KẾ HOẠCH TRIỂN KHAI

Phần Mềm Dịch Thuật Offline

CSV Translator Pro

Sử dụng mô hình AI TranslateGemma

Hỗ trợ 55 ngôn ngữ – Hoạt động hoàn toàn offline

Quy mô triển khai:	100 người dùng
Hạ tầng:	Mua mới GPU Server
Kiến trúc:	Client–Server (Web Application)
Mô hình AI:	TranslateGemma (4B / 27B tham số)

Ngày lập: Ngày 8 tháng 2 năm 2026

Mục lục

1 Tổng quan dự án

1.1 Mục tiêu triển khai

Triển khai hệ thống phần mềm dịch thuật offline **CSV Translator Pro** cho đơn vị với các mục tiêu:

1. **Dịch file CSV hàng loạt**: Hỗ trợ dịch tự động các file CSV chứa dữ liệu đa ngôn ngữ
2. **Dịch văn bản trực tiếp**: Cho phép nhập và dịch văn bản theo thời gian thực
3. **Dịch văn bản từ ảnh (OCR)**: Trích xuất và dịch văn bản nhúng trong hình ảnh
4. **Hoạt động offline hoàn toàn**: Không cần kết nối Internet sau khi triển khai
5. **Bảo mật dữ liệu**: Dữ liệu không rời khỏi mạng nội bộ đơn vị

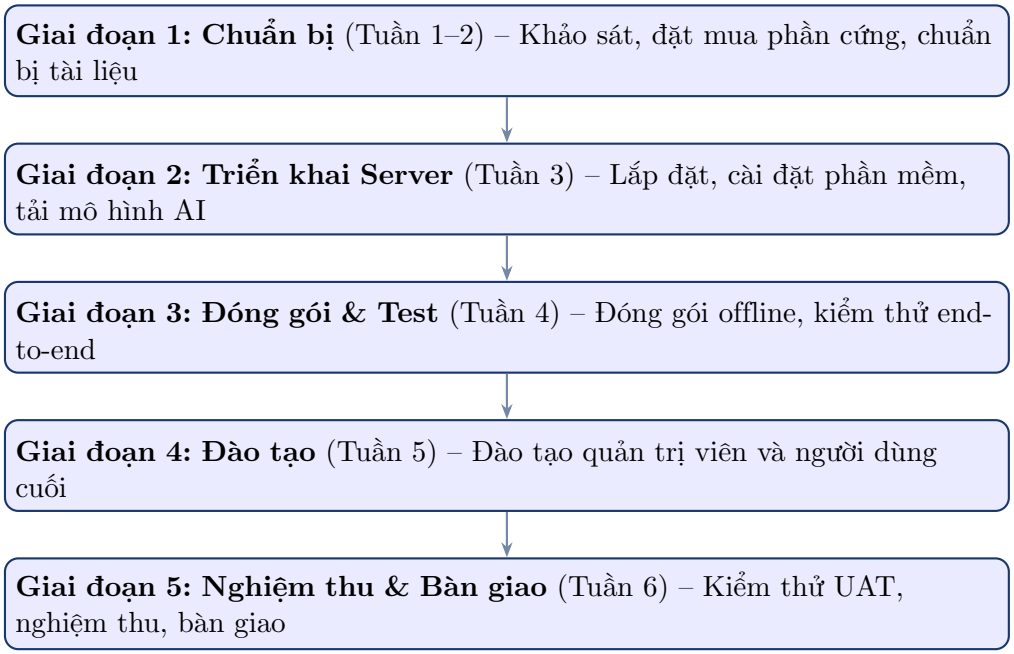
1.2 Phạm vi triển khai

Bảng 1: Phạm vi triển khai

Tiêu chí	Chi tiết
Số lượng người dùng	100 người
Số lượng Server	1 máy chủ GPU chuyên dụng
Ngôn ngữ hỗ trợ	55 ngôn ngữ (Ả Rập, Việt, Anh, Trung, Nhật, Hàn, v.v.)
Loại triển khai	Offline – Mạng nội bộ (LAN)
Thời gian dự kiến	4–6 tuần

2 Kế hoạch thời gian (Timeline)

2.1 Tổng quan các giai đoạn



Hình 1: Sơ đồ các giai đoạn triển khai

2.2 Chi tiết từng giai đoạn

2.2.1 Giai đoạn 1: Chuẩn bị (Tuần 1–2)

Bảng 2: Các công việc Giai đoạn 1

STT	Công việc	Thời gian	Phụ trách
1	Khảo sát hạ tầng mạng LAN hiện tại	Ngày 1–2	KTV Mạng
2	Khảo sát phòng đặt Server (điện, nhiệt độ)	Ngày 1–2	KTV Hạ tầng
3	Lập danh sách phần cứng cần mua	Ngày 3	Quản lý dự án
4	Phê duyệt ngân sách và đặt mua	Ngày 4–7	Ban lãnh đạo
5	Chuẩn bị tài liệu hướng dẫn sử dụng	Ngày 7–14	Đội triển khai
6	Nhận và kiểm tra phần cứng	Ngày 10–14	KTV Phần cứng

