

# Towards Culturally-Aware Text-to-Image Generation: Dataset Taxonomy and a Vietnamese Diffusion Model Pipeline

1<sup>st</sup> Van Quang Nghiem  
Viettel AI & Data Services Center  
Hanoi, Vietnam  
quangnv60@viettel.com.vn

**Abstract**—Text-to-image (T2I) generation is a rapidly evolving field in computer vision and natural language processing, enabling the automatic creation of images from textual descriptions. In this paper, we present a comprehensive survey of existing T2I datasets and benchmark tasks, focusing on the diversity of public datasets and the role of image captioning in prompt synthesis. Beyond the survey, we propose a novel taxonomy that categorizes T2I datasets based on language, domain, and annotation characteristics to support research in both English and low-resource languages. Furthermore, we introduce a Vietnamese-specific generation pipeline based on diffusion models, which integrates language-specific preprocessing and cultural context adaptation. Our work provides a solid foundation and practical guidance for future studies aiming to improve T2I generation in multilingual and culturally rich settings.

**Index Terms**—Text-to-Image Generation; Dataset Survey; Taxonomy; Diffusion Model; Vietnamese Language; Multimodal Learning; Prompt Engineering

## I. INTRODUCTION

In recent years, generative models have become a cornerstone of modern artificial intelligence, demonstrating remarkable capabilities across various domains such as text generation, image synthesis, and multimodal understanding. These models are largely empowered by advancements in deep learning architectures, particularly the Transformer [1], which has laid the foundation for state-of-the-art performance in both language and vision-related tasks.

Image generative models—including Variational Autoencoders (VAEs) [2], Generative Adversarial Networks (GANs) [3], and diffusion models like Stable Diffusion [4] and OpenAI’s DALL-E 3 [5]—have revolutionized the way machines generate high-fidelity, realistic visuals. On the language side, models such as OpenAI’s GPT series [6] and Google’s Gemini [7] showcase unprecedented fluency and contextual awareness in natural language generation. Meanwhile, the emergence of multimodal models, often referred to as Vision-Language Models (VLMs)—such as CLIP [8], Flamingo [9], and GPT-4 Vision [6]—enables AI systems to reason across both text and images, marking a significant leap toward generalized intelligence.

These generative models can be categorized into two main types: open-source models (e.g., Stable Diffusion, LLaMA [10], Qwen [11], InternLM [12]) and closed-source commercial models (e.g., DALL-E 3, Midjourney, GPT-4, Gemini).

Each category offers distinct advantages—open-source models provide flexibility for customization and research, while commercial models often deliver cutting-edge performance through proprietary optimization and large-scale training.

Among these, image generation remains one of the most impactful and widely discussed applications globally. Despite this progress, a notable challenge arises when adapting these technologies to non-English and non-Chinese linguistic and cultural contexts—particularly for Vietnamese. Most current generative models are developed and fine-tuned primarily for English or Chinese, limiting their effectiveness when applied to Vietnamese text inputs. Although there have been efforts to create Vietnamese language models such as PhoBERT [13] and VinTERN 3.5 [14] (built on Qwen 2.0.5B Instruct [11] and InternViT-300M-448px [12]), these initiatives primarily focus on text understanding or multimodal capabilities with minimal advancement in Vietnamese-specific image generation.

This report proposes a comprehensive solution to address this gap by building a Vietnamese-centered image generation pipeline. It includes constructing a culturally and linguistically representative dataset, selecting and training suitable generative models, and establishing a benchmark to evaluate performance. Our goal is to develop a system that not only understands Vietnamese prompts accurately but also generates images aligned with Vietnam’s cultural identity and visual aesthetics.

The remainder of this article is organized as follows: Section II reviews related work in the domains of generative models, Vietnamese datasets, and multimodal learning. Section III details our proposed approach, including the taxonomy construction and the Vietnamese-specific generation pipeline. Finally, Section IV presents our conclusions and outlines future work directions.

## II. RELATED WORK

In this section, we analyze related works, including text-to-image (T2I) datasets, benchmarks, image captioning models (used for prompt synthesis), and models for the T2I task.

### A. Datasets

This subsection is divided into two parts: first, an overview of popular public datasets commonly used for T2I tasks;

second, a presentation of some Vietnamese datasets.

1) *General Public Datasets*: Datasets are crucial for training, evaluating, and ensuring fairness in T2I models. Large-scale datasets enhance image generation capabilities. Table I summarizes key datasets in this field.

The COCO dataset [21] contains over 300,000 images, each with five human-generated captions covering diverse objects and scenes. It is widely used for image annotation, generation, and retrieval.

Flickr30k [18] and Flickr8k [18] provide thousands of images with multiple user-generated descriptions, supporting image generation and retrieval tasks.

The Oxford 102 Flower dataset [15] focuses on fine-grained image generation for 102 flower species with detailed descriptions.

The Conceptual Captions dataset [42] offers millions of image-caption pairs gathered from the web, emphasizing conceptual understanding for multimodal learning.

Large-scale datasets like LAION-400M [30] and LAION-5B [30] provide hundreds of millions to billions of image-text pairs, enabling training of large T2I models such as CLIP.

Other datasets such as Visual Genome [24] and SBU Captions [17] offer region-level annotations and contextual captions, enriching model capabilities.

Specialized datasets like DiffusionDB [33] and PaintSkills [39] focus on prompt diversity and artistic image generation, respectively.

Vietnamese datasets will be discussed in the following subsection.

2) *Vietnamese public dataset*: In recent years, the advancement of multimodal artificial intelligence, particularly in tasks such as image captioning and vision-language question answering (VQA), has highlighted the importance of high-quality language-specific datasets. For Vietnamese, a low-resource language, the availability of annotated image-caption datasets and OCR-based VQA corpora remains limited, which poses challenges for training robust models that can generalize across diverse contexts.

Several initiatives have addressed this gap by constructing Vietnamese image captioning datasets. One of the earliest is **UIT-ViIC** [44], derived from MS COCO images focused on sports, containing 3,850 images annotated with five human-written Vietnamese captions per image. This dataset laid the foundation for evaluating Vietnamese captioning models. Similarly, **KTVIC** [43] expands the context to everyday life scenes, consisting of 4,327 images and over 21,000 captions, written manually to reflect natural and contextual Vietnamese.

Another significant contribution is **UIT-OpenViIC** [45], which includes complex real-world scenes captured in Vietnam. This dataset emphasizes the diversity and cultural richness of Vietnamese environments and is particularly valuable for fine-tuning large captioning models on localized data. In addition to manually created datasets, translated datasets such as **Vietnamese COCO 2017** [46] and **Flickr8k Vietnamese** have extended existing English caption corpora into Vietnamese using automatic translation systems, notably VinAI

Translate. Although these help increase the scale, their reliance on machine translation may introduce noise or unnatural phrasing.

Despite these efforts, current Vietnamese image captioning datasets are still relatively few in number and lack diversity in both visual content and captioning styles. Most datasets focus on narrow domains (e.g., sports or daily life) and offer limited generalization across complex, real-world visual-linguistic reasoning.

Beyond captioning, OCR-based VQA has emerged as a vital task, especially for digitizing and understanding documents in Vietnamese. The **Viet-OCR-VQA** dataset by 5CD-AI [14] is a recent addition hosted on Hugging Face, featuring questions and answers grounded in OCR-detected content. Similarly, **ViOCR-VQA** [47] introduces a novel benchmark with vision-reading capabilities that include documents, bills, and ID cards—paving the way for applying language-specific VQA models in administrative and legal domains. Another resource is the **4,995 Vietnamese OCR Images Dataset** [48], which provides annotated Vietnamese text in natural scenes, further supporting the development of Vietnamese OCR engines.

Nevertheless, these OCR-VQA datasets remain fragmented and relatively small-scale. There is a need for standardized benchmarks with unified annotation guidelines and broader domain coverage to enable fair comparisons and reproducibility across models.

