Multimodal Encoder Models

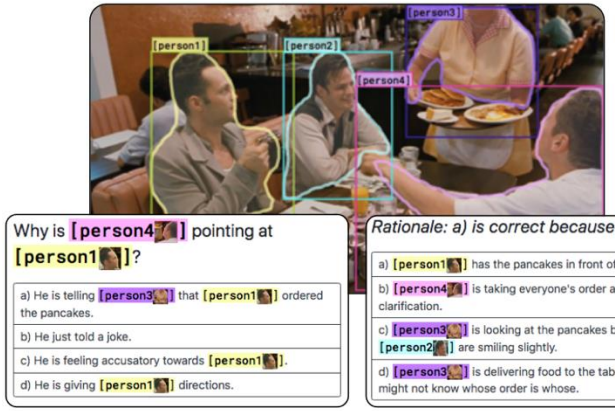
Advanced Large Language Models

ELL8299 · AIL861 · ELL881

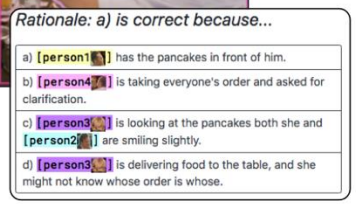
Manish Gupta
Principal Applied Scientist, Microsoft
https://sites.google.com/view/manishg/

Vision-and-Language Tasks

VQA



VCR Q→A

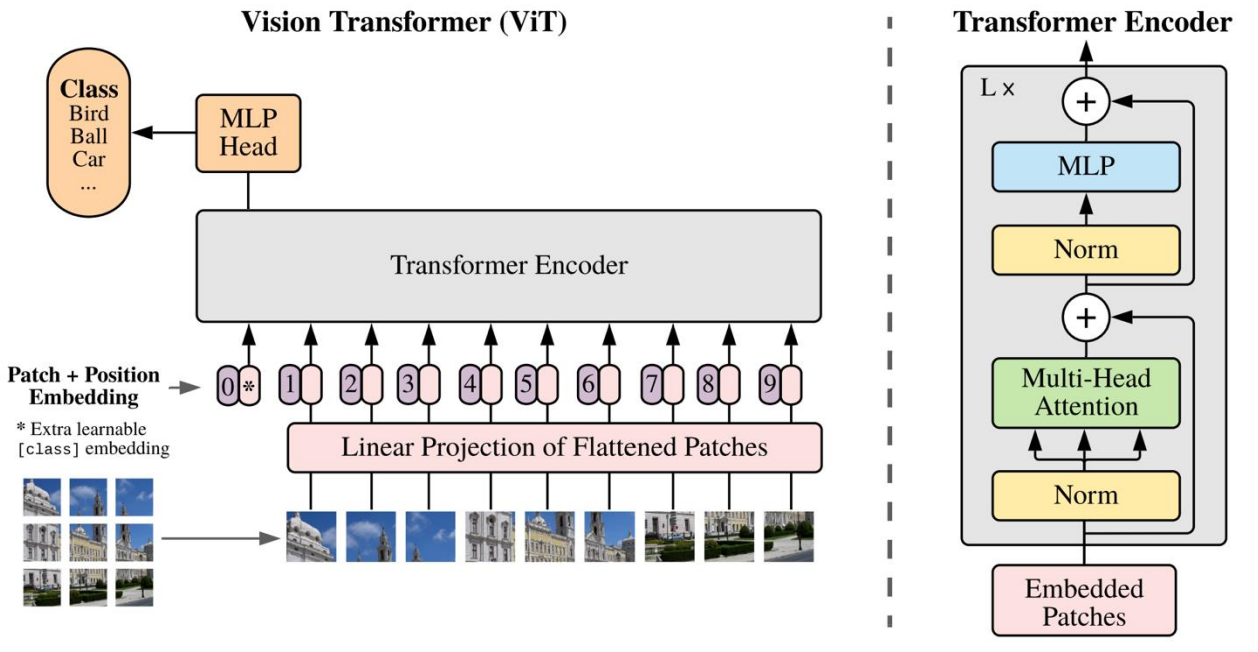


VCR QA→R

Referring Expressions

Caption-Based Image Retrieval

Vision Transformers



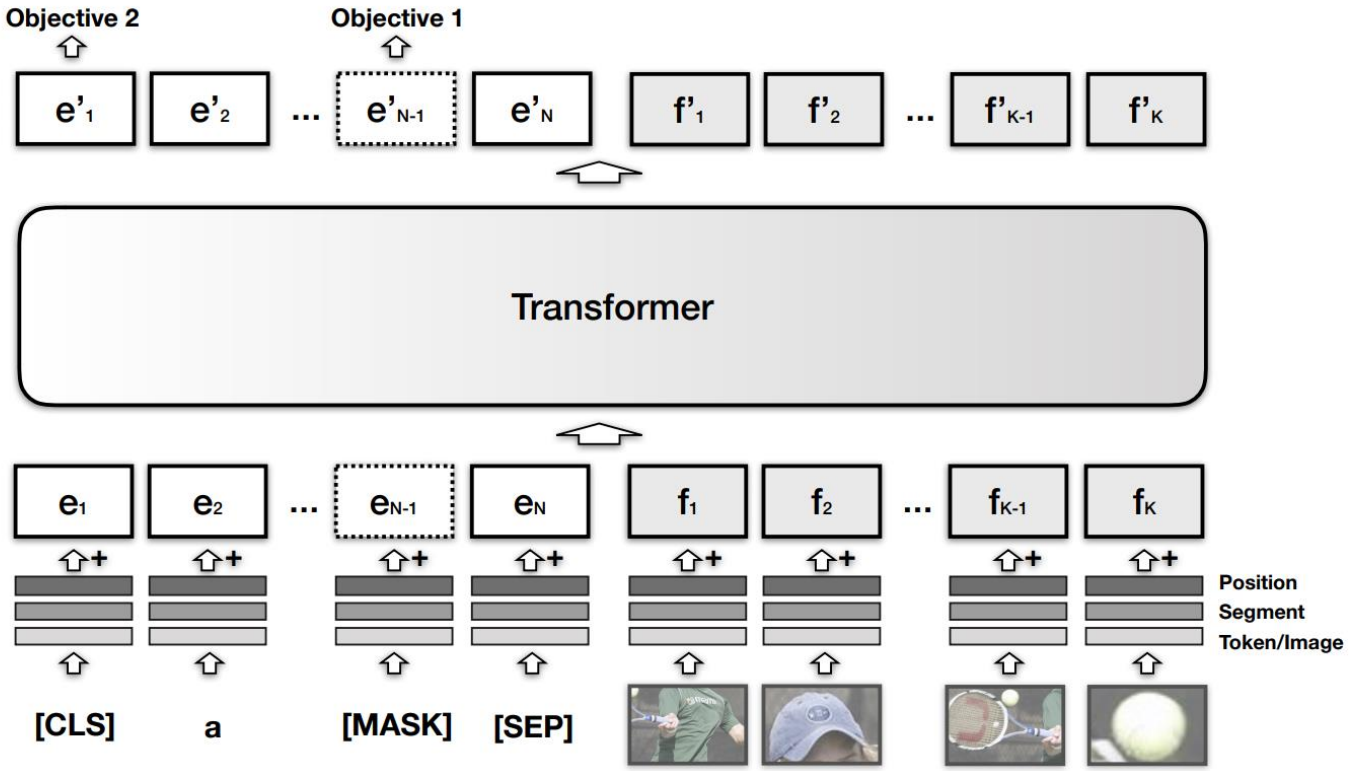
Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

- Split an image into fixed-size patches
- Linearly embed each of them
- Add 1D position embeddings
- Feed the resulting sequence of vectors (prepended by [CLS]) to a standard Transformer encoder.
- Classification MLP head with 1 hidden layer at pre-training and just a single linear layer at fine-tune time.
- Pretrain datasets: ImageNet-1K, ImageNet-21k, JFT
- ViT-L/16 means the "Large" variant with 16×16 input patch size. Smaller patch size ➔ larger seq length.
- Match or exceed accuracy of ResNets on many image classification datasets

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv:2010.11929 (2020).

