



\_akhaliq

Following

Message



578 posts

334 followers

0 following

AK

POSTS

REELS

TAGGED

## /visual Cortex and CortexBench

Website | Blog post | Paper



We're releasing CortexBench and our first Visual Cortex model, VC-1. CortexBench is a collection of 17 different AI tasks spanning locomotion, navigation, dexterous and mobile manipulation. We performed the largest and most comprehensive empirical study of pre-trained visual representations (PVFs) for Embodied AI (EAI), and find that none of the existing PVFs perform well across all tasks. Next, we trained VC-1 on a combination of over 4,000 ours of egocentric videos from 7 different sources and ImageNet, totaling over 5.6 million images. We show that when adapting VC-1 through task-specific losses or a small amount of in-domain data, VC-1 is competitive with the current state-of-the-art on all benchmark tasks.

## Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations

Jungo Kasai<sup>1</sup> Yubei Kasai<sup>2</sup> Keisuke Sakaguchi<sup>3</sup> Yutaro Yamada<sup>4</sup> Dragomir Radev<sup>5</sup>

<sup>1</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington  
<sup>2</sup>Sapporo Cardiovascular Clinic  
<sup>3</sup>Tokohu University  
<sup>4</sup>Yale University  
<sup>5</sup>Yale University

### Abstract

As large language models (LLMs) gain popularity among speakers of diverse languages, we believe that it is crucial to benchmark them to better understand model behaviors, failures, and limitations in languages beyond English. In this work, we evaluate LLM APIs (ChatGPT, GPT-3, and GPT-4) on the Japanese national medical licensing examinations from the past five years. Our team comprises native Japanese-speaking NLP researchers and a practicing cardiologist based in Japan. Our experiments show that GPT-4 outperforms ChatGPT and GPT-3 and passes all five years of the exams, highlighting LLMs' potential in a language that is typologically distant from English. However, our evaluation also exposes critical limitations of the current LLM APIs. First, LLMs sometimes select prohibited choices (禁選) that should be strictly avoided in medical practice in Japan, such as suggesting euthanasia. Further, our analysis shows that the API costs are generally higher and the maximum context size is smaller for Japanese because of the way non-Latin scripts are currently tokenized in the pipeline. We release our benchmark as JGAKU QA as well as all model outputs and examination data. We hope that our results and benchmark will spur progress on more diverse applications.

A 75-year-old man being alone was diagnosed with lung cancer with lung and bone metastases three months ago. He is currently receiving home care as per his wish to spend his last days at home peacefully. He has been offered bedridden care for the past few weeks and has been receiving visits from a visiting nurse service and a home helper as he cannot even go to the toilet. Five days ago, he developed lower back pain, and the visiting physician administered drug therapy, but his pain has worsened. Today, when the doctor visited him for treatment, the patient strongly expressed, "Please let me die today." "What is the appropriate response of the doctor to the patient's request today?"

- a. "Don't say things like that, let's do our best."  
b. "I will assist you in departing peacefully right away."  
c. "Let's consider hospitalization to a hospice right away."  
d. "There is a curative treatment for advanced lung cancer, so please don't worry."  
e. "Could you please tell me why you feel that way?"

This option shows consideration for the patient's wishes and emphasizes the importance of supporting the patient. Palliative measures, including euthanasia, are recognized under the Medical Practitioners' Act. However, depending on the patient's condition, it may be necessary to consider other options, such as treatment in a hospice specializing in palliative care, drug therapy to alleviate the patient's pain and anxiety, and psychological support. Regarding the choice of option, it is important to respect the patient's wishes. Therefore, option e, "Could you please tell me why you feel that way?", is the most appropriate response.