While there are several publicly available datasets for image captioning and visual question answering, the **majority** of them are in **English**. Only a **limited number** of datasets provide **Vietnamese-language** captions or QA pairs. Although it is possible to apply **machine translation** or **large language models (LLMs)** to convert English annotations into Vietnamese, the resulting captions often lack **cultural nuances** and **local context** that are essential for building more meaningful Vietnamese AI systems. Therefore, it is **crucial** to develop a **native Vietnamese dataset** that authentically reflects **Vietnamese culture** in both **images** and **captions**.

## B. Image Captioning for Creating T2I Dataset

1) *Image captioning*: In recent years, the task of image captioning has increasingly leveraged powerful **vision-language models (VLMs)** and **vision-language large models (VLLMs)**. These models combine the capability of understanding visual content and generating contextually relevant, coherent text, making them well-suited for generating diverse and natural captions. This process plays a crucial role in building high-quality **text-to-image (T2I)** datasets, where caption diversity and visual fidelity are both essential.

Several representative models have been widely adopted or explored for this task:

- **BLIP/BLIP-2** [49]: Open-source frameworks designed specifically for vision-language tasks, known for strong performance on both captioning and VQA tasks. However, their multilingual capacity—especially for Vietnamese—is still limited without additional fine-tuning.

TABLE I  
DATASETS IN THE T2I TASK.

Dataset	Year	Size	Source
Oxford 102 Flower Dataset [15]	2008	8,189	University of Oxford
UIUC Pascal Sentence [16]	2010	1,000	VOC2008
SBU Captions [17]	2011	1,000,000	Flickr.com
Flickr8K [18]	2013	8,092	Flickr.com
COCO [19]	2014	330,000+	Microsoft
ABSTRACT-50S [20]	2015	500	ASD
Flickr30k [18]	2015	31,783	Flickr.com
COCO Captions [21]	2015	204,721	Microsoft
PASCAL-50S [20]	2015	1,000	VOC2008
vQA [22]	2015	254,721	MS COCO & Abstract Images
Pinterest40M [23]	2016	40,000,000	Pinterest.com
Visual Genome [24]	2017	108,000	Stanford University
vQAv2.0 [25]	2017	204,721	MS COCO
CelebA [26]	2018	202,599	Chinese University of Hong Kong
Nocaps [27]	2019	15,100	Open Images V4 (Flickr.com)
VCR [28]	2019	110	LSMDC & YT
Conceptual Captions [29]	2018	3,369,218	Google Research
LAION-400M [30]	2021	400,000,000	LAION (Large-scale AI Open Network)
CC-500 [31]	2022	500	Synthetic prompts
Conceptual 12M [32]	2021	12,423,374	World Wide Web
LAION-5B [30]	2022	5,000,000,000	LAION
DiffusionDB [33]	2022	2,000,000	Open-source contributions
Winoground [34]	2022	800	Getty Images API
DrawBench [35]	2022	200	DALL-E & Reddit
ABC-6K [36]	2022	6,400	MS COCO
I2P [37]	2023	4,703	User generated prompts
T2I-CompBench [38]	2023	6,000	Generated prompts by GPT
PaintSkills [39]	2023	65,535	Synthetic prompts
RichHF-18K [40]	2024	18,000	Pick-a-Pic
MARIO-10M [41]	2024	10,061,720	LAION, TMDB, Open Library

TABLE II  
OVERVIEW OF VIETNAMESE IMAGE CAPTIONING AND OCR-VQA DATASETS

Dataset Name	#Images	#Captions / QA pairs	Description
KTVIC [43]	4,327	21,635 (5/img)	Manually written Vietnamese captions; covers various daily-life scenes.
UIT-ViC [44]	3,850	19,250 (5/img)	Sports-related images from MS COCO; among the first captioning datasets for Vietnamese.
UIT-OpenViC [45]	~3,000	~15,000	Real-world scenes in Vietnam with culturally contextualized captions.
Vietnamese COCO 2017 [46]	118,000 (train+val)	~590,000	Machine-translated from English COCO using VinAI Translate.
Flickr8k Vietnamese	8,000	40,000	Automatically translated from Flickr8k; useful for small-scale training.
Viet-OCR-VQA (SCD-AI) [14]	~3,000	~12,000	OCR-based VQA dataset with real-world scanned images.
ViOCRvQA [47]	28,282	123,781	OCR-VQA dataset on various documents. Introduced in: <i>ViOCRvQA: Novel Benchmark Dataset and Vision Reader for Vietnamese VQA</i> .
4,995 Vietnamese OCR Images Dataset [48]	4,995	N/A	258 natural scene images, 2,553 internet images, and 2,184 document scans. Annotated with line/column bounding boxes and text. Accuracy >97%.

- **GPT-4V** [6]: A closed-source, commercial VLLM developed by OpenAI that shows impressive capability in generating captions that are **rich in detail**, often identifying both **locations** and **cultural elements** (e.g., “Hoan Kiem Lake”, “ao dai”). Nonetheless, its use is restricted due to **API cost** and lack of open access.
  - **Gemini 2.5 Flash/Pro** [7]: Another commercial model with multilingual support and good captioning abilities, though it tends to provide **less specific context** compared to GPT-4V. Moreover, it suffers from the same closed-access limitations.
  - **Qwen2.5-VL-7B-Instruct** [11]: An open-source vision-language model that supports Vietnamese. It can interpret image content and generate coherent captions. However, due to being **not trained on culturally specific Vietnamese datasets**, it often lacks understanding of Vietnamese locations or attire.
  - **Vi-VLM/Vistral-V-7B** [50]: A recent Vietnamese-finetuned open-source model that demonstrates strong potential. It recognizes common Vietnamese landmarks (e.g., Ho Guom), attire (ao dai), and context, due to being **specifically trained on Vietnamese datasets**. This direction appears to be the most **feasible and scalable**, as such models can be further improved using **LoRA**, **QLoRA**, **PEFT**, or full fine-tuning—without requiring enormous hardware.
- Comparison Example.** Below is a qualitative comparison of image captioning outputs from different models (see Figure 1) on a set of Vietnamese cultural images:
- **GPT-4V** provides **highly descriptive and culturally aware** captions. It recognizes Ho Guom, ao dai, Ha Noi, and the setting, producing natural, human-like captions. Its drawback is high cost and lack of openness.
  - **Gemini 2.5** models produce grammatically correct cap-

tions but are **less contextualized**. The captions tend to generalize visual features and may not fully capture the Vietnamese context.

- **Qwen2.5-VL** shows good image understanding and supports Vietnamese prompts. However, due to **limited exposure to Vietnamese cultural data**, it often fails to recognize landmarks or local attire correctly.
- **Vistral-V-7B**, being **finetuned on Vietnamese datasets**, generates captions with better cultural alignment. It can name Kinh Thanh Hue, recognize Cho Noi, and describe scenes in a Vietnamese way. Its open-source nature also makes it suitable for further development.

2) *OCR Image Captioning*: Optical Character Recognition (OCR) plays an important role in image captioning, particularly when dealing with images that contain textual content. Understanding the embedded text in images allows for more accurate and meaningful captions, which is crucial when constructing high-quality text-to-image (T2I) datasets. This is especially relevant for applications such as generating advertisement posters or propaganda-style images, where text is a core part of the visual design.

Traditional OCR pipelines for Vietnamese have seen significant development. Among the most effective approaches are models fine-tuned from PaddleOCR [51] for Vietnamese, or systems that combine PaddleOCR with VietOCR [52] to extract both textual content and its corresponding bounding boxes. However, these conventional systems often assume consistent layouts or fixed formats, which limits their scalability for large-scale, diverse datasets where image structures vary significantly.

Recently, vision-language models (VLMs) have emerged as promising alternatives. For example, Vintern-1B-v2 [14] is a compact VLM with only one billion parameters, yet demonstrates impressive performance on OCR and VQA tasks for Vietnamese inputs. While this model excels at extracting text and preserving its visual structure, it lacks the capability to return precise bounding box coordinates for the detected text.

On the other hand, larger models such as GPT-4V, Gemini 2.5, or Qwen-VL-72B are capable of handling both OCR and spatial localization tasks. These models can not only recognize and understand text in images but also provide coordinates for each text span. The trade-off, however, lies in their cost — either requiring API calls or significant hardware resources for on-premise deployment.