Joint representation model for vision and language

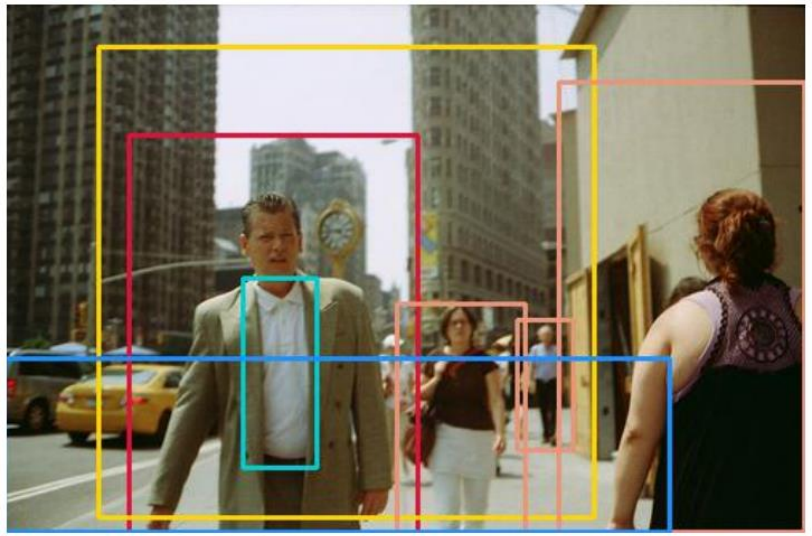
A person hits a ball with a tennis racket

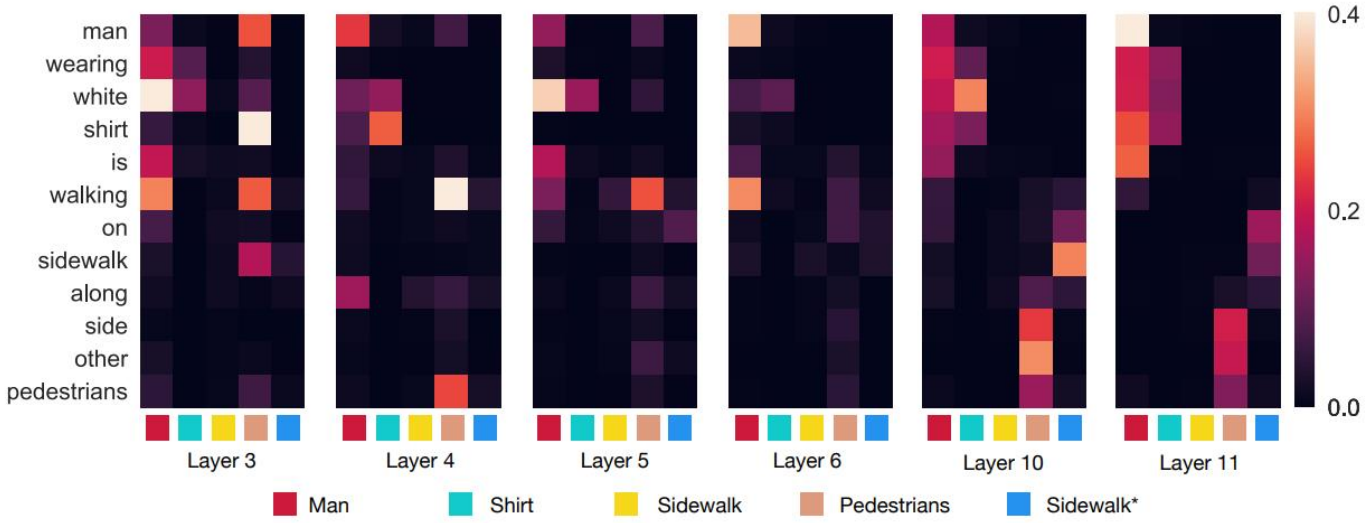


- MLM (Objective 1), and sentence-image prediction task (Objective 2)
- VisualBERT integrates BERT for NLP, and pretrained object proposals systems such as Faster-RCNN.

Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "Visualbert: A simple and performant baseline for vision and language." arXiv:1908.03557 (2019).

VisualBERT



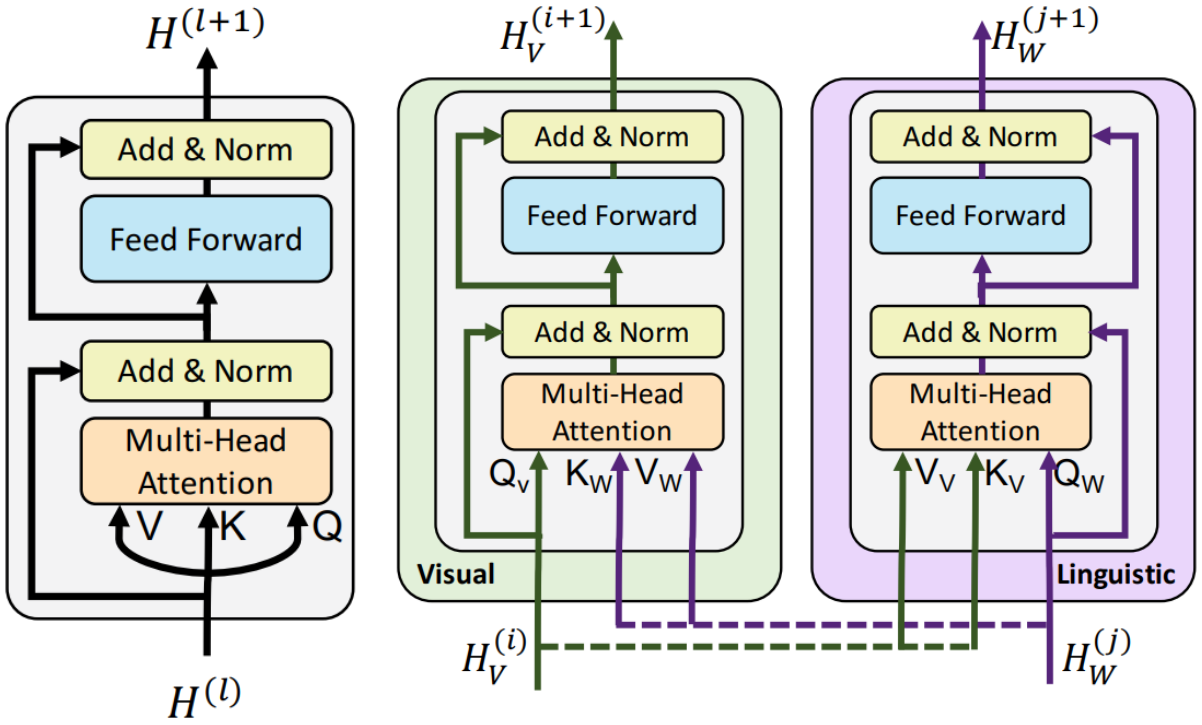


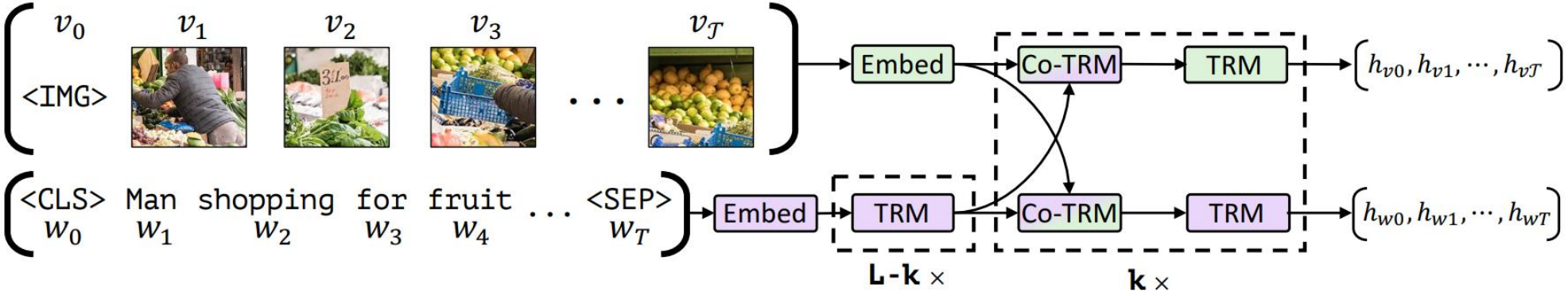
- Attention weights of some selected heads in VisualBERT.
- Model also refines its understanding over the layers
  - Incorrectly aligns “man” and “shirt” in the 3-rd layer but corrects them in higher layers.
  - In high layers (e.g., the 10-th and 11-th layer), it grounds visual concepts (e.g., “other pedestrians” and “man wearing white shirt”).

Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "Visualbert: A simple and performant baseline for vision and language." arXiv:1908.03557 (2019).

ViLBERT Architecture

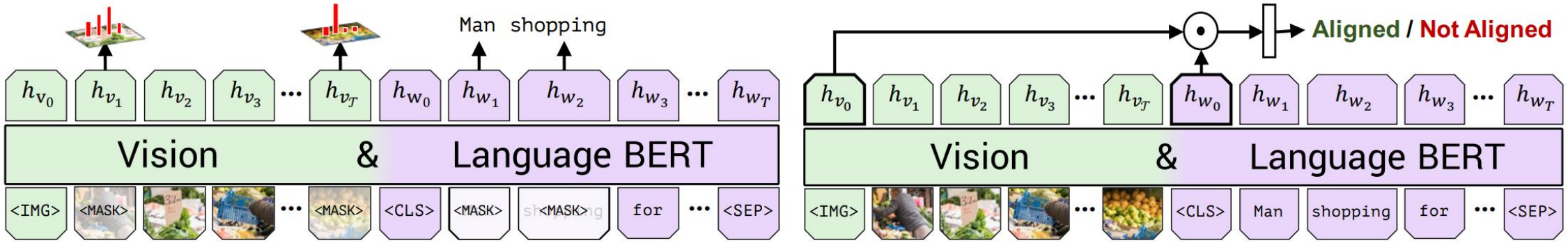
- Text stream has more processing before interacting with visual features.
- Initialize linguistic stream with BERT BASE.
- Use Faster R-CNN pretrained on Visual Genome dataset.





Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." arXiv:1908.02265 (2019).

ViLBERT Training Tasks and Objectives



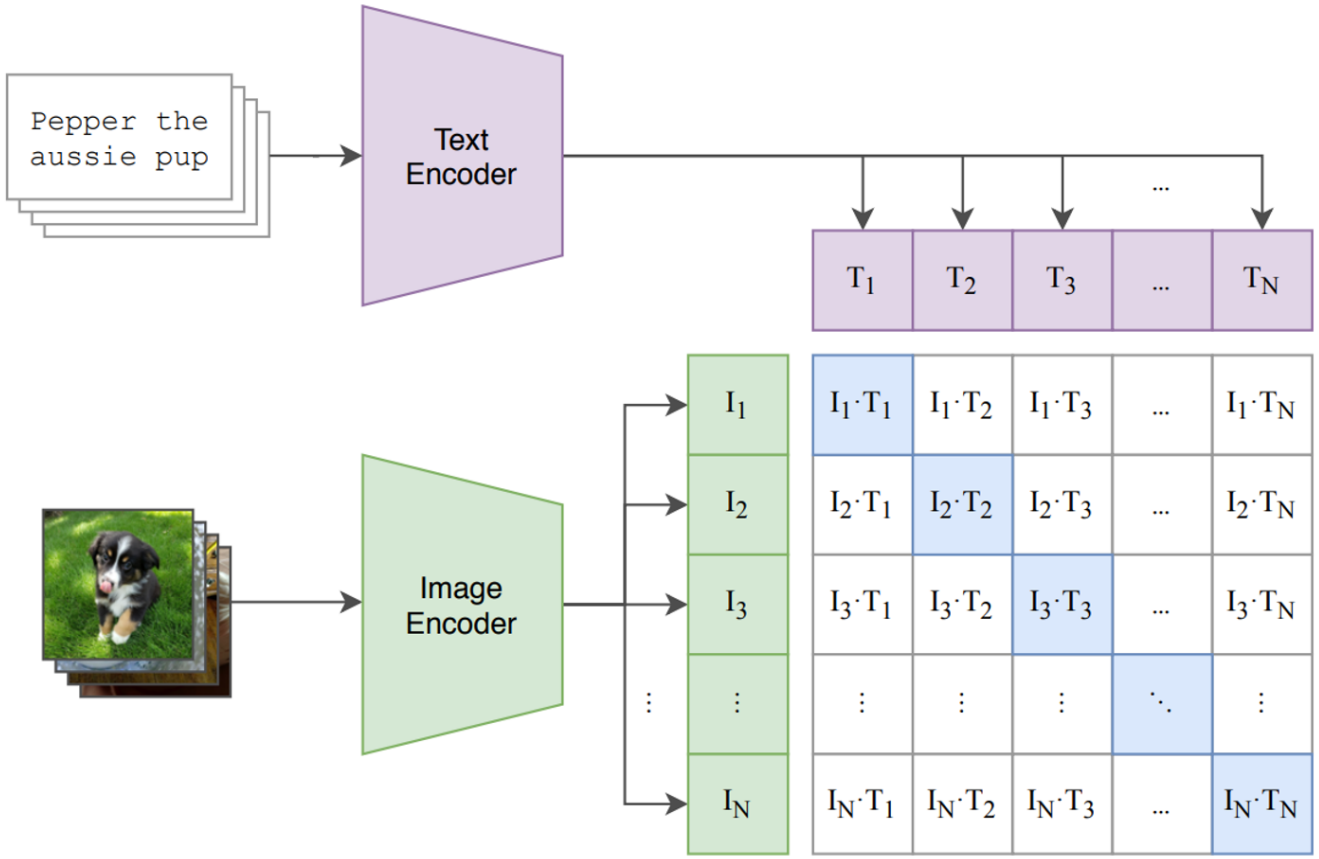
(a) Masked multi-modal learning (b) Multi-modal alignment prediction

- Train on Conceptual Captions (~3.3M images) to learn visual grounding.
- Masked multi-modal learning: reconstruct image region categories or words for masked inputs given the observed inputs.
- Multi-modal alignment prediction: predict whether or not the caption describes the image content.

Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." arXiv:1908.02265 (2019).

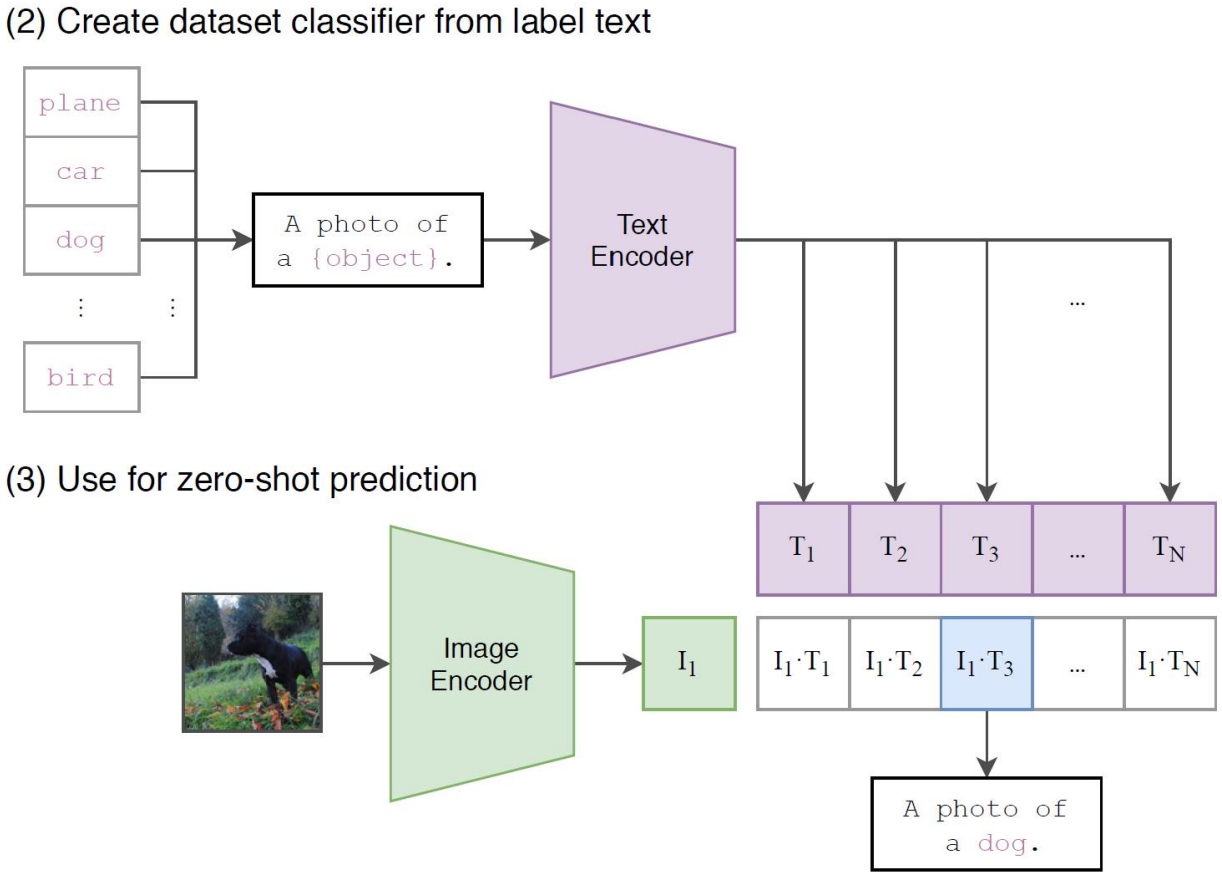
CLIP (Contrastive Language-Image Pre-training)

