**FIRST PASS OF THE PAPER**

1. Paper overview:

* Title: Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at any Resolution. It was given by **QwenTeam AlibabaGroup**
* Category: This is the research paper ,which are give the description of the advance vision-language model.
* Idea: 1st of all how it advance then former Vision transformer, so the former ViT are process the image which are come under their fixed threshold of the resolution of the image, but with this limitation problem is that, when higher resolution image are get as a input so it give partial knowledge of that image, means information loss are their, because the limit is exceed as per threshold so model performance is decrease so that it give the inconsistence results. So for that we introduce the Qwen2-VL which are perform the dynamic resolution ,means there is no limit of the resolution ,so when we provide the higher resolution image so it easily process this image and give the best result.

1. M-RoPE (Multimodal Rotary Position Embedding):

* The another demerits of the former ViT its perform static position embedding means if we provide the video as a input so it not track the movable object’s , so for resolving this problem we use the multimodal rotary position embedding , after using this we trace the object location,even if the objects are movable and its also work for combined modalities(text+image) match this by given text and produce the result.
* Multimodal Fusion: This will integrate the positional relationship of the text, images and videos. So that it generate the combined and coherent representation of the overall relation.
* After simulate this feature in the model, so it increase the inference length of the model (means increase the input size).

1. Scalability:

* Model have 3 possible size’s available 2B,8B and 72B parameters where B is Billion.  
  Now this size are define as per the model complexity and their working , means if the model are less complex and just working to process the text so in that we use 2B parameters, because parameters are playing a vary crucial role during the model training.
* Similarly if the model are process the multimodal like image+text so in that case we use 8B model.
* If the model is too complex and its process the image, text and video(>=20min) so for that case we prefer 72B modal.

1. **Approach:**

* visual encoder→ cross-modal connector→ LLM
* Now let we define in short , the working of components are used in above pipeline.
* Visual encoder: This is the encoder which are convert the images and videos in format which are model readable. Means it convert in embedding vectors.
* Cross-modal connector: This step are combine the textual tokens and visual tokens which are establish the relation in btw text and images and videos using the M-RoPE.
* LLM: This component of the pipeline are playing a major role, The coherent representation which are generate by the 2nd step is provide as input to the LLM, and LLM process that data and generate the intelligent outputs.
* Architecture: This model use the combination of ViT and LLM(Large Language model)
* Training: During the training 1.4 trillion tokens are used in that image-text pairs and videos are used.
* Benchmarks: This model are give the better performance as compare to others in document reading , visual question answering and video comprehension tasks.

1. **Conclusion:**

* It understood the multimodal data(text + image+ video) just like a human-perception.
* It set the new standards for vision-language tasks. Which we discuss above like dynamic resolution and Multimodal rotary position embedding.

**Now we Analyse the five Cs:**

1. Category: The category of this paper is “**A description of a research prototype”.**How?

Because this paper are explain the feature of new model Qwen2-VL. Like

* Dynamic Resolution Processing.
* Multimodal Rotary Position Embedding (M-RoPE)

This is new architectural enhancements , which is the part of prototype.

1. Context:

* Related Papers:

GPT-4o(open AI) [2410.21276](https://arxiv.org/pdf/2410.21276) and claude3.5(Anthropic) , this is compressions of Qwen2-VL.

* Theoretical bases: **multimodal learning**, **transformer architecture**, **vision-text alignment**, **Naïve Dynamic Resolution**, **dynamic resolution and M-RoPE.**

1. Correctness: In that paper the assumption which we used , like alignment if text and images, transformer architecture and dynamic resolution and all , this all are valid.

The all assumptions are based on proven techniques which are already successfully used in multimodal tasks.

1. Contributions:

* Vision-Language Integration: Qwen2-VL are propose the new technique which are effectively combine text and image data. And this technique is **“Naïve Dynamic Resolution”.**  This is the mechanism, with help of the this model are efficiently process the high-resolution images and combine the image and text with handling image resolution.
* M-PoER- This technique are used to track the object location and so that model easily relate the text and objects, it help to generate the context of the image. This technique are basically used to track the movable object as well as static objects.

1. Clarity: Yes this paper are well written , because the things which are required for well written paper is .

* Problem Statement
* Methodology
* Results
* Language and structure: language are technical ,but well understable.

These all things in this paper are well define so that we conclude that the given paper are well written.

**SECOND PASS OF THE PAPER**

1st of all we give the brief introduction of this paper , problem in former technique’s , what new revolution is it and how they work it.

**Problem in former technique’s:**

In the previous approaches like Qwen-VL and ViT , these techniques are basically used for the static data(unmovable) and fixed resolution images, so that due to this limitations problem is that , if the user provide the higher resolution image and video as input so that , in case of images, there is high chance of information loss cause the higher resolution , model have certain threshold of the resolution so above then that its not give the expected information and all. Similarly in case of video as input ,so in the object is movable so there are need of image position so it relate the text according to the location. So for that case it also give unpredictable results.

**What new revolution idea’s:**

Now to solving this problem we introduce the two new features which reduce the former demerits of this model, and hence this model is called as “**Qwen2-VL”**. Let we introduce this techniques.