Nevertheless, the high OCR capability of modern VLMs makes them viable tools for developing OCR-aware image captioning systems, which are essential in constructing robust T2I datasets that include embedded textual elements.

### C. Text to Image generated model

Modern text-to-image (T2I) generation models predominantly follow two architectural paradigms: diffusion-based models and autoregressive models. Each comes with its strengths, limitations, and implications when applied to low-resource languages like Vietnamese.

1) *Diffusion-based Models*: Diffusion models have become the mainstream approach for high-fidelity image synthesis. Prominent examples include **Stable Diffusion** [4], **SDXL** [53], and the upcoming **Stable Diffusion 3.5** [54], all of which are trained on large-scale English datasets. These models typically rely on a dual-encoder setup, where a text encoder (such as CLIP’s text model) maps the input prompt into a latent space, which then guides the iterative denoising process to generate an image.

While these models can generate visually appealing content, they face challenges when applied to Vietnamese inputs. Several strategies have been explored to bridge this gap, such as replacing the English text encoder with a Vietnamese model like **PhoBERT** [13], or adapting the entire CLIP [8] backbone to a Vietnamese-compatible version. Although such methods improve the model’s ability to interpret Vietnamese prompts, the *latent space* of the pretrained diffusion model remains inherently biased towards English-language semantics. As a result, the generated images often fail to capture the unique cultural, contextual, and linguistic elements of Vietnamese.

A critical limitation of diffusion models—especially those based on the Stable Diffusion architecture—is their poor handling of **text rendering** within images. For example, when tasked with generating posters or flyers that require accurate Vietnamese text placement, these models often produce distorted or unreadable characters. This shortcoming hinders their direct applicability in design-oriented tasks such as advertising or educational content creation.

To address this, techniques such as **ControlNet** [55] have been introduced to guide the generation process using additional conditions (e.g., segmentation maps, pose skeletons, or text layouts). However, integrating such constraints still requires a multi-stage pipeline, often involving OCR extraction, layout specification, and fine-grained prompt engineering. Therefore, building a robust Vietnamese T2I system likely necessitates decomposing the problem into sub-tasks with modular solutions.

2) *Autoregressive Models*: On the other side of the spectrum, autoregressive models like **DALL·E** [5], and newer research efforts such as **Flamingo** [9] and **FluX** [56] (from OpenAI and other labs) adopt a token-by-token generation approach. These models often treat both text and images as sequences of discrete tokens and generate visual outputs in a left-to-right (or block-wise) manner.

Autoregressive models generally offer finer control over structure and layout and have shown greater promise in tasks requiring the faithful reproduction of text, logos, and graphical elements. However, they typically lag behind diffusion models in terms of photorealism and visual fidelity. Additionally, training such models requires massive computational resources and access to tokenized multi-modal datasets, which are often unavailable or limited for low-resource languages like Vietnamese.

Nonetheless, given their strength in layout-sensitive tasks, autoregressive models may provide a valuable foundation for future systems where text rendering and spatial alignment are



Input Prompt: "Tạo mô tả ngắn cho ảnh sau:"





				
GPT4v:	Bức ảnh ghi lại khoảnh khắc ba người phụ nữ mặc áo dài truyền thống, tươi cười bên nhau bên bờ Hồ Gươm (Hồ Hoàn Kiếm), với Tháp Rùa mờ ảo phía xa. Một người trong số họ cầm bó hoa tím, tạo nên không khí dịu dàng và đậm chất Việt Nam.	Bức ảnh ghi lại khung cảnh quen thuộc của một quán trà đá vỉa hè tại Việt Nam, nơi nhiều người ngồi trên ghế nhựa nhỏ, trò chuyện và thư giãn. Phía sau là một cửa hàng tạp hóa với các loại nước giải khát và đồ ăn nhẹ, tạo nên không gian gần gũi, đời thường đậm chất phố phường Việt Nam.	Bức ảnh ghi lại cảnh đoàn người mặc áo dài truyền thống tham gia diễu hành tại Đại Nội Huế. Các thiếu nữ duyên dáng trong tà áo dài, tay cầm hoa sen và quạt, bước đi trong không gian cổ kính của cố đô, tạo nên một khung cảnh vừa trang nghiêm vừa đậm đà bản sắc văn hoá dân tộc.	Bức ảnh ghi lại khung cảnh nhộn nhịp tại một chợ nổi miền Tây Nam Bộ, nơi người dân buôn bán nông sản trên những chiếc ghe thuyền chèo đầy trái cây tươi ngon. Màu sắc rực rỡ, không khí sôi động cùng hình ảnh quen thuộc của chiếc nón lá tạo nên nét đặc trưng văn hoá sông nước Việt Nam.
Gemini2.5-flash	Ba người phụ nữ mặc áo dài truyền thống đang đứng cạnh nhau, mỉm cười và nhìn về phía bên phải. Họ đang cầm những bó hoa màu tím. Phía sau họ là một hồ nước yên tĩnh và một kiến trúc cổ kính mờ ảo trong sương sớm hoặc hoàng hôn.	Một nhóm người đang ngồi trên những chiếc ghế nhựa nhỏ, trò chuyện và uống nước trước một quán trà đá vỉa hè.	Một đoàn người, bao gồm cả những người mặc áo dài và trang phục truyền thống khác, đang diễu hành trước một công trình kiến trúc cổ kính với mái ngói vàng.	Chợ nổi tập nập với nhiều thuyền bè chèo đầy hàng hóa, chủ yếu là trái cây và rau củ, cùng hoạt động mua bán sôi nổi của người dân trên sông.
Gemini2.5-Pro	Ba người phụ nữ duyên dáng trong tà áo dài truyền thống, tay ôm bó hoa tím dịu dàng, cùng nhau tươi cười bên hồ Gươm thơ mộng, phía xa thấp thoáng bóng Tháp Rùa cổ kính.	Một quán trà đá vỉa hè bình dị, quen thuộc ở Hà Nội, nơi mọi người đủ mọi lứa tuổi ngồi trên những chiếc ghế nhựa nhỏ, quây quần trò chuyện và thưởng thức đồ uống. Phía sau là cửa hàng tạp hóa với nhiều mặt hàng được bày bán.	Các thiếu nữ duyên dáng trong tà áo dài truyền thống, đầu đội khăn, đang tham gia một buổi diễu hành hoặc lễ hội. Một số cô cầm hoa sen, làm tăng thêm vẻ thanh lịch. Phía sau là một công trình kiến trúc cổ kính, gợi liên tưởng đến cung đình Huế. Khung cảnh trang trọng và mang đậm nét văn hóa truyền thống.	Một phiên chợ nổi tập nập trên sông, đặc trưng của vùng Đồng bằng sông Cửu Long. Những chiếc ghe, thuyền chèo đầy trái cây và nông sản tươi ngon như dưa, xoài, và rau củ, cùng với những người bán hàng đội nón lá đang trao đổi, mua bán sôi nổi. Hai bên bờ là những hàng dừa xanh mát, tạo nên khung cảnh sông nước trữ tình và đậm đà bản sắc miền Tây.
Qwen2.5VL-32B	Hình ảnh cho thấy ba người phụ nữ đang đứng bên bờ hồ, trong tay mỗi người cầm một bó hoa màu tím. Họ đang nhìn về phía xa, bên cạnh họ có một công trình kiến trúc cổ kính được phản chiếu trên mặt nước. Khung cảnh yên bình và đẹp mắt, tạo nên một không gian tĩnh lặng và thư giãn.	Hình ảnh cho thấy một nhóm người đang ngồi trên các chiếc ghế nhựa màu đỏ và vàng, trước một cửa hàng nhỏ. Họ có vẻ như đang thư giãn, trò chuyện và thưởng thức các sản phẩm bày bán bên ngoài cửa hàng, như nước ngọt và các mặt hàng khác. Cửa hàng nhỏ này nằm trong một khu vực đô thị, với một số xe máy đỗ cạnh đó.	Hình ảnh cho thấy một nhóm người mặc trang phục truyền thống Việt Nam đang diễu hành trước một công trình kiến trúc cổ kính và hoành tráng.	Hình ảnh cho thấy một chợ nổi sôi động với hàng trăm con thuyền chèo đầy trái cây và hàng hóa, người dân mua bán trên sông.
Vistral-V-7B	Trong bức ảnh, có ba người phụ nữ mặc áo dài, cầm hoa và tươi cười. Phía xa là Hồ Gươm, biểu tượng của Hà Nội.	Đây là một quán trà đá vỉa hè có phong cách Hà Nội với nhiều người ngồi trên ghế nhựa, nói chuyện.	Hình ảnh nhiều người mặc trang phục truyền thống thời xưa đang đi bộ trước Kinh thành Huế.	Đây là một phiên chợ nổi ở Nam Bộ, Việt Nam, với nhiều thuyền chèo trái cây trên sông nước.