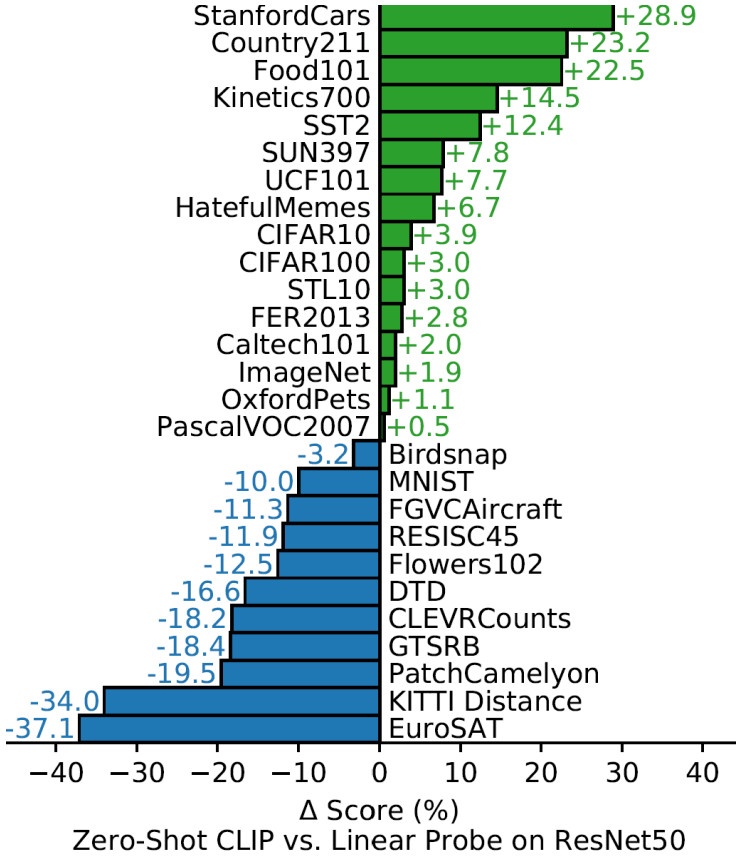
- Pre-trained using WebImageText (WIT) 400M (image, text) pairs.
- Text encoder is a 12L Transformer.
- 5 ResNets: ResNet-50, ResNet-101, RN50x4, RN50x16, and RN50x64
- 3 ViTs: ViT-B/32, a ViT-B/16, and a ViT-L/14
- Maximize cos-sim of the image and text embeddings of N real pairs in the batch
- Minimize cos-sim of the embeddings of the N × N − N incorrect pairings.
- Tested on 30+ CV tasks like OCR, action recognition in videos, geo-localization,…
- 0-shot CLIP is often ≡ fully supervised baseline



Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *ICML*, pp. 8748-8763. PmLR, 2021.

CLIP (Contrastive Language-Image Pre-training)





A zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

Classification using CLIP

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels

- ✓ a photo of **guacamole**, a type of food.
- × a photo of **ceviche**, a type of food.
- × a photo of **edamame**, a type of food.
- × a photo of **tuna tartare**, a type of food.
- × a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397

- ✓ a photo of a **television studio**.
- × a photo of a **podium indoor**.
- × a photo of a **conference room**.
- × a photo of a **lecture room**.
- × a photo of a **control room**.

YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23

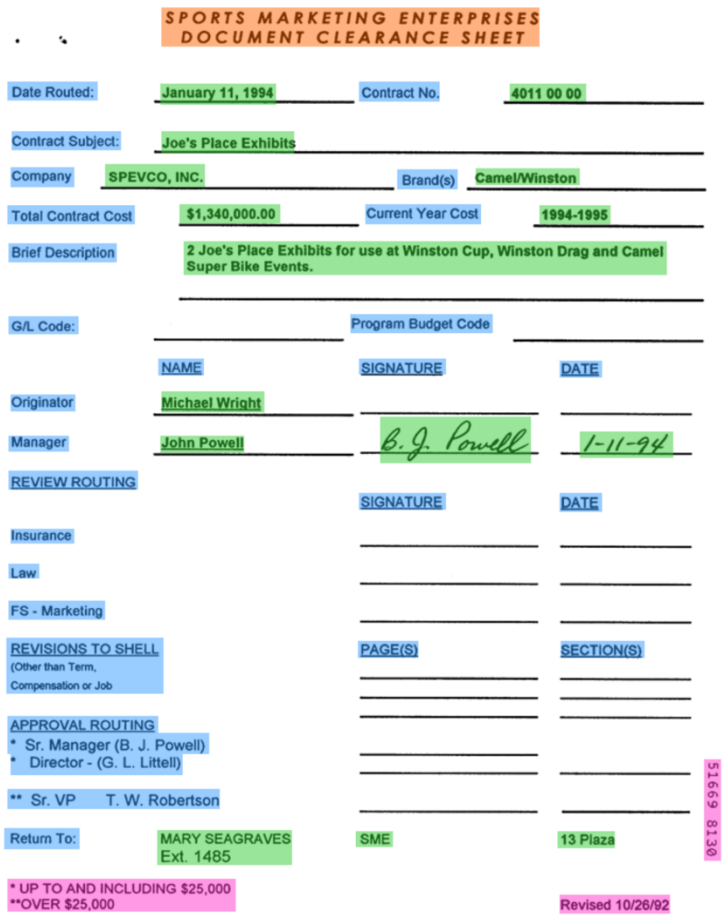
- ✓ a photo of a **airplane**.
- × a photo of a **bird**.
- × a photo of a **bear**.
- × a photo of a **giraffe**.
- × a photo of a **car**.

EUROSAT

annual crop land (12.9%) Ranked 4 out of 10

- × a centered satellite photo of **permanent crop land**.
- × a centered satellite photo of **pasture land**.
- × a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- × a centered satellite photo of **brushland or shrubland**.

Visually-rich Document Understanding



SPORTS MARKETING ENTERPRISES
DOCUMENT CLEARANCE SHEET

Date Routed: January 11, 1994 Contract No. 4011 00 00

Contract Subject: Joe's Place Exhibits

Company SPEVCO, INC. Brand(s) Camel/Winston

Total Contract Cost $1,340,000.00 Current Year Cost 1994-1995

Brief Description 2 Joe's Place Exhibits for use at Winston Cup, Winston Drag and Camel Super Bike Events.

G/L Code: Program Budget Code

	NAME	SIGNATURE	DATE
Originator	Michael Wright		
Manager	John Powell	B. J. Powell	1-11-94

REVIEW ROUTING

	SIGNATURE	DATE
Insurance		
Law		
FS - Marketing		

REVISIONS TO SHELL (Other than Term, Compensation or Job	PAGE(S)	SECTION(S)

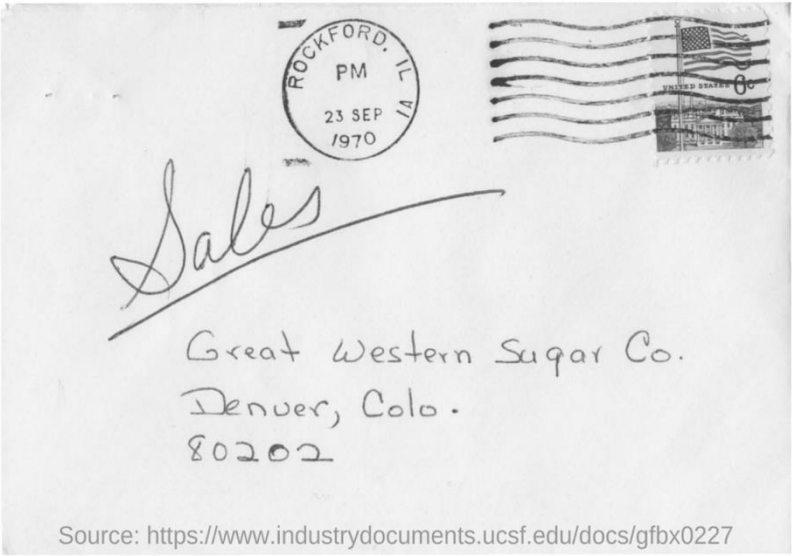
APPROVAL ROUTING
* Sr. Manager (B. J. Powell)
* Director - (G. L. Littell)
** Sr. VP T. W. Robertson

Return To: MARY SEAGRAVES Ext. 1485 SME 13 Plaza

51669 8130

* UP TO AND INCLUDING $25,000
**OVER $25,000

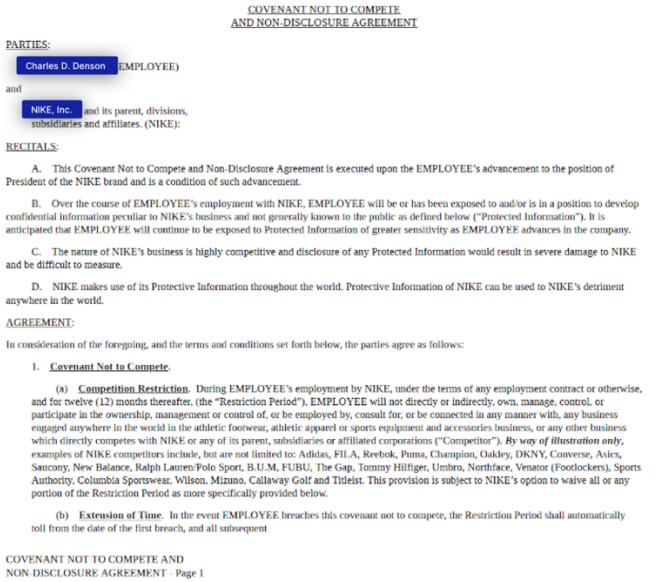
Revised 10/26/92



Source: https://www.industrydocuments.ucsf.edu/docs/gfbx0227

Q: Mention the ZIP code written?
A: 80202
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970
Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

First page:



COVENANT NOT TO COMPETE
AND NON-DISCLOSURE AGREEMENT

PARTIES:

Charles D. Denson EMPLOYEE)

and

NIKE, Inc. and its parent, divisions, subsidiaries and affiliates. (NIKE):

RECITALS:

A. This Covenant Not to Compete and Non-Disclosure Agreement is executed upon the EMPLOYEE's advancement to the position of President of the NIKE brand and is a condition of such advancement.