2.2.2 Giai đoạn 2: Triển khai Server (Tuần 3)

Bảng 3: Các công việc Giai đoạn 2

STT	Công việc	Thời gian	Phụ trách
1	Lắp đặt Server tại phòng máy	Ngày 1	KTV Phần cứng
2	Cài đặt hệ điều hành (Ubuntu/Windows Server)	Ngày 1–2	KTV Hệ thống
3	Cài đặt NVIDIA Driver + CUDA Toolkit	Ngày 2	KTV Hệ thống
4	Cài đặt Python, PyTorch, Dependencies	Ngày 2–3	KTV Phần mềm
5	Tải mô hình TranslateGemma từ Hugging Face	Ngày 3–4	KTV Phần mềm
6	Cấu hình mạng, firewall, IP tĩnh	Ngày 4–5	KTV Mạng
7	Khởi chạy và kiểm tra Server hoạt động	Ngày 5–7	Đội triển khai

2.2.3 Giai đoạn 3: Đóng gói & Kiểm thử (Tuần 4)

Bảng 4: Các công việc Giai đoạn 3

STT	Công việc	Thời gian	Phụ trách
1	Build Frontend và cấu hình API URL	Ngày 1	KTV Phần mềm
2	Đóng gói Offline Package (nếu cần)	Ngày 1–2	KTV Phần mềm
3	Kiểm thử dịch CSV với dữ liệu mẫu	Ngày 2–3	QA/Tester
4	Kiểm thử dịch văn bản và OCR	Ngày 3–4	QA/Tester
5	Kiểm thử với 10–20 người dùng đồng thời	Ngày 4–5	Đội triển khai
6	Sửa lỗi và tối ưu hiệu năng	Ngày 5–7	KTV Phần mềm

2.2.4 Giai đoạn 4: Đào tạo (Tuần 5)

Bảng 5: Các công việc Giai đoạn 4

STT	Công việc	Thời gian	Phụ trách
1	Đào tạo Quản trị viên (2–3 người)	Ngày 1–2	Đội triển khai
2	Đào tạo người dùng cuối (đợt 1: 50 người)	Ngày 3–4	Đào tạo viên
3	Đào tạo người dùng cuối (đợt 2: 50 người)	Ngày 5–6	Đào tạo viên
4	Hỗ trợ giải đáp thắc mắc	Ngày 6–7	Đội triển khai

2.2.5 Giai đoạn 5: Nghiệm thu & Bàn giao (Tuần 6)

Bảng 6: Các công việc Giai đoạn 5

STT	Công việc	Thời gian	Phụ trách
1	Kiểm thử UAT với đại diện các phòng ban	Ngày 1–3	QA + Đơn vị
2	Hoàn thiện tài liệu bàn giao	Ngày 3–4	Đội triển khai
3	Họp nghiệm thu và ký biên bản	Ngày 5	Các bên liên quan
4	Bàn giao và kết thúc dự án	Ngày 6–7	Quản lý dự án

3 Chi phí dự toán

3.1 Chi phí phần cứng Server

Bảng 7: Dự toán chi phí phần cứng Server GPU – NVIDIA A100 80GB

Thành phần	Thông số kỹ thuật	SL	Thành tiền (VNĐ)
Cấu hình Production: NVIDIA A100 80GB – Full Precision BF16			
GPU	NVIDIA A100 80GB PCIe (HBM2e, 2TB/s)	1	450.000.000
CPU	AMD EPYC 7313 (16C/32T, 3.0GHz)	1	55.000.000
Mainboard	Supermicro H12SSL-i (SP3 Socket)	1	25.000.000
RAM	DDR4-3200 ECC REG 128GB (8x16GB)	1 bộ	35.000.000
SSD	NVMe Gen4 2TB (Samsung PM9A3 Enterprise)	1	12.000.000
PSU	1200W 80+ Platinum Redundant	1	8.000.000
Case	4U Rackmount Server Chassis	1	10.000.000
Cooling	Hệ thống tản nhiệt Server	1	5.000.000
UPS	3000VA Online Double Conversion	1	25.000.000
Tổng chi phí phần cứng			**625.000.000**