## Procedure-Aware Pretraining for Instructional Video Understanding

Honglu Zhou<sup>1,2</sup>, Roberto Martín-Martín<sup>1,3</sup>, Mubbasir Kapadia<sup>3</sup>, Silvio Savarese<sup>3</sup> and Juan Carlos Niebles<sup>1</sup>  
<sup>1</sup>Salesforce Research, <sup>2</sup>Rutgers University, <sup>3</sup>UT Austin  
{hzhou, mml353}@cs.rutgers.edu, robertom@cs.utexas.edu, {ssavarese, jniebles}@salesforce.com

### Abstract

Our goal is to learn a video representation that is useful for downstream procedure understanding tasks in instructional videos. Due to the small amount of available annotations, a key challenge in procedure understanding is to be able to extract from unlabeled videos the procedural knowledge such as the identity of the task (e.g., "make latte"), its steps (e.g., "pour milk"), or the potential next steps given partial progress in its execution. Our main insight is that instructional videos depict sequences of steps that repeat between instances of the same or different tasks, and that this structure can be well represented by a Procedural Knowledge Graph (PKG), where nodes are discrete steps and edges connect steps that occur sequentially in the instructional activities. This graph can then be used to generate pseudo labels to train a video representation that encodes the procedural knowledge in a more accessible form to generalize to multiple procedure understanding tasks. We build a PKG by combining information from a text-based procedural knowledge database and an unlabeled instructional video corpus and then use it to generate training pseudo labels with four novel pre-training objectives. We call this PKG-based pre-training procedure and the resulting model *PaperLika*. Procedure-Aware Pre-training for Instructional Knowledge Acquisition. We evaluate *PaperLika* on COIN and CrossTask for procedure understanding tasks such as task recognition, step recognition, and step forecasting. *PaperLika* yields a video representation that improves over the state of the art: up to 11.33% gains in accuracy in 12 evaluation settings. Implementation is available at <https://github.com/salesforce/paperlika>.

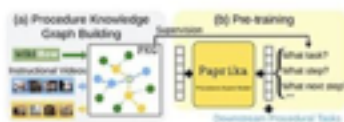


Figure 1. Training a video representation for procedure understanding with supervision from a procedural knowledge graph: the structure observed in instructions for procedures (from text, from videos) corresponds to sequences of steps that repeat between instances of the same or different tasks; this structure is well represented by a Procedural Knowledge Graph (PKG). (a) We build a PKG combining text instructions with unlabeled video data, and (b) obtain a video representation by encoding the human procedural knowledge from the PKG into a more general procedure-aware model (*PaperLika*) generating pseudo labels with the PKG for several procedure understanding objectives. *PaperLika* can then be easily applied to multiple downstream procedural tasks.

tasks such as the identification of the task, its steps, or forecasting the next steps. An agent that has acquired procedural knowledge is said to have gained procedure understanding of instructional videos, which can be then exploited in multiple real-world applications such as instructional video labeling, video characterization, process mining and, when connected to a robot, robot task planning.

Our goal is to learn a novel video representation that can be applicable to a variety of procedure understanding tasks in instructional videos. Unfortunately, prior methods for video representation learning are inadequate for this goal, as they lack the ability to capture procedural knowledge. This is because most of them are trained to learn the (weak) correspondence between visual and text modalities, where the text contains either human annotated descriptions (e.g., COIN)

## SoftCLIP: Softer Cross-modal Alignment Makes CLIP Stronger

Yuting Gao<sup>1\*</sup> Jinfeng Liu<sup>1,2,\*</sup> Zihan Xu<sup>1,\*</sup>  
Tong Wu<sup>1</sup> Wei Liu<sup>2</sup> Jie Yang<sup>2</sup> Ke Li<sup>1</sup> Xing Sun<sup>1</sup>

<sup>1</sup>Tencent YouTu Lab <sup>2</sup>Shanghai Jiaotong University  
{yutinggao, lanxxu}@tencent.com, {jfl19991226@sjtu.edu.cn}

### Abstract

In the preceding biennium, vision-language pre-training has achieved noteworthy success on several downstream tasks. Nevertheless, acquiring high-quality image-text pairs, where the pairs are entirely exclusive of each other, remains a challenging task, and noise exists in the used datasets. To address this issue, we propose a novel approach that relaxes the strict one-to-one constraint and achieves a soft cross-modal alignment by using a softened target, which is generated from the self-supervised intra-modal self-similarity. The intra-modal self-similarity is indicative to enable two pairs have some local similarities and model many-to-many relationships between modalities. Besides, since the positive still dominates the softened target distribution, we disentangle the positive in the distribution to further boost the relation with the negatives in the cross-modal learning. Experiments demonstrate the effectiveness of SoftCLIP on ImageNet zero-shot classification and COCO-MaskRCNN at pre-training dataset. SoftCLIP achieves a top-1 accuracy improvement of 6.8%/7.2% CLIP baseline.