B. Over the course of EMPLOYEE's employment with NIKE, EMPLOYEE will be or has been exposed to and/or is in a position to develop confidential information peculiar to NIKE's business and not generally known to the public as defined below ("Protected Information"). It is anticipated that EMPLOYEE will continue to be exposed to Protected Information of greater sensitivity as EMPLOYEE advances in the company.

C. The nature of NIKE's business is highly competitive and disclosure of any Protected Information would result in severe damage to NIKE and be difficult to measure.

D. NIKE makes use of its Protective Information throughout the world. Protective Information of NIKE can be used to NIKE's detriment anywhere in the world.

AGREEMENT:

In consideration of the foregoing, and the terms and conditions set forth below, the parties agree as follows:

1. **Covenant Not to Compete.**

(a) **Competition Restriction.** During EMPLOYEE's employment by NIKE, under the terms of any employment contract or otherwise, and for twelve (12) months thereafter, (the "Restriction Period"), EMPLOYEE will not directly or indirectly, own, manage, control, or participate in the ownership, management or control of, or be employed by, consult for, or be connected in any manner with, any business engaged anywhere in the world in the athletic footwear, athletic apparel or sports equipment and accessories business, or any other business which directly competes with NIKE or any of its parent, subsidiaries or affiliated corporations ("Competitor"). ***By way of illustration only,*** examples of NIKE competitors include, but are not limited to: Adidas, FILA, Reebok, Puma, Champion, Oakley, DKNY, Converse, Asics, Saucony, New Balance, Ralph Lauren/Polo Sport, B.U.M, FUBU, The Gap, Tommy Hilfiger, Umbro, Northface, Venator (Footlockers), Sports Authority, Columbia Sportswear, Wilson, Mizuno, Callaway Golf and Titleist. This provision is subject to NIKE's option to waive all or any portion of the Restriction Period as more specifically provided below.

(b) **Extension of Time.** In the event EMPLOYEE breaches this covenant not to compete, the Restriction Period shall automatically toll from the date of the first breach, and all subsequent

COVENANT NOT TO COMPETE AND
NON-DISCLOSURE AGREEMENT - Page 1

Signature page:

(d) **Applicable Law/Jurisdiction.** This Agreement, and EMPLOYEE's employment hereunder, shall be construed according to the laws of the State of Oregon. EMPLOYEE further hereby submits to the jurisdiction of, and agrees that exclusive jurisdiction over and venue for any action or proceeding arising out of or relating to this Agreement shall lie in the state and federal courts located in Oregon.

EMPLOYEE

By: /s/ Charles D. Denson

Name: Charles D. Denson
Title: President, NIKE Brand

DATE: 3.26.01

NIKE, Inc.

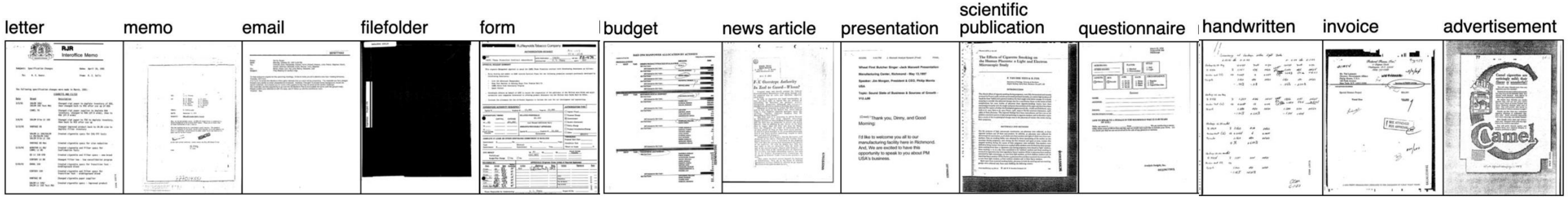
By: /s/ PHILIP H. KNIGHT

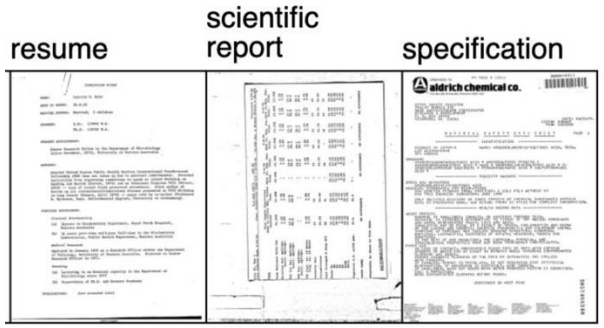
Name: Philip H. Knight
Title: President & CEO

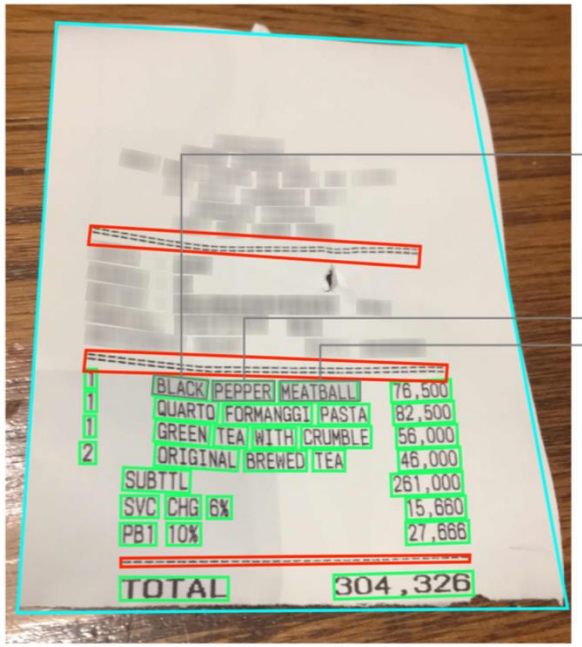
COVENANT NOT TO COMPETE AND
NON-DISCLOSURE AGREEMENT - Page 4

Xu, Yang, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu et al. "LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding." ACL, pp. 2579-2591. 2021.