Fig. 1. Fine-Tuning Process Overview

critical — especially if adapted to incorporate Vietnamese tokenizer support and multilingual training objectives.

3) *Summary*.: Overall, while diffusion models dominate in terms of popularity and image quality, their limitations in multilingual text rendering — especially Vietnamese — suggest the need for hybrid systems. Combining the strengths of both paradigms, possibly through modular pipelines or cross-modal fusion (e.g., CLIP-guided layout control combined with OCR-informed generation), could offer a viable path forward for generating Vietnamese-specific visual content with embedded text.

#### D. Model Evaluation

Evaluating the performance of vision-language models is essential for determining their quality, reliability, and suitability for downstream tasks. This section describes evaluation

metrics and methodologies for three main components: Image Captioning, OCR, and Text-to-Image generation using diffusion models.

##### 1) Image Captioning Evaluation:

a) *Automatic Metrics*.: Image captioning models are typically evaluated using a set of standard metrics that compare the generated caption with human-written references. The most common include:

- **BLEU (Bilingual Evaluation Understudy)**: BLEU-n compares the n-gram precision between a candidate and reference captions. It is computed as:

$$\text{BLEU-n} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad (1)$$

where  $p_n$  is the modified n-gram precision,  $w_n$  is a weight

(usually  $1/N$ ), and BP is the brevity penalty:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r \end{cases} \quad (2)$$

with  $c$  and  $r$  being the lengths of the candidate and reference captions.

- **METEOR**: This metric aligns unigrams based on exact, stemmed, or synonym matches. The score is computed as:

$$METEOR = F_{mean} \cdot (1 - Penalty) \quad (3)$$

where  $F_{mean}$  is the harmonic mean of unigram precision and recall, and Penalty depends on the number of chunks (non-contiguous matches).

- **CIDEr**: Weights n-gram similarity by TF-IDF across references:

$$CIDEr(c, S) = \frac{1}{|S|} \sum_{s \in S} \text{sim}_{tf-idf}(c, s) \quad (4)$$

where  $c$  is the candidate and  $S$  the set of references.

- **SPICE**: Converts captions into scene graphs and compares tuples (object, attribute, relation).
- **CLIPScore**: Measures semantic similarity using CLIP embeddings:

$$CLIPScore(I, T) = \cos(\phi_I(I), \phi_T(T)) \quad (5)$$

where  $\phi_I$  and  $\phi_T$  are the CLIP image and text encoders.

#### Example (BLEU-1):

Reference: ``A dog is playing with a ball.''   
 Candidate: ``A dog plays with a ball.''

Unigram matches: “A”, “dog”, “with”, “a”, “ball”  $\Rightarrow$  5/6 = 0.83 BLEU-1.

b) *Human Evaluation.*: Captions are judged on:

- *Relevance*: Accuracy in describing image content.
- *Fluency*: Grammar and readability.
- *Richness*: Descriptive and contextual depth.

Annotators rate captions on a 5-point Likert scale. Agreement is measured with Cohen’s kappa.

2) *OCR Model Evaluation*:

a) *Benchmark Datasets.*: Typical benchmarks include:

- **ICDAR Robust Reading Challenges (2013–2019)**
- **Vietnamese datasets: UIT-ViODC, Vietnamese OCR (VietOCR-Benchmark)**

b) *Evaluation Metrics.*:

- **Character Error Rate (CER)**:

$$CER = \frac{S + D + I}{N} \quad (6)$$

where  $S$ ,  $D$ , and  $I$  are the counts of substitutions, deletions, and insertions needed to match the predicted string to the reference, and  $N$  is the number of characters in the reference.

- **Word Error Rate (WER)**: Similar to CER, but counts word-level operations.

- **Bounding Box Metrics**: If OCR includes detection:

– **IoU (Intersection over Union)**:

$$IoU = \frac{\text{Area}(B_{pred} \cap B_{gt})}{\text{Area}(B_{pred} \cup B_{gt})} \quad (7)$$

– **Precision, Recall, F1-score**: Computed over matched boxes using IoU  $\geq$  threshold (e.g., 0.5).

#### Example (CER):

Reference: ``Hanoi''   
 Prediction: ``Hanio''

One substitution (i  $\rightarrow$  o)  $\Rightarrow$  CER = 1/6 = 0.167

c) *Additional Aspects.*: When evaluating OCR for documents or multimodal tasks, structure preservation (e.g., reading order, table layout) may also be measured using tree or graph alignment scores.

3) *Diffusion Model Evaluation*:

a) *Automatic Metrics.*:

- **CLIPScore**: Measures prompt-image semantic alignment.
- **Frechet Inception Distance (FID)**:

$$FID = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (8)$$

where  $(\mu_r, \Sigma_r)$  and  $(\mu_g, \Sigma_g)$  are the mean and covariance of real and generated image features.

- **Inception Score (IS)**: Combines confidence and diversity:

$$IS = \exp(\mathbb{E}_x[KL(p(y|x)||p(y))]) \quad (9)$$

where  $p(y|x)$  is the class probability of image  $x$  and  $p(y)$  is the marginal distribution.

- **Structural Similarity Index (SSIM)**:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (10)$$

where  $\mu_x, \mu_y$  are mean intensities,  $\sigma_x^2, \sigma_y^2$  variances, and  $\sigma_{xy}$  the covariance.

b) *Human Evaluation.*: Crucial for assessing perceptual quality and alignment:

- 1) Sample 50–100 prompts.
- 2) Generate corresponding images.
- 3) Ask raters to evaluate on:

- **Prompt Alignment**: Image content matches textual input.
- **Visual Realism**: Image appears photorealistic or stylistically correct.
- **Text Correctness**: For prompts involving embedded text.

- 4) Use Likert scale or pairwise ranking to compute Mean Opinion Score (MOS).

### III. METHODS

#### A. Vietnam Dataset Taxonomy for Diffusion Model Training

This section outlines the design of a multi-dimensional taxonomy framework to organize image data related to Vietnam (Alg. 1). This taxonomy will support the efficient training of diffusion-based generative models, with emphasis on historical, cultural, geographic, and economic representations of Vietnam.

1) *Temporal Taxonomy (\$T\$): Historical Periods of Vietnam*: Vietnamese history is long and complex. We divide the temporal axis into three major periods, each with different objectives for dataset construction:

- **Pre-1858: Dynastic and Ancient Civilizations**

- Focus: major historical events, ancient culture, and pre-modern society.
- Themes: Bronze Drums (Dong Son culture), wars against Northern invaders (e.g., Bach Dang River battles), Ly–Tran dynasty temples, Confucianism, feudal architecture.
- Data: Illustrative paintings, dioramas, heritage reconstructions, archaeological findings.

““

- **1858–1975: Colonial and Wartime Eras**

- Focus: French and American colonial periods, wars of independence.
- Themes: anti-colonial resistance, Indochinese architecture, military uniforms, revolution posters, refugee migrations.
- Data: Archival photographs, war documentaries, historical reconstructions.

- **Post-1975–Present: Socialist and Contemporary Vietnam**

- Focus: Modernization, economic reform (Doi Moi), globalization.
- Themes: contemporary architecture, public housing, traditional festivals in modern settings, infrastructure, tourism.
- Data: Photographs from 1975 to today, street life, industry, landscape, contemporary fashion.