Figure 1. (a) Three image-text pairs randomly sampled from the CC3M dataset have some local similarities, suggesting the existence of many-to-many relationships. (b) Using fine-grained intra-modal self-similarity as the softened target can allow for the existence of some similarities among unpaired image and text.

two unpaired samples. However, acquiring high-quality image-text pairs is a challenging task, owing to the fact that the majority of image-text pairs are obtained through

## $\infty$ -Diff: Infinite Resolution Diffusion with Subsampled Mollified States

Sam Bond-Taylor, Chris G. Willcocks  
Department of Computer Science  
Durham University

{samuel.e.bond-taylor, christopher.g.willcocks}@durham.ac.uk

### Abstract

We introduce  $\infty$ -Diff, a generative diffusion model which directly operates on infinite resolution data. By randomly sampling subsets of coordinates during training and learning to denoise the content at those coordinates, a continuous function is learned that allows sampling at arbitrary resolutions. In contrast to other recent infinite resolution generative models, our approach operates directly on the raw data, not requiring latent vector compression for context, using hypernetworks nor relying on discrete components. As such, our approach achieves significantly higher sample quality, as evidenced by lower FID scores, as well as being able to effectively scale to higher resolutions than the training data while retaining detail.

## Language Models can Solve Computer Tasks

Guanyao Kim  
University of California, Irvine  
kgwu@uci.edu

Pierre Baldi  
University of California, Irvine  
pfbaldi@uci.edu

Stephen McAleer<sup>\*</sup>  
Carnegie Mellon University  
smcaleer@cs.cmu.edu

### Abstract

Agents capable of carrying out general tasks on a computer can improve efficiency and productivity by automating repetitive tasks and assisting in complex problem-solving. Ideally, such agents should be able to solve new computer tasks presented to them through natural language commands. However, previous approaches to this problem require large amounts of expert demonstrations and task-specific reward functions, both of which are impractical for new tasks. In this work, we show that a pre-trained large language model (LLM) agent can execute computer tasks guided by natural language using a simple prompting scheme where the agent Recursively Critiques and Improves its output (RCI). The RCI approach significantly outperforms existing LLM methods for automating computer tasks and surpasses supervised learning (SL) and reinforcement learning (RL) approaches on the MiniWoB++ benchmark. RCI is competitive with the state-of-the-art SL+RL method, using only a handful of demonstrations per task rather than tens of thousands, and without a task-specific reward function. Furthermore, we demonstrate RCI prompting's effectiveness in enhancing LLMs' reasoning abilities on a suite of natural language reasoning tasks, outperforming chain of thought (CoT) prompting. We find that RCI combined with CoT performs better than either separately.

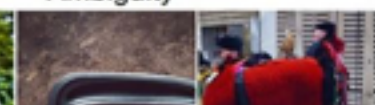
## Scaling Up Visual Speech Recognition With Synthetic Sup

Lakshminarayanan Lakomkin<sup>2</sup>, Konstantinos Vougioukas<sup>2</sup>, Pingchuan Ma<sup>2</sup>, Honglie Chen<sup>2</sup>

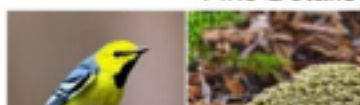
## SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer

Guanyao Chen<sup>1,2,\*</sup> Zhijian Liu<sup>4,\*</sup> Haotian Tang<sup>4</sup> Li Yi<sup>1,3</sup> Hang Zhao<sup>1,3</sup> Song Han<sup>4</sup>

## Ambiguity



## Fine Details





# MORE TO COME

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.