Ưu điểm NVIDIA A100 80GB – Full Precision

- **VRAM 80GB HBM2e**: Chạy mô hình TranslateGemma-27B với **Full Precision BF16** – chất lượng dịch tối ưu nhất
- **Không cần Quantization**: Không giảm chất lượng do INT8/INT4, giữ nguyên độ chính xác mô hình
- **Tensor Cores thế hệ 3**: Tăng tốc inference AI lên đến 312 TFLOPS (TF32)
- **Băng thông 2TB/s**: Giảm bottleneck khi xử lý batch lớn
- **Thiết kế Datacenter**: Hoạt động 24/7, bảo hành enterprise, độ tin cậy cao

- **ECC Memory**: Đảm bảo tính toàn vẹn dữ liệu trong quá trình suy luận

So sánh hiệu năng mô hình 27B

GPU	Precision	Tokens/giây	Chất lượng
RTX 4090 (24GB)	INT4 (4-bit)	~20–35	Giảm nhẹ
RTX A6000 (48GB)	INT8 (8-bit)	~25–40	Tốt
A100 (80GB)	**BF16 (Full)**	**~40–70**	**Tối ưu**

3.2 Chi phí triển khai và đào tạo

Bảng 8: Dự toán chi phí triển khai

Hạng mục	Mô tả	Chi phí (VNĐ)
Triển khai phần mềm	Cài đặt, cấu hình, kiểm thử (2 tuần)	30.000.000
Đào tạo quản trị viên	2–3 người, 2 ngày	5.000.000
Đào tạo người dùng	100 người, 2 đợt	10.000.000
Tài liệu hướng dẫn	Biên soạn, in ấn	3.000.000
Hỗ trợ kỹ thuật ban đầu	1 tháng sau nghiệm thu	5.000.000
Tổng chi phí triển khai		**53.000.000**

3.3 Tổng hợp chi phí

Bảng 9: Tổng hợp chi phí dự án – NVIDIA A100 80GB

Hạng mục	Chi phí (VNĐ)
Phần cứng Server (A100 80GB + hệ thống)	625.000.000
Triển khai & Đào tạo	53.000.000
Dự phòng (10%)	67.800.000
TỔNG CỘNG	**745.800.000**

Lợi ích đầu tư NVIDIA A100 80GB

Với tổng chi phí **745,8 triệu VNĐ**, đơn vị được:

- **Chất lượng dịch tối ưu**: Mô hình 27B chạy Full Precision BF16, không mất độ chính xác
- **Hiệu năng cao**: 40–70 tokens/giây, xử lý nhanh file CSV lớn

- **Độ tin cậy enterprise**: GPU datacenter, hoạt động 24/7, bảo hành dài hạn
- **Không giới hạn VRAM**: 80GB HBM2e đủ cho mọi tình huống sử dụng
- **Khả năng mở rộng**: Có thể nâng cấp lên multi-GPU hoặc model lớn hơn trong tương lai

4 Phân công nhân sự

4.1 Cơ cấu đội dự án

Bảng 10: Danh sách nhân sự dự án

STT	Vai trò	Trách nhiệm	Số lượng
1	Quản lý dự án (PM)	Điều phối, báo cáo tiến độ, giao tiếp các bên	1
2	KTV Phần cứng	Lắp đặt Server, kiểm tra thiết bị	1
3	KTV Hệ thống	Cài đặt OS, Driver, CUDA	1
4	KTV Phần mềm	Cài đặt ứng dụng, cấu hình, debug	1–2
5	KTV Mạng	Cấu hình mạng, firewall, IP	1
6	QA/Tester	Kiểm thử chức năng, hiệu năng	1
7	Đào tạo viên	Đào tạo người dùng cuối	1–2
		Tổng cộng	**8–10 người**

4.2 Ma trận phân công RACI

Bảng 11: Ma trận RACI (R=Responsible, A=Accountable, C=Consulted, I=Informed)

Công việc	PM	HW	SYS	SW	NET	QA	ĐT
Khảo sát hạ tầng	A	R	C	C	R	I	I
Đặt mua phần cứng	A/R	C	I	I	I	I	I
Lắp đặt Server	A	R	C	I	I	I	I
Cài đặt OS/CUDA	A	I	R	C	I	I	I
Cài đặt phần mềm	A	I	C	R	I	I	I
Cấu hình mạng	A	I	C	I	R	I	I
Kiểm thử hệ thống	A	I	C	C	C	R	I
Đào tạo người dùng	A	I	I	C	I	I	R
Nghiệm thu	A/R	I	I	C	I	C	I

4.3 Yêu cầu từ phía đơn vị

Bảng 12: Nhân sự đơn vị cần phối hợp

Vai trò	Trách nhiệm	Số lượng
Đầu mối phối hợp	Điều phối nội bộ, phê duyệt	1
Quản trị viên IT	Tiếp nhận bàn giao, vận hành hệ thống	2–3
Đại diện phòng ban	Tham gia UAT, phản hồi yêu cầu	5–10