““

2) *Spatial Taxonomy (\$L\$): Location-Based Structuring*: Vietnam’s diversity in geography and human development can be captured using three main spatial strategies:

a) *Administrative Division: 63 Provinces and Cities*:

Vietnam is divided into 63 first-level administrative units:

- 58 provinces (tinh)
- 5 centrally-controlled municipalities (thanh pho truc thuoc trung uong)

This allows precise modeling of localized phenomena (e.g., Hue imperial city, Sa Pa market).

b) *Climatic and Geographic Zones: 7 Macro Regions*:

These zones represent shared cultural and environmental conditions:

- 1) Northern Midlands and Mountainous Area (Tay Bac – Dong Bac)
- 2) Red River Delta (Dong Bang Song Hong)
- 3) North Central Coast (Bac Trung Bo)
- 4) South Central Coast (Nam Trung Bo)
- 5) Central Highlands (Tay Nguyen)
- 6) Southeast Region (Dong Nam Bo)
- 7) Mekong River Delta (Dong Bang Song Cuu Long)

Suitable for stylistic clustering (e.g., highland clothing, delta rice farming).

c) *Economic Zones*: Vietnam has key economic regions which focus on development:

- **Northern Key Economic Zone (NKEZ)** – Ha Noi, Hai Phong, Bac Ninh...
- **Central Key Economic Zone (CKEZ)** – Da Nang, Quang Nam...
- **Southern Key Economic Zone (SKEZ)** – Ho Chi Minh City, Binh Duong...
- Special administrative/economic zones – Van Don, Phu Quoc, Thu Duc City...

Use cases: modeling industrialization, transport, new cities.

d) *Hybrid Tagging Structure*:

```
tags = {
    time = "1990{2000",
    region = "Nam Trung Bo",
    province = "Phu Yen",
    econ\_zone = "CKEZ",
    category = "Architecture",
    ethnic = "Kinh"
}
```

3) *Category Taxonomy (\$C\$): Cultural and Symbolic Categories*: We define visual categories reflecting Vietnam’s rich cultural ecosystem:

- **Architecture**: pagodas, temples, French colonial buildings, post-war socialist housing, skyscrapers
- **Costumes and Textiles**: Ao Dai, ethnic minorities’ dress, war uniforms, modern fashion
- **Festivals and Rituals**: Tet, Mid-Autumn Festival, rural weddings, religious ceremonies
- **Daily Life and Labor**: farming, fishing, tea making, street vendors, factories
- **Transportation**: bicycles, cyclos, war-era tanks, trains, motorbikes, highways
- **Symbols and Flags**: national flags, traditional patterns, propaganda posters

Each image is tagged by one or more visual-cultural categories.

4) *Ethnic Taxonomy (\$E\$): Ethnolinguistic Groups*: Vietnam has 54 officially recognized ethnic groups. While most images may belong to the Kinh majority, it is important to label data from ethnic minorities:

- Major minorities: Tay, Thai, Muong, Hmong, Khmer, Hoa, Nung, Cham, Ede, Bahnar, etc.
- Use in fashion, festival, and architectural generation.
- Helps with fair and diverse cultural representation.

5) *Summary: Tagging Structure*: Each image (or group of sequential images) will be tagged as follows:

```
tags = {
  time = "1980 - 1990",
  location = "Dong Nai",
  region = "Dong Nam Bo",
  econ\_zone = "SKEZ",
  category = \["Daily Life",
  "Architecture"],
  ethnic = "Kinh"
}
```

This multi-dimensional tagging structure allows both dataset filtering and conditional image generation across space, time, culture, and people.

---

**Algorithm 1** Vietnam Multi-Dimensional Data Collection

---

**Input:**  $T$ : time periods,  $L$ : locations,  $C$ : categories,  $A$ : attributes

**Output:**  $D$ : collected dataset with tagged images

```
1:  $D \leftarrow \emptyset$ 
2: for all  $t \in T$  do
3:   for all  $l \in L$  do
4:     for all  $c \in C$  do
5:       for all  $a \in A$  do
6:          $q \leftarrow \text{GenQuery}(t, l, c, a)$   $\triangleright$  compose
search query
7:          $imgs \leftarrow \text{Crawl}(q)$   $\triangleright$  collect raw image
set
8:         for all  $img \in imgs$  do
9:           if  $\text{IsValid}(img)$  then
10:             $meta \leftarrow \{time : t, location : l, category : c, attribute : a\}$ 
11:             $D \leftarrow D \cup \{(img, meta)\}$ 
12:          end if
13:        end for
14:      end for
15:    end for
16:  end for
17: end for
18: return  $D$ 
```

---

### B. Vietnam People's Army Imagery Subset

The Vietnam People's Army (VPA) plays a crucial role in the country's modern and historical narrative, making it a high-priority subset in the Vietnamese image taxonomy for diffusion model training. This subset should reflect the evolution of military culture, uniform design, tactical organization, and equipment across key time periods and operational units. Below we outline the multi-dimensional taxonomy structure for the VPA image dataset.

1) *Historical Phases*: We categorize VPA data into three major historical segments, each reflecting distinct visual and symbolic characteristics:

- **Anti-French Resistance (1945–1954)**: Focus on the Viet Minh troops, rudimentary gear, guerilla warfare uniforms, and important events such as the Dien Bien Phu campaign.
- **Anti-American War (1955–1975)**: Includes North Vietnamese Army (NVA) operations, Ho Chi Minh Trail logistics, iconic gear like the pith helmet (mu cung), and major battles such as Khe Sanh and Tet Offensive.
- **Post-1975 to Present**: Emphasizes modernization: camouflage uniforms, mechanized units, parades, peacekeeping operations, and military exercises with new equipment such as tanks, aircraft, and missile systems.

2) *Unit-Based Taxonomy*: Each military unit within the VPA has distinct visual features that aid in diverse image collection:

- **Infantry Divisions**: Common soldiers, basic weaponry, camouflage styles based on terrain.
- **Artillery and Air Defense**: Includes mobile rocket systems, radar stations, and associated personnel.
- **Air Force**: Pilots, aircraft (e.g., Su-30MK2), and flight uniforms.
- **Navy and Marine Units**: Naval uniforms, ships, submarines, and amphibious operations.
- **Border Guards and Special Forces**: Tactical gear, night-vision equipment, and training scenarios.
- **Cadets and Training Schools**: Students in military academies with characteristic green training attire.

3) *Clothing and Equipment Dimensions*: The tagging should cover distinct military items and symbols, such as:

- **Uniforms**: Seasonal styles, ceremonial vs. field gear, rank insignias.
- **Helmets and headgear**: Standard VPA green helmet, berets for special units.
- **Vehicles**: Soviet-style tanks, trucks, modern APCs, naval vessels, aircraft.
- **Weapons**: AK-47 variants, RPGs, shoulder-mounted missiles, machine guns.

4) *Media and Cultural Sources*: High-quality references can be drawn from both state media and military-themed entertainment, such as:

- **Military Newspapers**: Each major unit or regional command often maintains its own publication or website (e.g., baoquankhu4.com.vn, quankhu7.vn).
- **National Broadcast Archives**: Programs from QPVN (National Defense Channel), especially military parades, news features, and training drills.
- **Reality Shows**:
  - **Sao nhap ngu (Military Rookie Show)**: Offers real-life footage of new recruits undergoing training in various military environments.
  - **Chung toi la chien si (We Are Soldiers)**: A long-running VTV program showcasing both cultural life and combat simulation drills of enlisted personnel.
- **YouTube & Fanpages**: Channels that upload clips from military shows or veteran groups sharing wartime photos.





This method allows for greater control and flexibility when generating images that must include specific textual content. However, it also adds complexity to the pipeline and further increases the need for annotated data that includes visual text cues.

Overall, this approach offers a high-potential direction for creating culturally and linguistically aligned Vietnamese text-to-image systems, though it requires significant investment in data collection and training infrastructure.

2) *Approach 2: Add conditions to SD via ControlNet.*: This pipeline generates culturally grounded Vietnamese poster-style images using Stable Diffusion (SD) with ControlNet and reference images. Instead of retraining SD, the approach adds modular control using ControlNet and enhances generation fidelity with real Vietnamese references. See Fig. 3 and Alg. 2.

---

**Algorithm 2** Vietnamese Poster Generation Pipeline

---

**Input:**  $x$ : user prompt

**Output:** Stylized image  $\hat{y}$

**Data:** VectorDB of Vietnamese references

```

1:  $p \leftarrow \text{SystemPrompt} + \text{UserPrompt}$ 
2:  $s \leftarrow \text{LLM}(p)$  ▷ parsed scene
3:  $R \leftarrow \text{VectorDB.search}(s)$  ▷ retrieve references
4: for  $r \in \{\text{pose, cloth, bg}\}$  do
5:    $C_r \leftarrow \text{ControlNet}_r(R_r)$ 
6: end for
7: if add_text then
8:    $T \leftarrow \text{ControlNet}_{\text{text}}(\text{glyph})$ 
9: end if
10:  $z \leftarrow \text{SD}(C_{\text{pose}} \oplus C_{\text{cloth}} \oplus C_{\text{bg}} \oplus T \oplus p)$ 
11: if RL enabled then
12:    $r \leftarrow \text{RewardModel}(z)$ 
13:   Update model using  $r$ 
14: end if
15: return  $z$ 

```

---

Explanation of Pseudo-Code:

- User Prompt is processed with a System Prompt to guide LLMs.
- LLMs parse the prompt into structured phrases (e.g., "couple eating Pho").
- These phrases are used to search a VectorDB, retrieving relevant Vietnamese reference images.
- The images are fed into multiple ControlNets (e.g., for pose, background, clothing).
- SD combines all signals and generates an image.
- A glyph-image ControlNet can optionally add text.
- A Reward model evaluates the image (optional RL feedback).

a) *Prompt Parsing via Lightweight LLM*: To extract meaningful Vietnamese visual concepts from user prompts, we employ a lightweight large language model (LLM) with approximately 7 billion parameters. The model is guided by a fixed system prompt:

**Prompt Template:**

You are a visual concept extractor for Vietnamese cultural content in image captions. Given a sentence, extract two types of information:

1. **\*\*Vietnamese Visual Keywords\*\***: iconic Vietnamese food, clothing, architecture, festivals, or items.
2. **\*\*Contextual Vietnamese Phrases\*\***: short phrases including actions or descriptions involving Vietnamese visual elements.

Respond in JSON with two fields: 'keywords' and 'phrases'.

User: [User Prompt]

The LLM parses the user’s natural language description into a structured JSON output containing two fields: keywords and phrases. For instance, given a prompt such as "A girl wearing ao dai, standing next to Cho Ben Thanh", the model may extract "ao dai" and "Cho Ben Thanh" as keywords, and "a girl wearing ao dai" as a contextual phrase. These parsed outputs are then used for querying a visual reference database in the next stage of the pipeline.

b) *Image Retrieval via Phrase-based Embedding.*: After extracting semantically rich Vietnamese visual phrases from textual prompts using a large language model (LLM), we embed these phrases into a shared image-text representation space using a retrained CLIP model tailored for Vietnamese cultural concepts. Unlike generic keyword-based search, our approach preserves contextual information (e.g., "a couple eating Pho" instead of simply "Pho"), which enhances semantic alignment between text and images. These embedded phrases are used to retrieve the most relevant images from a large-scale image database via similarity search. The retrieved images—containing culturally significant visual elements—serve as visual conditions to guide a downstream generative model. Specifically, they are used as reference or conditioning inputs for a Stable Diffusion (SD) model, enabling it to generate images that are not only text-aligned but also visually grounded in authentic Vietnamese aesthetics and iconography.

c) *Conditions for Diffusion.*: To enhance controllability and preserve visual faithfulness in image generation, we employ ControlNet modules to inject structured visual conditions into the diffusion process. The choice of ControlNet type depends on the nature of the extracted Vietnamese visual phrases. For instance, when the phrase includes human interactions such as "a couple eating Pho," we utilize a pose-based ControlNet (e.g., OpenPose) to guide the human body layout and interaction. This ensures that the generated image captures realistic and culturally appropriate human gestures.

In cases where clothing elements such as "ao dai" are emphasized, we optionally combine pose ControlNet with a segmentation-based ControlNet (e.g., semantic segmentation or soft edge) to preserve the flowing structure and silhouette

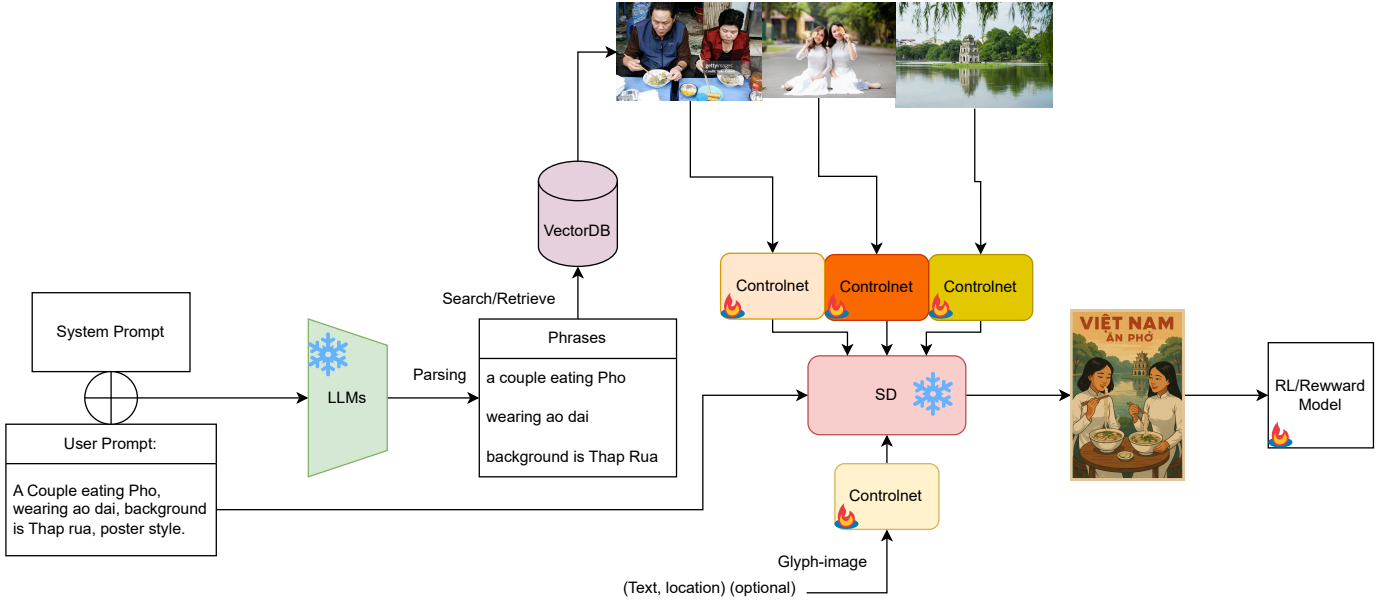


Fig. 3. Approach 2: Overview

of traditional Vietnamese garments. This combination helps maintain the characteristic appearance of the ao dai in the generated imagery.

For background-specific elements like “Thap Rua” (Turtle Tower), we apply depth-based or canny-edge ControlNet modules. These models are effective in constraining architectural structures and spatial layouts, ensuring that the generated background remains faithful to the iconic landmark.

In addition to these visual conditions, we incorporate textual guidance—such as the original caption or refined visual phrases—using text-to-image capabilities or image inpainting when parts of the scene need to be conditionally completed or corrected.

Depending on the domain-specific requirements, these ControlNet models can either be used as-is (zero-shot) or fine-tuned on a curated dataset containing Vietnamese-specific visual patterns. For improved fidelity, we explore lightweight fine-tuning strategies (e.g., LoRA or DreamBooth) on each ControlNet branch to better adapt to cultural nuances and ensure consistency with the conditioning phrases.

*d) Diffusion Model.:* Our final image synthesis step leverages a Stable Diffusion model conditioned on both visual control signals and user-defined prompts. The model receives multi-modal conditions from ControlNet—such as pose maps, edge maps, depth maps, or semantic segmentations—that encode structural and contextual information extracted from the original input. These control conditions act as spatial priors that constrain the generation process, allowing the model to produce images that are compositionally accurate and visually coherent.

