Zalo AI Challenge / 2023

Team Name: T-Sharp

Members

Thang Huu Nguyen, Tuan Hau Tran

INPUT / OUTPUT

This methodology introduces an effective strategy for crafting advertising banners in the Vietnamese language through the leverage of generative AI capabilities. Our approach seamlessly integrates linguistic and visual models to enhance the system's adaptability, enabling it to accommodate various inputs and accurately generate the target image. The prompt builder adeptly transforms product descriptions into key information comprehensible to the text-to-image model. Ultimately, the image generator component takes charge of producing the advertising banner tailored to the specific product in focus.

SYSTEM ARCHITECTURE / PIPLELINE

The comprehensive workflow comprises two primary components: the Prompt builder operating atop **Qwen-7B**[1] and the Image generator running on the foundation of **Kandinsky-2.0**[2].

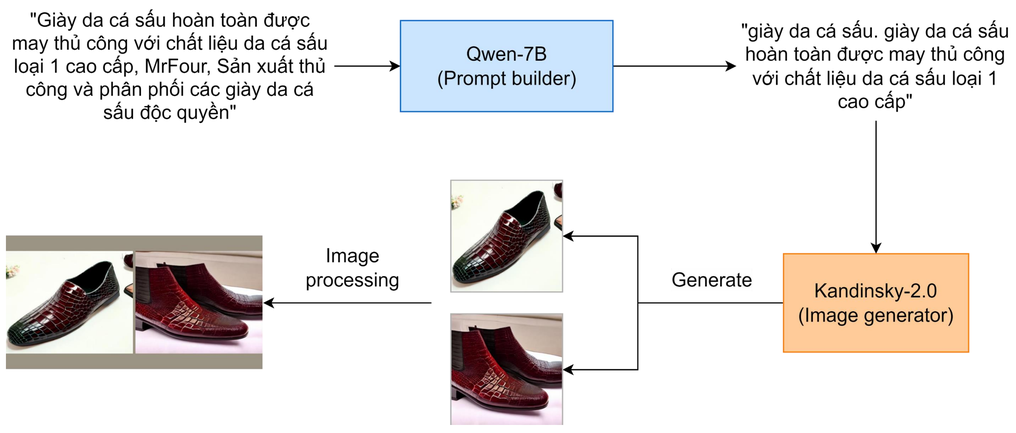

Figure 1: Advertising banner generation pineline

[1] VillaLabs, Qwen-7b-chat-vietnamese , https://huggingface.co/VillaLabs/Qwen-7b-chat-vietnamese
[2] Arseniy S., et al., Kandinsky_2.0 , https://huggingface.co/ai-forever/Kandinsky_2.0

Prompt builder

The prompt builder allows the pipeline to process diverse Vietnamese language contexts effectively, leveraging the substantial capabilities of **Qwen-7B-chat-vietnamse**, a LLM **fine-tuned on Vietnamese** by **VillaLabs**.

Qwen is built with architecture similar to LLaMA, with modifications:

- Using untied embedding
- Using rotary positional embedding
- No biases except for QKV in attention
- RMSNorm instead of LayerNorm
- SwiGLU instead of ReLU
- Adopting flash attention to accelerate training.

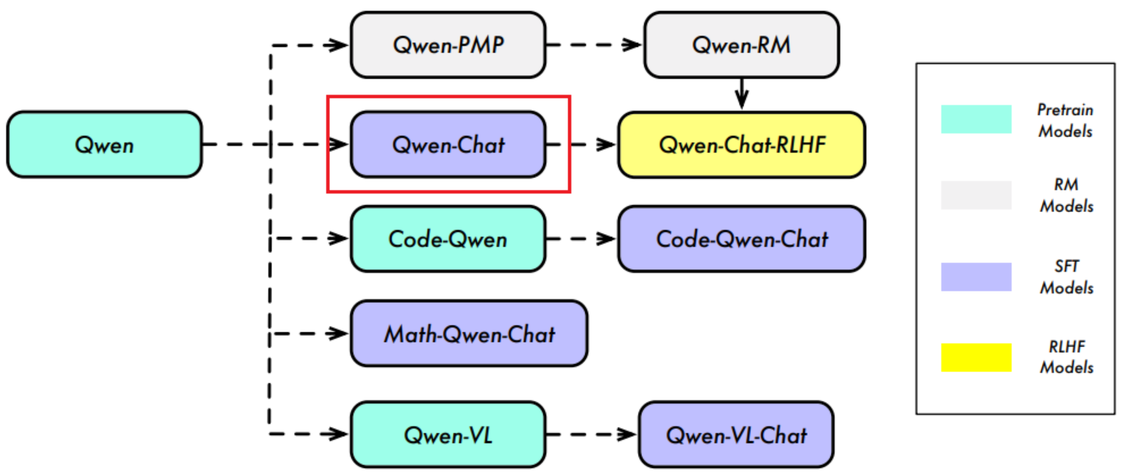

Figure 2: Model Lineage of the Qwen Series

# of Params	Hidden size	Heads	Layers
7B	4096	32	32

Table 1: Qwen model sizes, architectures

The main reason we selected Qwen-7B-chat-vietnamese is its capability to understand context with missing information, which other LLM models like Llama-2-7b-chat struggle with. Our prompt was built as follows:

"""You are a good assistant in extracting key information from ads in Vietnamese. Your task is to extract product categories from ads or infering product categories that are not directly mentioned in the content. Ads: "{ads}".
Your response must be in JSON format:
{"product_type": "", // product type mentioned in content, this should be noun, you infer product type if not mentioned
"description": "" // description of appearance including shape, colors, materials, details, features, reference images.} """

After receiving the response, we concatenate its values to generate prompt for the image generator.

Image generator

We propose **Kandinsky 2.0** as the image generator instead of other open-source models like Kandinsky 2.1, 2.2, Stable Diffusion 1.5, and SDXL. Its multilingual training, including Vietnamese, enables seamless alignment with the challenge's original caption.

Kandinsky 2.0 is based on an improved Latent Diffusion approach with several significant differences:

- Two multilingual text encoders, mT5-small (146M) and XLMR-CLIP (560M)
- Deployment of a custom UNet (1.2B parameters).
- Dynamic trash holding during the sampling process

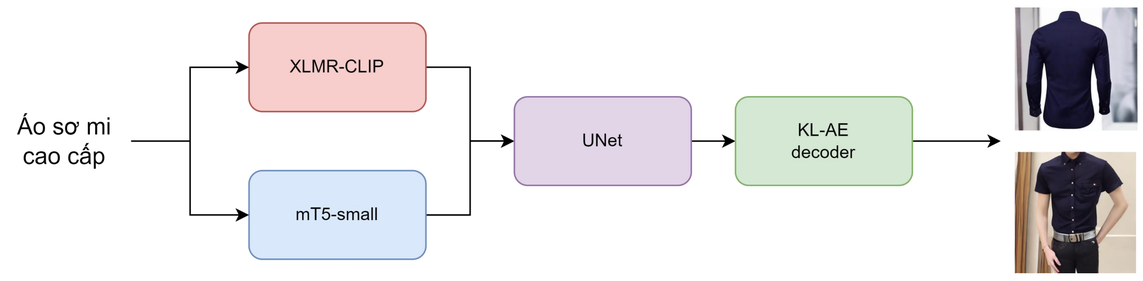

Figure 3: Kandinsky 2.0 model architecture

images_per_prompt	height	width	num_steps	guidance_scale
2	512	512	30	10

Table 2: Model usage parameters

Image processing

We significantly enhanced the quality of generated images through advanced post-processing. Following common practice, we normalized outputs using training data statistics enhanced similarity to originals. Combining multiple images further improved InceptionV3's object recognition.

Custom normalization formula:

$$X = X \times \frac{\sigma_{train}}{\sigma_X} + (\mu_{train} - \mu_X \times \frac{\sigma_{train}}{\sigma_X})$$

where σ_X and μ_X are standard deviation and mean of image X, σ_{train} and μ_{train} are the standard deviation and mean, respectively, of the training set for each color channel,

EXPERIMENTS

We explored two image generators: Stable Diffusion XL (SD-XL) represents English image generator and Kandinsky 2.0 (multilingual). For SD-XL, we tested Vietnamese-to-English translation with EnViT5 and VinAI translate. **Kandinsky 2.0 combined with Qwen-7B-chat-Vietnamese** outperformed other setups by 0.004, and further improved with post-processing. This outcome highlights the potential synergy between Kandinsky 2.0 and Qwen-7B-chat-Vietnamese for achieving superior results in our image generation experiments.

Translator	Prompt builder	Image generator	Image size	Image processing	Score
EnViT5	-	SD-XL	(1024,533)	-	0.40818
VinAI	-	SD-XL	(1024,533)	-	0.40521
-	-	Kandinsky 2.0	(512,512)	-	0.40664
-	Qwen-7b-chat-vietnamese	Kandinsky 2.0	(512,512)	-	0.40151
-	Qwen-7b-chat-vietnamese	Kandinsky 2.0	2 x (512,512)	Normalize and concatenate	**0.37855**

Table 3: Experiments score

SUCESSFUL TIPS

- Utilizing a Prompt builder to succinctly and enhance the input for the Image generator.
- Utilizing an Image generator that has been trained to understand local vocabulary by integrating precise linguistic cues and contextual.
- Enhancing the resulting outputs through the application of image processing

CONCLUSIONS

In this challenge, we proposed the pipeline designed for the proficient generation of advertising banner generation using Vietnamese prompt- condition by an seamless integration combination of prompt builder and multilingual text-to-image based-Diffusion model. Additionally, we incorporated advanced image processing techniques to enhance the overall outcome. The approach demonstrated favorable results compared to single text-to-image models.