Visually-rich Document Understanding



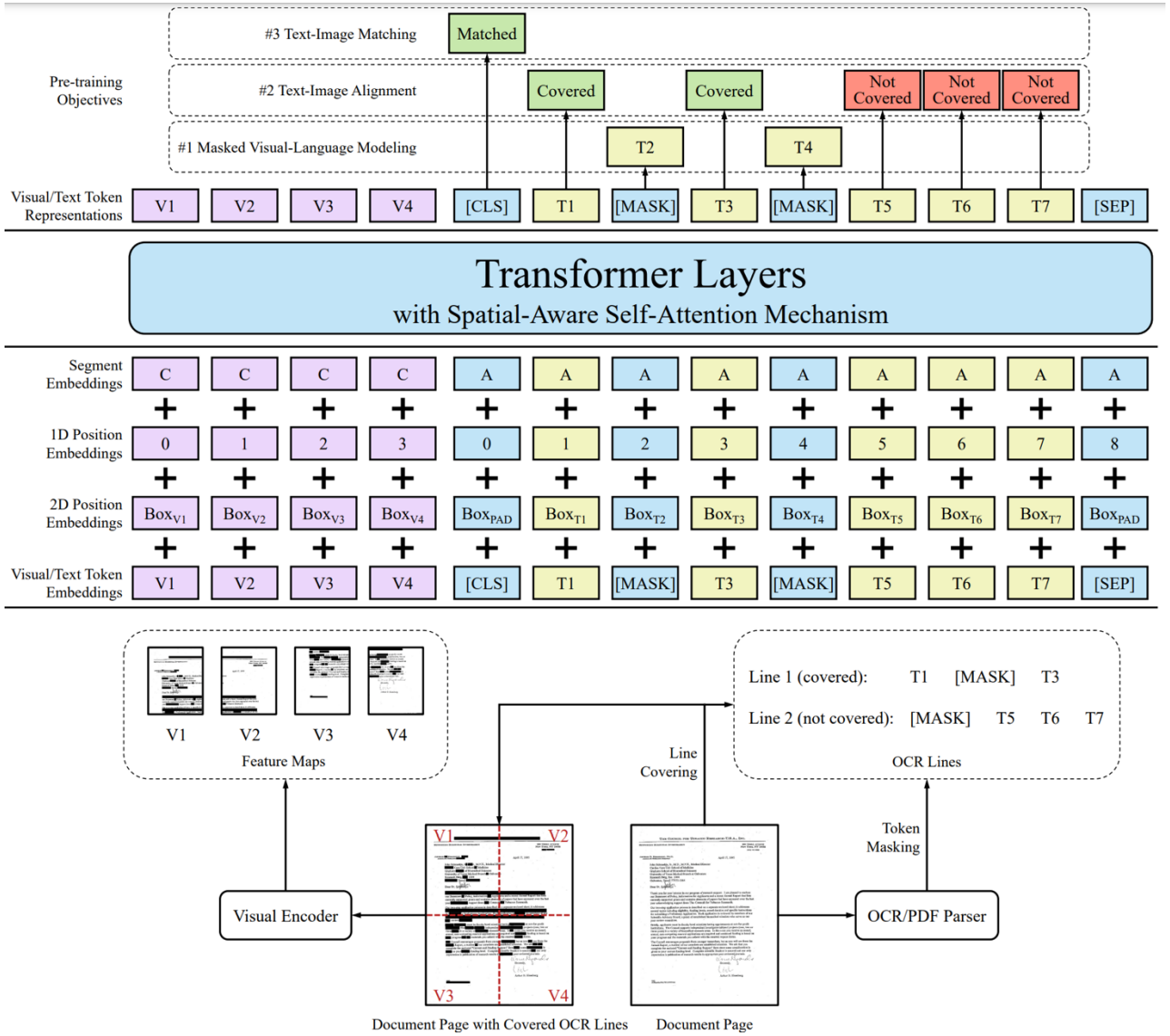





```
...
{
    "words":[
        {
         "quad":{"x2":  837,"y3": 2260,"x3":  829,"y4": 2238,
                 "x1":  625,"y1": 2156,"x4":  617,"y2": 2177},
         "is_key": 0,
         "row_id": 1631316,
         "text": "BLACK"
        },
        {
         "quad":{"x2": 1105,"y3": 2261,"x3": 1105,"y4": 2261,
                 "x1":  857,"y1": 2180,"x4":  857,"y2": 2180},
         "is_key": 0,
         "row_id": 1631316,
         "text": "PEPPER"
        },
        {
         "quad":{"x2": 1441,"y3": 2264,"x3": 1443,"y4": 2269,
                 "x1": 1125,"y1": 2184,"x4": 1126,"y2": 2179},
         "is_key": 0,
         "row_id": 1631316,
         "text": "MEATBALL"
        }
    ],
    "category": "menu.nm",
    "group_id": 3
}
...
```

LayoutLMv2 Architecture

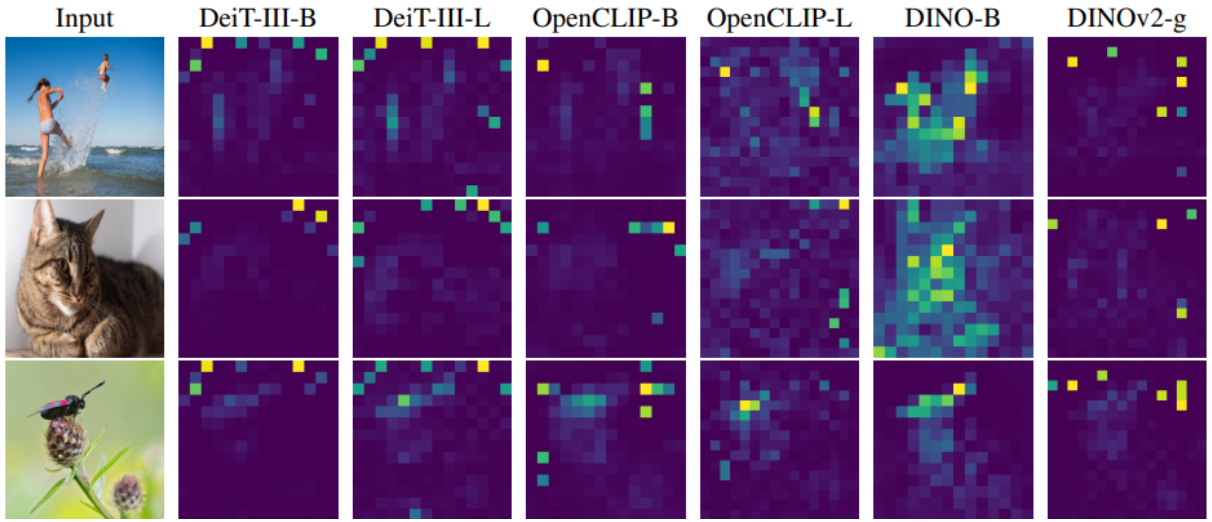
- Text: Initialized using UniLMv2
- Visual encoder
  - ResNeXt-FPN architecture with MaskRCNN backbone.
  - Use output feature map (W=H=7).
- Embed spatial layout of token bounding boxes from OCR results
  - Concat(PosEmb2Dx($x_{\min}, x_{\max}$, width), PosEmb2Dy($y_{\min}, y_{\max}$, height))
- LayoutLMv2: Base (12L, 200M), Large (24L, 426M)
- 3 tasks: MVLM, TIA, TIM
- Dataset: 11M scanned docs. Text OCR: Microsoft Read API

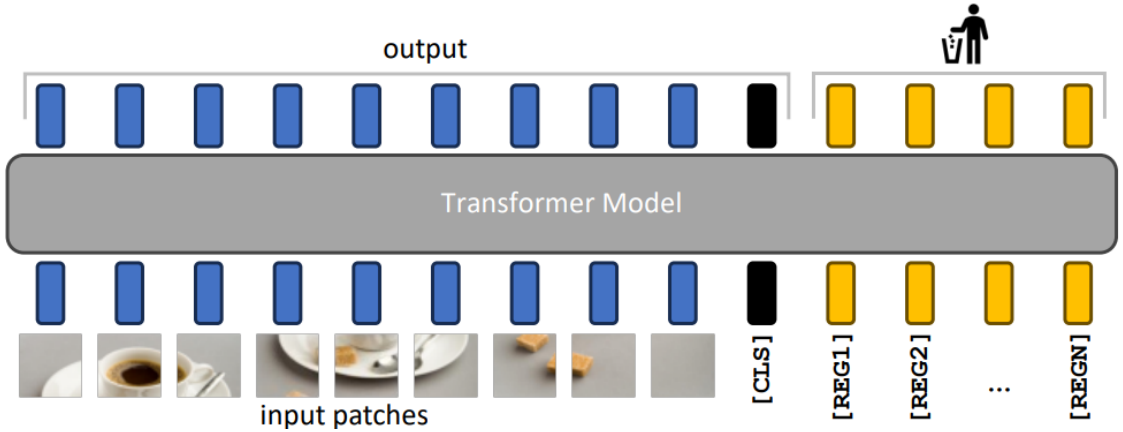


Xu, Yang, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu et al. "LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding." ACL, pp. 2579-2591. 2021.

ViT with Registers

- Several ViT models exhibit peaky outliers in attention maps.
- Artifacts
  - 2.37% tokens with ~10x higher norm at output
  - appear around the middle layers of ViT
  - appear after a sufficiently long training of a sufficiently big transformer.
  - appear in patches similar to their neighbors, meaning patches that convey little additional information
- Append register tokens to input
  - Improved performance in dense prediction and object discovery.
  - Removes artifacts leading to smooth attention maps.

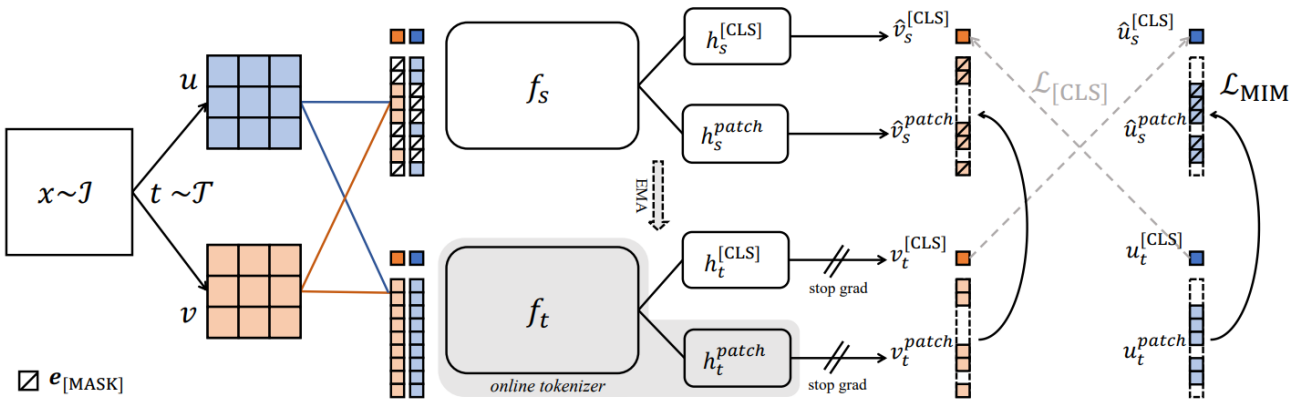




Darcet, Timothée, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. "Vision Transformers Need Registers." In ICLR. 2024

DINOv3

- Vision encoder: ViT-7B distilled to ViT (21M to 840M) and ConvNeXt (29M to 198M).
- Self-supervision dataset (17B images).
- Learning Objectives
  - Image-level objective as in DINOv2
  - Patch-level latent reconstruction iBOT.
  - Koleo regularizer encourages features within a batch to spread uniformly in the space.
- Variant of RoPE
- 2 global crops and 8 local crops per image.
- Post-Training
  - Resolution Scaling: 256 → 512, 768
  - Distillation: 7B teacher instead of EMA teacher
  - Aligning DINOv3 with Text: CLIP style.



Teacher model	DINOv3
Backbone	ViT-7B
#Params	6.7B
#Blocks	40
Patch Size	16
Pos. Embeddings	RoPE
Registers	4
Embed. Dim.	4096
FFN Type	SwiGLU
FFN Hidden Dim.	8192
Attn. Heads	32
Attn. Heads Dim.	128

$$\theta_t \leftarrow \tau \cdot \theta_t + (1 - \tau) \cdot \theta_s$$

$$\mathcal{L}_{DINO} = -\sum p_t \log p_s$$

$$\mathcal{L}_{iBOT} = -\sum_i p_{ti} \log p_{si}$$

$$L_{\text{KoLeo}} = -\frac{1}{B} \sum_{i=1}^{B} \log(\|x_i - x_{NN(i)}\| + \epsilon)$$

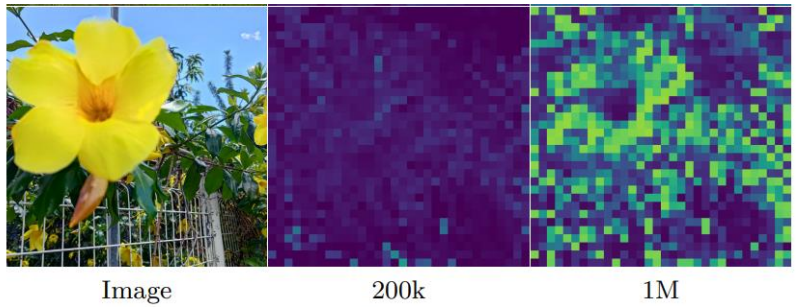
$$\mathcal{L}_{\text{Pre}} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + 0.1 * \mathcal{L}_{\text{DKoleo}}$$

Siméoni, Oriane, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov et al. "Dinov3." arXiv:2508.10104 (2025).

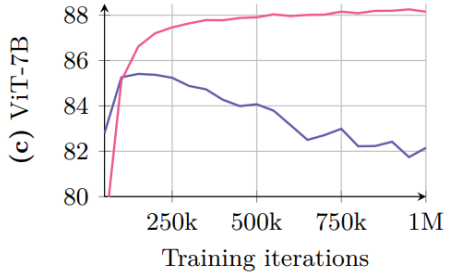
Gram Anchoring (Regularization for Dense Features) in DinoV3

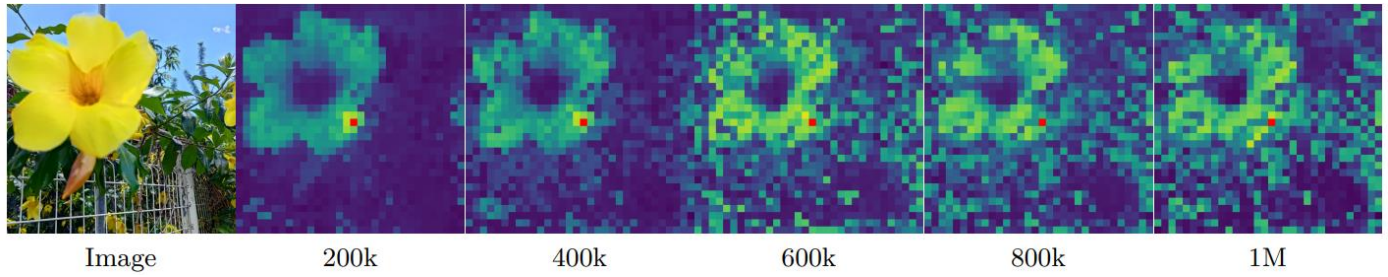
- Prolonged training leads to improvements on global benchmarks but degradation on dense tasks.
- Loss of Patch-Level Consistency Over Training.
- Gram Anchoring Objective.
  - $L_{Gram} = \left|\left|X_S.X_S^T - X_G.X_G^T\right|\right|_F^2$
  - Only on global crops and after 1M iters.
- Update Gram teacher every 10k iterations at which Gram teacher becomes same as main EMA teacher
- L_{Href}: higher-resolution features for the Gram objective

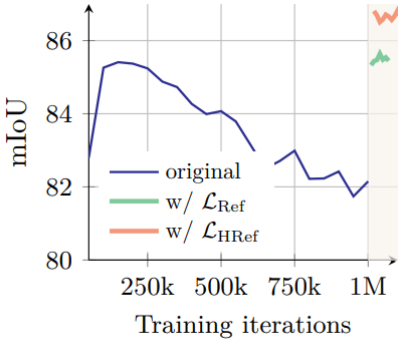
$$\mathcal{L}_{\text{Ref}} = w_{\text{D}}\mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + w_{\text{DK}}\mathcal{L}_{\text{DKoleo}} + w_{\text{Gram}}\mathcal{L}_{\text{Gram}}$$



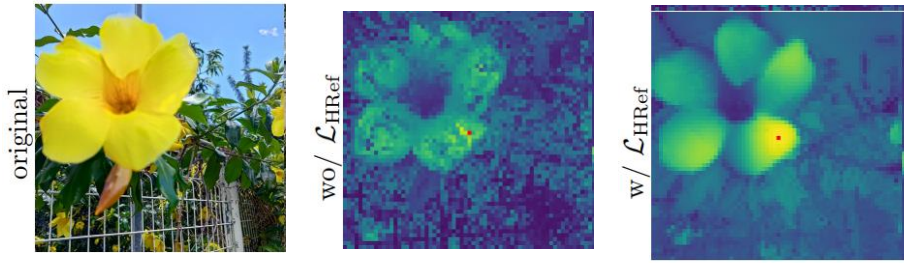
(a) Cosine similarities between the CLS and output patches.







(a) VOC



Siméoni, Oriane, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov et al. "Dinov3." arXiv:2508.10104 (2025).

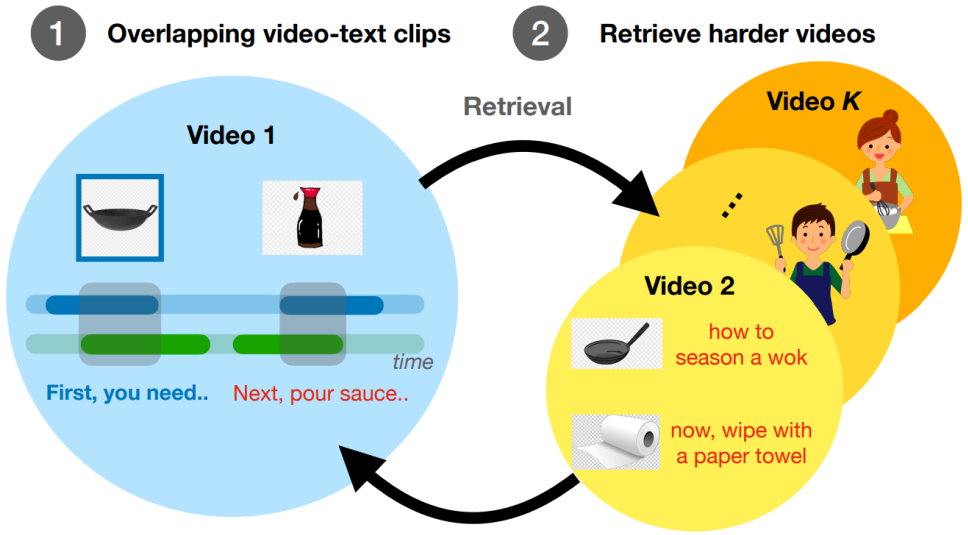
Video tasks

- Text→Video Retrieval
  - Given text and a collection of videos, find relevant ones.
- Multiple-choice VideoQA.
  - Given video, a question and multiple candidate answers, choose the best one.
- Action Segmentation/Action Step Localization
  - Assign each frame with one of the pre-defined labels to separate meaningful segments of videos.
  - Similar to sequence labeling (e.g. NER) in NLP.