**For resolution:**

Naïve Dynamic Resolution Mechanism:

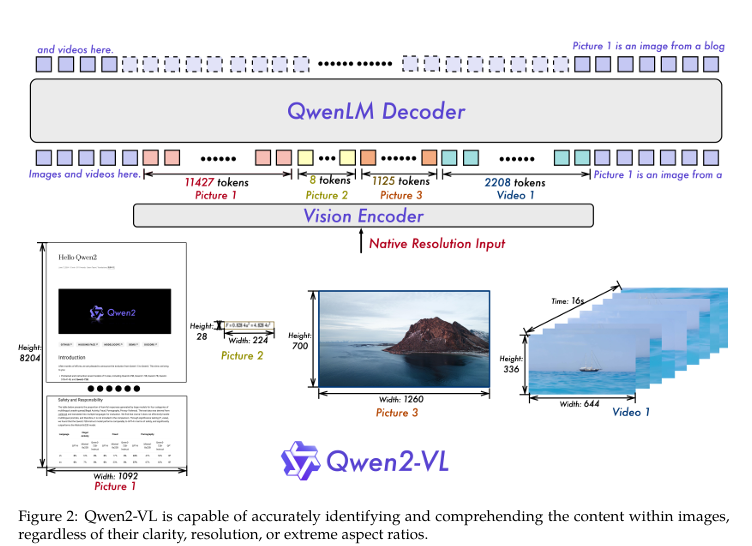
* It will process the image on their original size either its higher resolution or lower resolution.
* Using the MLP layer compress the tokens (part of the image) so that memory is save.

**For movable objects:**

Multimodal Rotary Position Embedding(M-RoPE):

* The working of that part is to understand the positional information(image start on which corner or part).
* In videos it merge the details of temporal(time-based) and spatial(space-based).
* Temporal : means to understand the changes with the time. Like we see the video and in that every frame is the snapshot of the moment. So the temporal detail analyse the changes in btw the frames.
* Spatial: means to understand the position and structure of the object in the space. Means give the every details like where object is stand , which direction is move and all about their location and structure.

**How it’s work:**

****

1. Naïve Dynamic Resolution Mechanism:

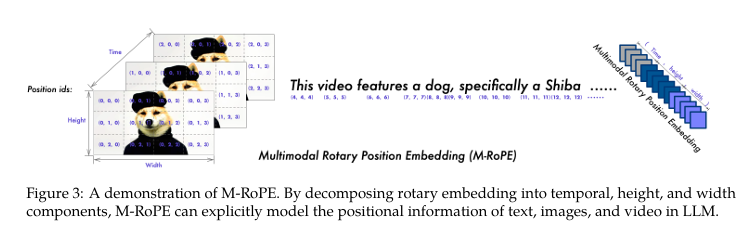
Step-1: 1st it will get the image as a input.

Step-2: After that divide the image in the patches using conv2 technique. After that convert the every patch into visual tokens and then store in the form of token embedding. Using vision encoder.  
Step-3: after that compress the visual tokens, because in the complex image there are lot of tokens which are create the memory overload.

So to optimise the memory we use MLP layer.  
step-4: After that arrange the tokens in a sequence, and this sequence give the vison transformer and it will generate the context and details of the image.

1. M-RoPE:

* It will understand the positional details which are critical for multimodal tasks(e.g., video question answering, navigation).
* It will perform the fusion of the temporal and spatial details which are useful to clear the relation of each modality.

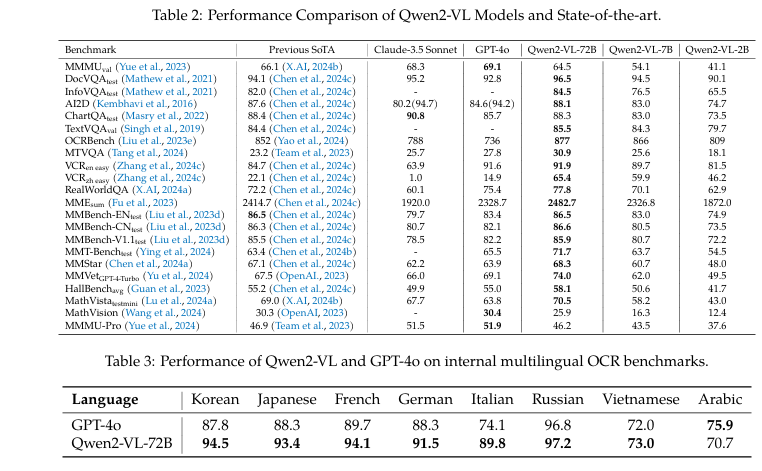


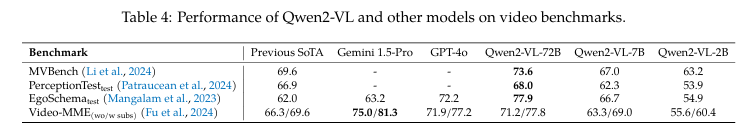
**Now How we train this modal:**

1. ViT: This is for understanding the relation in btw text and images
2. To train the complete model: for train the complete model we use Image-text pairs and video also we used, OCR(optical Character Recognition) and use the dataset of visual question-answering(VQA).
3. It was Train using special format ChatML ,so that it perform the multimodal conversations.

**Performance of the Model:**

1. DocVQA and InfoVQA: If we provide the document(image and text) of this model it will give the best result.
2. Multilingual OCR: It will understood the different types of languages in the images, like English, Chinese, Japanese, Arabic and French.
3. Video Understanding: Its capable understand the 20min + videos, its analyse the every frame in single flow, means not break in single frame.





**3 versions of the Model:**

1. Qwen2-VL-2B:It will useful for less complex or resource-limited device.
2. Qwen2-VL-7B: It will use for medium performance and cost balance.
3. Qwen2-VL-72B: This is the most powerful version , which handle the complex tasks.

Capability of the model: