Foundation Models

Paolo Rota
paolo.rota@unitn.it

Marco Bronzini
marco.bronzini-1@unitn.it

University of Trento
November 7, 2025

Overview

Friday, October 10	**Lab 1**	• Metrics for Natural Language Generation (NLG) • Real-world use cases (MT, QA with RAG)
Wednesday, October 15	**Lab 2**	• Extract latent features from LLM embeddings by training a classification model
Friday, November 7	**Lab 3**	• Generate video captions using: a) CLIP-inspired models (**Co**ntrastive **Ca**ptioners) b) Vision-Language Models (**VLMs**)
Wednesday, November 12	**Lab 4**	• Agents using *LangChain*

Multi-modal setting: **Video Captioning**

VIDEO: collection of images

e.g., 30 frames per second (FPS)

Generative models

TEXT: Short textual description

A person is changing a car tire.

Approach: ***Co**ntrastive **Ca**ptioners*

It is a unified model trained from scratch on image-text pairs, which combines:

1. **contrastive learning** (e.g., CLIP)
2. **caption generation** (e.g., GPT-style decoding).

TRAINING

Vision Encoder

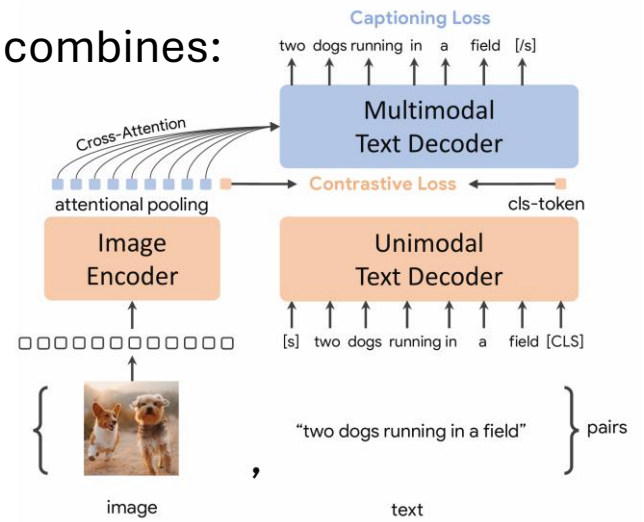
A transformer-based visual backbone	It extracts visual tokens

Text Encoder/Decoder:

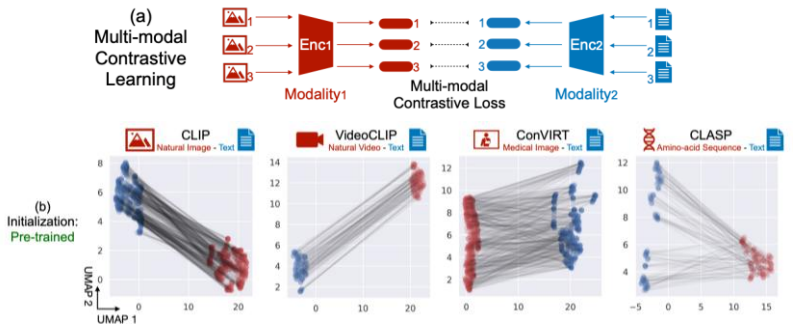
A transformer that serves both as a:	(1) contrastive encoder for alignment	(2) generative decoder for captioning

Training

Contrastive loss aligns image and text embeddings (like CLIP).	**Captioning loss** (cross-entropy) trains the model to generate descriptive text.



CoCa: Contrastive Captioners are Image-Text Foundation Models



Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning

Approach: **V**ision-**L**anguage **M**odels (VLMs)

A generative model able to understand images and text inputs using:

a) a vision encoder

b) a large language model

PRE-TRAINED

TRAINING

Vision Encoder

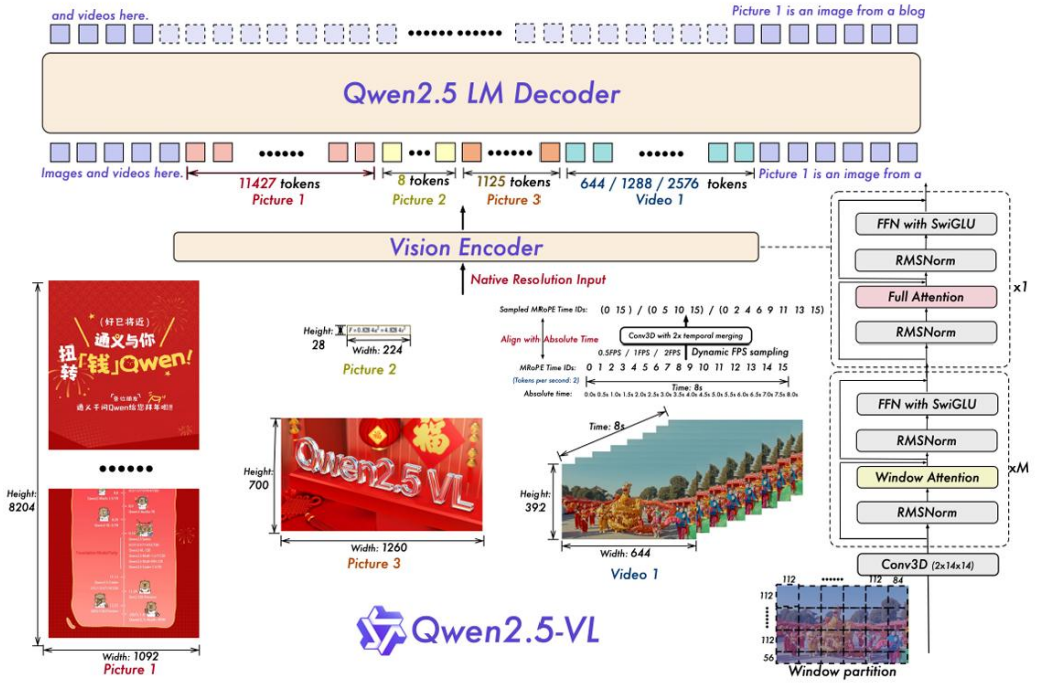
Pretrained visual backbone (e.g., CLIP-ViT, EVA, Swin)	It extracts visual tokens

Large Language Model

A pretrained textual model (e.g., T5, GPT, Llama)	It can be frozen or fine-tuned

Adapter / Connector

A lightweight module (e.g., MLP or cross-attention)	It maps vision features to language token space



Qwen2.5-VL Technical Report

Comparison

Aspect	***Contrastive Captioners*** (CoCa)	**Vision-Language Models** (VLMs)
Vision Encoder	Trained from scratch	Pretrained (e.g., CLIP-ViT)
Language Model	Trained from scratch	Pretrained (e.g., LLaMA)
Connector	None (joint fusion)	Explicit projection
Training Objective	Contrastive + Captioning (joint)	Alignment + instruction tuning
Integration	End-to-end unified	Modular (vision + language + adapter)
Flexibility	Less modular but efficient	Highly modular and extensible
Compute Efficiency	Heavy one-time training	Cheaper fine-tuning with reused parts

Practical Tutorial

Please, open the following notebook:
https://colab.research.google.com/github/saturnMars/FM_2025/blob/main/Lab3_video.ipynb

CoCa versus **VLMs**
using NLG metrics

QUIZ

MENTI **(2153 2489)**

https://www.menti.com/al4ika5zijuu