Xu, Hu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer, "VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding." In EMNLP, pp. 6787-6800. 2021.

VideoCLIP architecture and pretraining

- Contrastive approach
  - Loosely temporally overlapping positive video-text pairs, instead of enforcing strict start/end timestamp overlap.
  - Hard negatives using nearest neighbor retrieval that uses video clusters to form batches with mutually harder videos.
- BERT-base-uncased for both video (6L) and text (12L).
- Video: frozen pretrained CNN, projected to video tokens using a MLP layer.
- Average pooling over the sequence of tokens for video and text.
- Pretraining data: HowTo100M



VideoCLIP: Contrastive learning with hard-retrieved negatives and overlapping positives for video-text pre-training.

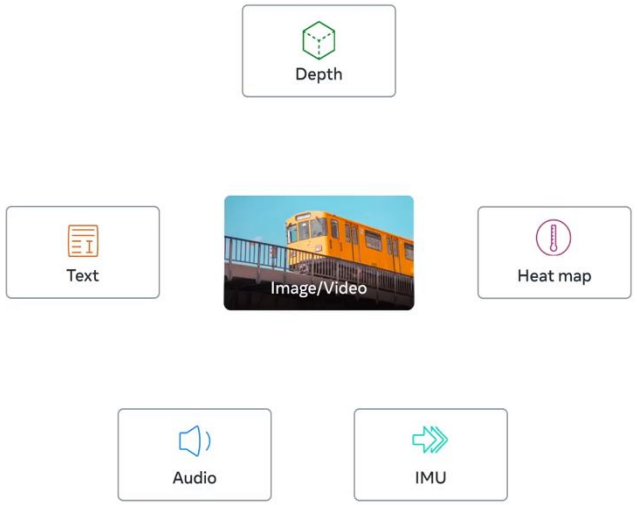
$$\mathcal{L} = -\sum_{(v,t)\in B}\Big(\log \mathrm{NCE}(z_v, z_t) + \log \mathrm{NCE}(z_t, z_v)\Big)$$

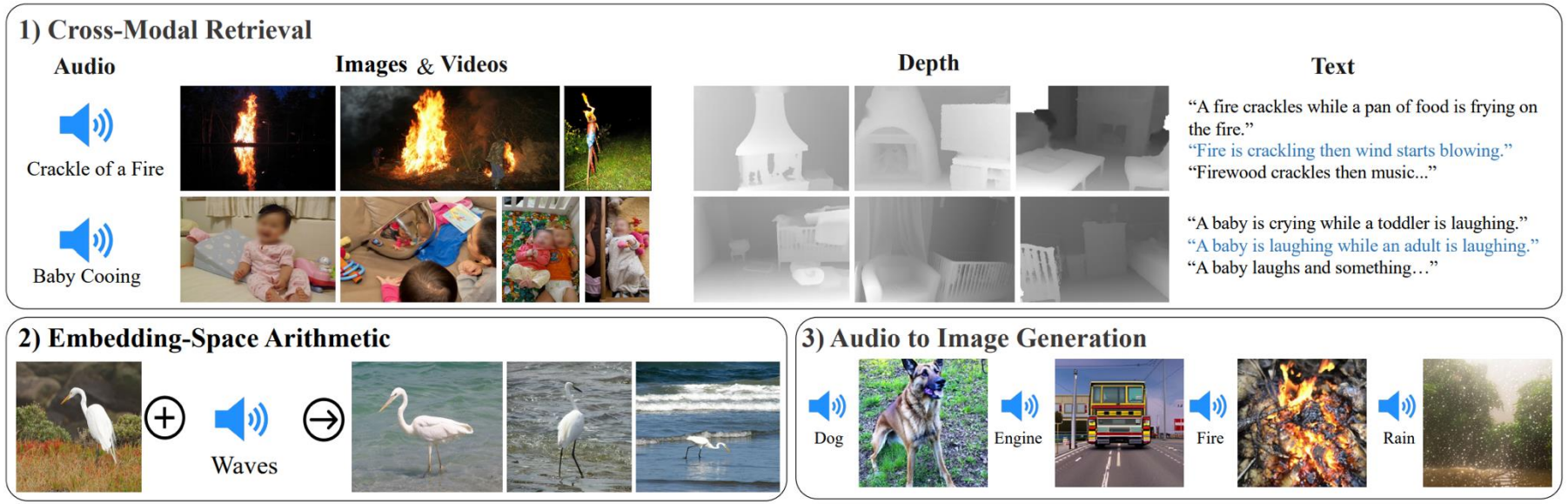
$$\mathrm{NCE}(z_v, z_t) = \frac{\exp\left(z_v \cdot z_t^+/\tau\right)}{\sum_{z\in\{z_t^+, z_t^-\}}\exp\left(z_v \cdot z/\tau\right)}$$

Xu, Hu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. "VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding." In EMNLP, pp. 6787-6800. 2021.

ImageBind

- An image of a beach can remind us of the sound of waves, the texture of the sand, a breeze, or even inspire a poem.
- Aligns 6 modalities' embedding into a common space: images, text, audio, depth, thermal, and Inertial Measurement Unit (IMU).
- Image-paired data is sufficient to bind the modalities together.

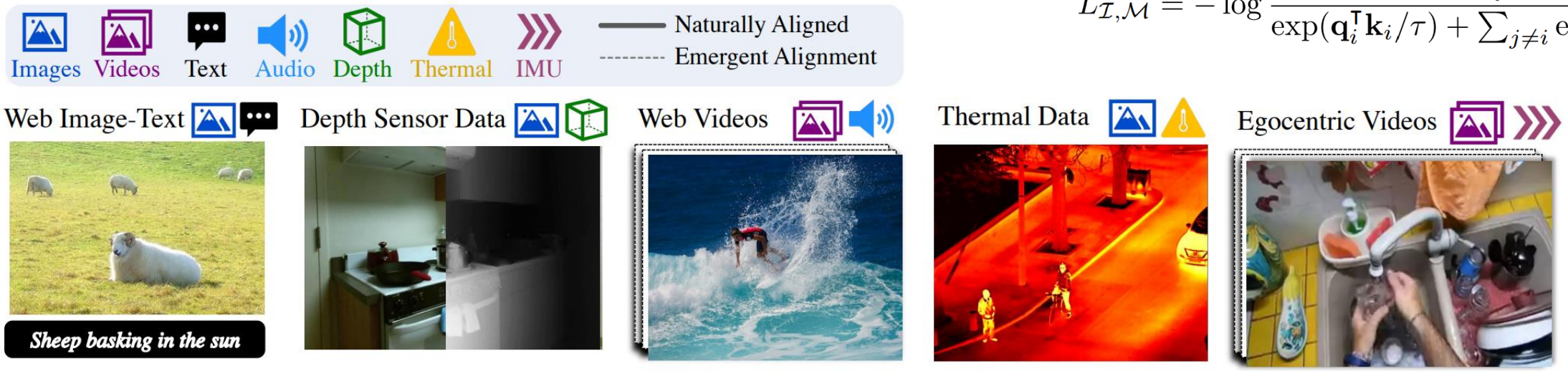




Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. "Imagebind: One embedding space to bind them all." In CVPR, pp. 15180-15190. 2023.

ImageBind Pretraining

$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$



- $q_i = f(I_i)$ and $k_i = g(M_i)$ where f and g are deep networks.
- InfoNCE loss. Symmetric loss $L_{I,M} + L_{M,I}$
- ViT-H 630M params; text encoders (302M params) from OpenCLIP (frozen)
- Same encoder for images+videos. Treat videos as multi-frame images.

- Datasets:
  - (video, audio) pairs from Audioset
  - (image, depth) pairs from SUN RGB-D
  - (image, thermal) pairs from LLVIP
  - (video, IMU) pairs from Ego4D
  - (image, text) pairs from large-scale web data.

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. "Imagebind: One embedding space to bind them all." In CVPR, pp. 15180-15190. 2023.

Summary

- ViT
- VisualBERT
- VilBERT
- CLIP
- LayoutLMv2
- ViT with registers
- DinoV3
- DoRA
- VideoCLIP
- ImageBind

Thanks!

- HomePage: https://sites.google.com/view/manishg/
- Google Scholar: https://scholar.google.co.in/citations?user=eX9PSu0AAAAJ
- LinkedIn: http://aka.ms/manishgupta
- YouTube (Data Science Gems): https://www.youtube.com/@dlByManish