dr. Marie-Francine Moens
- **Coordinator:** Katholieke Universiteit Leuven, Belgium

## 2 Objective

The CALCULUS project, drawing on human capabilities of imagination and commonsense for natural language understanding (NLU), aims to advance machine-based NLU by integrating traditional AI concepts with contemporary machine learning techniques. It focuses on developing anticipatory event representations from both textual and visual data, connecting language structure to visual spatial organization and incorporating broad knowledge domains. Anticipatory event representations refer to representations that are able to predict what content is highly probable to be communicated next in a discourse. CALCULUS tests these models in NLU tasks and uses real-world metrics to evaluate their ability to predict untrained spatial and temporal details. CALCULUS employs machine learning methods, including Bayesian techniques and artificial neural networks, especially in data-sparse scenarios. The project's culmination involves the interdisciplinary studies in natural language processing, visual data analysis and cognitive neuroscience

## 3 Relation with Machine Translation

In the CALCULUS project, we are broadening the horizons of machine translation by delving into the essence of transforming the formats of data distribution while keeping the meaning. This innovative approach involves converting information from one modality into another, transcending traditional linguistic boundaries. Our project includes novel work on translating text into images, videos and layouts and brain signals to stimuli as illustrated below. The proposed models for multimodal translation can be a source of inspiration for future language translation (e.g., noise reduction in diffusion models, loss functions that preserve the structure of the source input).

### 3.1 Text to Image/Video Translation

Creating images and videos from text, which can also be seen as translating text into visual signals, is a key aspect of advancing artificial-intelligence-generated content (AIGC), with diffusion models standing out for their effectiveness. However, these models face the challenge of exposure bias, which refers to the training inference discrepancy. To address this, we introduce the time shift sampler (Li et al., 2024), a novel sampling method that reduces bias without needing to retrain the model and can be seamlessly integrated into existing algorithms like denoising diffusion





*Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 16–17
June 24-27, 2024 ©2024 European Association for Machine Translation

probabilistic model (DDPM) and denoising diffusion implicit model (DDIM), enhancing text-to-image translation efficiency with minimal computational increase. On the other hand, converting text to videos is more complex due to the larger output space, demanding more sophisticated models to generate natural videos. We propose the scene and motion conditional diffusion (SMCD) approach, incorporating scene semantics, motion dynamics, and textual information to improve text-to-video translation. Specifically, we leverage the first frame as semantic conditioning and the sequence of bounding box of objects as motion dynamic conditioning. The diffusion UNet incorporates both semantic and motion dynamic conditioning via gated self-attention and cross-attention layers. SMCD employs an advanced motion conditioning module and various scene integration methods, fostering synergy between modalities for dynamic and coherent video generation that aligns with the input text and motion dynamics.

### 3.2 Text to Layout Translation

Translating text into a 2D spatial layout involves understanding both language and spatial organization, a crucial step in text-to-image synthesis that allows for precise and controlled image generation. Our study (Nuyts et al., 2024) reveals that layouts can be predicted from language representations that incorporate sentence syntax, whether implicitly or explicitly, especially when sentences describe entity relationships similar to those encountered during training. We add explicit syntax by encoding a sentence "*John hits the ball*" with its constituent structure marked by brackets: "*(S (NP John) (VP hits (NP the ball)))*".

However, when testing models with grammatically correct sentences describing novel combinations of known entities and relations, we observe a significant drop in performance. This decline indicates that current models mainly rely on training data correlations instead of on a disentangled understanding of the structural complexity of input sentences. To address this challenge, we introduce a novel contrastive loss function that pulls 2D-layout representations towards an encoding of the syntax of the sentence they depict. Hence, the syntactic structure of input sentences is retained more effectively in the outputs, especially when structure was already explicitly present in the input sentences (cf. the example above). Our approach demonstrates marked improvements in predicting 2D spatial layouts from textual descriptions.

### 3.3 Brain Signals to Image Translation

We delve into the groundbreaking task of translating brain signals into images (Sun et al., 2023a; Sun et al., 2023b). This task is notably complex due to the noisy nature of fMRI (functional magnetic resonance imaging) brain signals and the sophisticated visual patterns they represent. Our methodology introduces a two-phase framework for fMRI data representation learning. Initially, we use a double-contrastive mask auto-encoder to pre-train a feature learner, effectively extracting representations by denoising data, since the noises inherent in fMRI will severely influence the reconstruction quality. The subsequent phase fine-tunes this learner, honing in on neural activation patterns vital for visual reconstruction, guided by an image auto-encoder. Our approach has demonstrated exceptional capability, significantly surpassing existing models in semantic classification accuracy. We believe that such technology will be highly useful in the future when multi-modal translation is expected to conduct directly on human's brain signals to ensure seamless and real-time experiences.

## References


Li, Mingxiao, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. 2024. Alleviating exposure bias in diffusion models through sampling with shifted time steps. *International Conference on Learning Representations*.

Nuyts, Wolf, Ruben Cartuyvels, and Marie-Francine Moens. 2024. Explicitly representing syntax improves sentence-to-layout prediction of unexpected situations. *Transactions of the Association for Computational Linguistics*, 12:264–282.

Sun, Jingyuan, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. 2023a. Contrast, attend and diffuse to decode high-resolution images from brain activities. In Oh, A., T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 12332–12348. Curran Associates, Inc.

Sun, Jingyuan, Mingxiao Li, and Marie-Francine Moens. 2023b. Decoding realistic images from brain activity with contrastive self-supervision and latent diffusion. In *European Conference in Artificial Intelligence 2023*, volume Volume 372: ECAI 2023, pages 2250 – 2257.