Simultaneously, a natural language prompt provided by the user (e.g., “a couple eating Pho in front of Thap Rua”) guides the model’s semantic and stylistic output through cross-

attention mechanisms within the diffusion architecture. This fusion of textual and structured visual conditions enables the model to generate high-fidelity images that are faithful to both the intended narrative and culturally specific visual details.

We build upon the Stable Diffusion 3 architecture, optionally extending it with ControlNet modules and fine-tuned LoRA weights to better align with Vietnamese aesthetics. This approach ensures that generated images not only follow the prompt accurately, but also integrate culturally meaningful elements like “Pho”, “ao dai”, and “Thap Rua” in a visually grounded and artistically coherent manner.

*e) Reward Model.:* To further enhance the quality and cultural relevance of the generated images, we integrate a reward model based on Reinforcement Learning with Human Feedback (RLHF). The core idea is to fine-tune the diffusion process by introducing a classification model that evaluates whether the generated image contains distinctive Vietnamese cultural features. This classifier is trained with two classes: *Vietnamese characteristic present* and *Vietnamese characteristic absent*, based on a dataset of culturally rich images.

During the training phase, the reward model assesses the output of the Stable Diffusion model conditioned by ControlNet. If the generated image is classified as having strong Vietnamese characteristics (e.g., “Pho”, “Ao Dai”, or “Thap Rua”), it receives a positive reward, encouraging the model to focus more on these cultural elements in future generations. Conversely, images that fail to capture these features are penalized. This feedback loop is utilized to refine the ControlNet conditioning process, leading to better optimization of the spatial and semantic conditions for culturally grounded image generation.

By using RLHF, we enable continuous improvements in the model’s ability to align its outputs with both user prompts and

culturally specific visual attributes, ensuring that the generated images maintain high fidelity to Vietnamese aesthetics while enhancing the overall quality and relevance of the generated content.

*f) Model Evaluation.*: Evaluating the performance of a diffusion-based generative model is crucial to ensure that the generated images are both high-quality and culturally aligned with the desired attributes. To assess the effectiveness of our model, we employ a multi-faceted evaluation approach that includes both automated metrics (such as CLIP scores) and human feedback mechanisms.

Firstly, we leverage a CLIP-based evaluation method, where we utilize a CLIP model retrained on Vietnamese-specific data to compute similarity scores between the generated images and their corresponding textual prompts. This retrained CLIP model, which has been fine-tuned to recognize culturally relevant visual features (e.g., “Pho”, “Ao Dai”, and “Thap Rua”), is capable of evaluating how well the model’s output aligns with the semantic and visual cues embedded in the prompt. The CLIP score, a measure of cosine similarity between the image and text embeddings, serves as a quantitative metric for evaluating the degree of alignment between the generated content and the cultural context defined by the user’s prompt. Higher CLIP scores indicate better alignment with Vietnamese aesthetics and features, while lower scores highlight areas where the model may need further refinement in capturing cultural nuances.

In addition to CLIP-based evaluation, we also design a human feedback pipeline to assess the subjective quality and cultural fidelity of the generated images. This evaluation system allows human evaluators to rate the images on several key criteria, including but not limited to:

- **Cultural Authenticity:** Does the image reflect Vietnamese cultural elements accurately (e.g., traditional attire like Ao Dai, famous landmarks like Thap Rua, or iconic foods like Pho)?
- **Visual Aesthetics:** Is the image visually pleasing, with proper lighting, composition, and realistic rendering?
- **Prompt Consistency:** How well does the image adhere to the given text prompt? Are the key elements of the prompt (e.g., “a couple eating Pho”) clearly visible and well-represented in the image?
- **Clarity and Detail:** Are the image details sufficiently sharp, and does it avoid visual artifacts or blurring, especially in complex areas like faces or landmarks?

Evaluators rate each image on a scale from 1 to 5 for each criterion, with 1 indicating poor performance and 5 indicating excellent performance. These ratings are then aggregated to generate an overall quality score, which can be used to guide further refinements in the model’s training process. This human-centered evaluation allows for a more nuanced understanding of how well the model performs in real-world scenarios, especially when it comes to cultural sensitivity and subjective visual preferences.

To optimize the model based on human feedback, we integrate the ratings from the human feedback pipeline into

a reinforcement learning framework, specifically using Reinforcement Learning with Human Feedback (RLHF). The ratings from evaluators are treated as rewards or penalties for the model’s outputs, further refining the model’s ability to generate culturally accurate and aesthetically pleasing images. By incorporating both objective metrics (e.g., CLIP scores) and subjective human evaluation, we can ensure that the diffusion model produces outputs that are not only quantitatively accurate but also culturally relevant and visually appealing.

This dual evaluation approach—using both automated metrics and human feedback—provides a comprehensive framework for assessing the quality and relevance of the generated images. It ensures that the model is continuously improved and optimized for producing images that align with user expectations, both in terms of cultural fidelity and visual quality.

#### IV. CONCLUSION AND FUTURE WORK

In this article, we have conducted a comprehensive review and exploration of text-to-image generation with a strong focus on embedding rich Vietnamese cultural characteristics into the generative process. Beginning with a detailed analysis of existing general-purpose datasets and public Vietnamese datasets, we examined the limitations and potential of current image-captioning practices in constructing high-quality Vietnamese text-to-image (T2I) datasets. Furthermore, we reviewed the architecture and progression of modern diffusion-based text-to-image models, identifying both their strengths and cultural limitations when applied to underrepresented regions like Vietnam.

In the latter half of the paper, we proposed a structured and domain-specific pipeline aimed at enhancing the cultural fidelity of generated images. Central to this pipeline is a well-defined data taxonomy that classifies image content along several Vietnamese-specific dimensions: traditional customs, historical periods, regional diversity, festivals, and national identity. We expanded this taxonomy to cover specialized subsets, most notably a focused effort on visual representations of the Vietnam People’s Army (VPA), which encapsulates decades of military evolution, diverse unit-specific imagery, and high-context cultural elements.

Our work also proposed a practical generation pipeline integrating scene parsing, reference retrieval, and control-guided generation based on multimodal alignment. This approach not only enables stylistically consistent and culturally aware generation but also provides a robust framework for training and evaluating future Vietnamese-centric diffusion models.

**Future Work.** In the future, we plan to modularize this project by dividing the development into smaller, manageable components—such as focused dataset curation for each cultural domain (e.g., clothing, festivals, military), controlled model fine-tuning, and feedback-driven reward optimization—to evaluate feasibility at each step. By validating the cultural and generative quality incrementally, we aim to build a scalable foundation for a comprehensive Vietnamese T2I



diffusion framework that can serve as both a creative tool and a digital archive of national identity.

## REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [2] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, p. 307–392, 2019. [Online]. Available: <http://dx.doi.org/10.1561/22000000056>
- [3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: <https://arxiv.org/abs/1406.2661>
- [4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: <https://arxiv.org/abs/2112.10752>
- [5] OpenAI, "Improving image generation with better captions," <https://cdn.openai.com/papers/dall-e-3.pdf>, 2023, accessed: 2025-05-15.
- [6] OpenAI and Others, "Gpt-4 technical report," <https://arxiv.org/abs/2303.08774>, 2023, accessed: 2025-05-15.
- [7] G. DeepMind, "Gemini: A family of highly capable multimodal models," <https://arxiv.org/abs/2312.11805>, 2023, accessed: 2025-05-15.
- [8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
- [9] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, F. Momeni, S. Milani, P.-Y. Huang, I. Laptev *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022. [Online]. Available: <https://arxiv.org/abs/2204.14198>
- [10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. [Online]. Available: <https://arxiv.org/abs/2302.13971>
- [11] B. Yang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.10671>
- [12] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.
- [13] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," 2020. [Online]. Available: <https://arxiv.org/abs/2003.00744>
- [14] K. T. Doan, B. G. Huynh, D. T. Hoang, T. D. Pham, N. H. Pham, Q. T. M. Nguyen, B. Q. Vo, and S. N. Hoang, "Vintern-1b: An efficient multimodal large language model for vietnamese," 2024. [Online]. Available: <https://arxiv.org/abs/2408.12480>
- [15] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. [Online]. Available: <https://www.robots.ox.ac.uk/~vgg/data/flowers/102/>
- [16] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147. [Online]. Available: <https://hockenmaier.cs.illinois.edu/pubs/rashtchian-et-al10.pdf>
- [17] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in neural information processing systems*, 2011, pp. 1143–1151. [Online]. Available: <https://www.cs.unc.edu/~vso/publication/im2text.html>
- [18] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," 2016. [Online]. Available: <https://arxiv.org/abs/1505.04870>
- [19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *European conference on computer vision*, pp. 740–755, 2014. [Online]. Available: <https://cocodataset.org>
- [20] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [21] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," 2015. [Online]. Available: <https://arxiv.org/abs/1504.00325>
- [22] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433. [Online]. Available: <https://visualqa.org>
- [23] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," *arXiv preprint arXiv:1611.08321*, 2017. [Online]. Available: <https://arxiv.org/abs/1611.08321>
- [24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. E. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," in *International journal of computer vision*, vol. 123, no. 1, 2017, pp. 32–73. [Online]. Available: <https://visualgenome.org>
- [25] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017. [Online]. Available: <https://visualqa.org>
- [26] H. Zhang, W. Chen, J. Tian, Y. Wang, and Y. Jin, "Show, attend and translate: Unpaired multi-domain image-to-image translation with visual attention," 2019. [Online]. Available: <https://arxiv.org/abs/1811.07483>
- [27] H. Agrawal, K. Desai, Y. Wang, G. Chechik, A. Berg, D. Parikh, and D. Batra, "nocaps: novel object captioning at scale," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8948–8957.
- [28] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6720–6731.
- [29] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypemymed, image alt-text dataset for automatic image captioning," 2018. [Online]. Available: <https://arxiv.org/abs/1803.09123>
- [30] C. Schuhmann, R. Beaumont, R. Vencu, R. Wightman, M. Cherti, T. Coombes, R. Rombach, P. Jenicek, P. Wilke, C. Schulz *et al.*, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," 2021. [Online]. Available: <https://arxiv.org/abs/2111.02114>
- [31] W. Feng, X. He, T.-J. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Concept conjunction 500 (cc-500)," <https://conceptconjunction.github.io/>, 2024, dataset published by the authors.
- [32] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021. [Online]. Available: <https://github.com/google-research-datasets/conceptual-12m>
- [33] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models," *arXiv preprint arXiv:2210.14896*, 2022. [Online]. Available: <https://arxiv.org/abs/2210.14896>
- [34] J. Bitton, A. Diwan, N. Goyal, A. Kembhavi, A. Farhadi, D. Parikh, and D. Batra, "Winoground: Probing vision and language models for visio-linguistic compositionality," *arXiv preprint arXiv:2204.03162*, 2022. [Online]. Available: <https://arxiv.org/abs/2204.03162>
- [35] C. Saharia, W. Chang, J. Ho, T. Salimans, J. Ho, T. Salimans, J. Ho, T. Salimans, J. Ho, and T. Salimans, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022. [Online]. Available: <https://arxiv.org/abs/2205.11487>
- [36] Y. Zhang, P. Yu, and Y. N. Wu, "Abc-6k dataset," <https://doi.org/10.57702/nc9rsac4>, 2024, dataset released by the authors.
- [37] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," 2023. [Online]. Available: <https://arxiv.org/abs/2211.05105>



- [38] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 723–78 747, 2023. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract-Datasets\\_and\\_Benchmarks.html](https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract-Datasets_and_Benchmarks.html)
- [39] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," 2023. [Online]. Available: <https://arxiv.org/abs/2202.04053>
- [40] Y. Liang, J. He, G. Li, P. Li, A. Klimovskiy, N. Carolan, J. Sun, J. Pont-Tuset, S. Young, F. Yang, J. Ke, K. Dvijotham, K. Collins, Y. Luo, Y. Li, K. J. Kohlhoff, D. Ramachandran, and V. Navalpakkam, "Rich human feedback for text-to-image generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. [Online]. Available: <https://arxiv.org/abs/2312.10240>
- [41] J. Chen, Y. Huang, T. Lv, and F. Wei, "Textdiffuser: Diffusion models as text painters," *arXiv preprint arXiv:2305.10855*, 2023. [Online]. Available: <https://arxiv.org/abs/2305.10855>
- [42] S. Changpinyo, P. Sharma, N. Ding, and S. Goodman, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," *arXiv preprint arXiv:2102.08981*, 2021. [Online]. Available: <https://arxiv.org/abs/2102.08981>
- [43] A.-C. Pham, V.-Q. Nguyen, T.-H. Vuong, and Q.-T. Ha, "Ktvic: A vietnamese image captioning dataset on the life domain," *arXiv preprint arXiv:2401.08100*, 2024. [Online]. Available: <https://arxiv.org/abs/2401.08100>
- [44] Q. H. Lam, Q. D. Le, K. V. Nguyen, and N. L.-T. Nguyen, "Uit-viic: A dataset for the first evaluation on vietnamese image captioning," *arXiv preprint arXiv:2002.00175*, 2020. [Online]. Available: <https://arxiv.org/abs/2002.00175>
- [45] D. C. Bui, N. H. Nguyen, and K. Nguyen, "Uit-openviic: A novel benchmark for evaluating image captioning in vietnamese," *arXiv preprint arXiv:2305.04166*, 2023. [Online]. Available: <https://arxiv.org/abs/2305.04166>
- [46] dinhanhx, "VisualRoBERTa," 9 2022. [Online]. Available: <https://github.com/dinhanhx/VisualRoBERTa>
- [47] H. Q. Pham, T. K.-B. Nguyen, Q. V. Nguyen, D. Q. Tran, N. H. Nguyen, K. V. Nguyen, and N. L.-T. Nguyen, "Viocrvqa: Novel benchmark dataset and vision reader for visual question answering by understanding vietnamese text in images," *arXiv preprint arXiv:2404.18397*, 2024. [Online]. Available: <https://arxiv.org/abs/2404.18397>
- [48] Nexdata-AI, "4,995 vietnamese ocr images dataset," <https://www.nexdata.ai/datasets/ocr/1059?source=Github>, Nexdata-AI, China, 2024, accessed: 2025-05-15. [Online]. Available: <https://www.nexdata.ai/datasets/ocr/1059?source=Github>
- [49] J. Li, D. Yang, P. Su, C. Lu, X. Li, and Q. V. Le, "Blip-2: Bootstrapped language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.
- [50] V. A. Team, "Vistral-v-7b: Vietnamese vision-language model," 2024, <https://github.com/Vistral/vistral-v-7b>.
- [51] PaddlePaddle Authors, "Paddleocr: An open-source ocr system based on paddlepaddle," <https://github.com/PaddlePaddle/PaddleOCR>, 2021, accessed: 2025-05-15.
- [52] H. Nguyen, "Vietocr: Open-source optical character recognition for vietnamese," <https://github.com/duyquang/vietocr>, 2020, accessed: 2025-05-15.
- [53] S. AI and Collaborators, "Stable diffusion xl: Scaling latent diffusion models to 1 billion parameters," <https://stability.ai/blog/stable-diffusion-xl-release>, 2023, accessed: 2025-05-15.
- [54] S. AI, "Stable diffusion 3.5," <https://stability.ai/blog/stable-diffusion-3-5-release>, 2024, accessed: 2025-05-15.
- [55] L. Zhang, M. A. Yang, and A. A. Efros, "Controlnet: Adding conditional control to text-to-image diffusion models," 2023, <https://arxiv.org/abs/2302.05543>.
- [56] C. Yang, C. Liu, X. Deng, D. Kim, X. Mei, X. Shen, and L.-C. Chen, "1.58-bit flux," 2024. [Online]. Available: <https://arxiv.org/abs/2412.18653>
- [57] L. Hamdi, A. Tamasna, P. Boisson, and T. Paquet, "Vista-ocr: Towards generative and interactive end to end ocr models," 2025. [Online]. Available: <https://arxiv.org/abs/2504.03621>
- [58] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," 2021. [Online]. Available: <https://arxiv.org/abs/2010.11934>