5 Kế hoạch đào tạo

5.1 Đối tượng đào tạo

Bảng 13: Phân loại đối tượng đào tạo

Đối tượng	Số lượng	Mục tiêu đào tạo	Thời lượng
Quản trị viên	2–3 người	Khởi động/dừng Server, xử lý sự cố, backup, restore	1 ngày (8h)
Người dùng cuối	100 người	Sử dụng giao diện web dịch CSV, text, OCR	2 giờ/đợt

5.2 Nội dung đào tạo Quản trị viên

1. **Kiến trúc hệ thống** (1 giờ)
   - Mô hình Client–Server
   - Các thành phần: Backend (FastAPI), Frontend (React), Model AI
2. **Vận hành Server** (3 giờ)
   - Khởi động/dừng dịch vụ Backend
   - Kiểm tra trạng thái GPU, RAM, CPU
   - Xem log và xử lý lỗi cơ bản
3. **Xử lý sự cố** (2 giờ)
   - Server không phản hồi
   - Lỗi GPU out of memory
   - Lỗi mạng/kết nối
4. **Backup và Restore** (2 giờ)
   - Backup model cache
   - Backup cấu hình
   - Khôi phục khi cần

5.3 Nội dung đào tạo Người dùng cuối

1. **Truy cập hệ thống** (15 phút)
   - URL truy cập (VD: `http://192.168.x.x:8000`)
   - Không cần đăng nhập
2. **Dịch file CSV** (45 phút)
   - Chuẩn bị file CSV đúng định dạng (cột “Text”)

- Upload file và chọn ngôn ngữ nguồn/đích
- Theo dõi tiến trình và tải file kết quả

3. **Dịch văn bản trực tiếp** (30 phút)
   - Nhập văn bản cần dịch
   - Chọn ngôn ngữ và nhấn Dịch
   - Copy kết quả
4. **Dịch từ ảnh (OCR)** (30 phút)
   - Upload ảnh chứa văn bản
   - Hệ thống nhận dạng và dịch tự động

5.4 Tài liệu đào tạo

- Sổ tay Quản trị viên (PDF, 20–30 trang)
- Hướng dẫn sử dụng nhanh (PDF, 5–10 trang)
- Video hướng dẫn (10–15 phút)
- FAQ – Câu hỏi thường gặp

6 Đánh giá rủi ro và phương án dự phòng

6.1 Ma trận rủi ro

Bảng 14: Ma trận đánh giá rủi ro (Xác suất × Tác động)

ID	Rủi ro	XS	TĐ	Mức	Phương án giảm thiểu
R1	Chậm giao hàng phần cứng GPU	TB	Cao	Cao	Đặt hàng sớm, có nhà cung cấp dự phòng
R2	GPU lỗi hoặc không tương thích	Thấp	Cao	TB	Kiểm tra kỹ trước khi mua, bảo hành
R3	Mô hình AI quá nặng, thiếu VRAM	TB	Cao	Cao	Sử dụng quantization 4-bit, chọn model 4B
R4	Mạng LAN không ổn định	Thấp	TB	Thấp	Kiểm tra và nâng cấp mạng trước
R5	Người dùng khó tiếp cận công nghệ	TB	Thấp	Thấp	Đào tạo kỹ, tài liệu đơn giản
R6	Mất điện đột ngột	Thấp	Cao	TB	UPS, generator dự phòng
R7	Nhân sự triển khai không đủ	Thấp	TB	Thấp	Lên kế hoạch nhân sự sớm
R8	Bảo mật: Truy cập trái phép	Thấp	Cao	TB	Firewall, chỉ mở trong LAN, không public

Chú thích: XS = Xác suất, TĐ = Tác động, TB = Trung bình

6.2 Phương án dự phòng chi tiết

6.2.1 R1: Chậm giao hàng GPU

- **Phòng ngừa**: Đặt hàng ngay từ tuần 1, liên hệ 2–3 nhà cung cấp
- **Xử lý**: Nếu chậm > 1 tuần, sử dụng GPU cloud tạm thời (AWS, GCP)
- **Trách nhiệm**: Quản lý dự án

6.2.2 R3: Thiếu VRAM cho mô hình AI

- **Phòng ngừa**: Chọn GPU đủ VRAM (RTX A6000 48GB)
- **Xử lý**: Chuyển sang model 4B (thay vì 27B), sử dụng quantization 4-bit
- **Tác động**: Chất lượng dịch giảm nhẹ nhưng vẫn hoạt động

6.2.3 R6: Mất điện đột ngột

- **Phòng ngừa**: UPS 1500VA, duy trì 15–20 phút để shutdown an toàn
- **Xử lý**: Nếu mất điện kéo dài, thông báo người dùng dừng sử dụng
- **Khuyến nghị**: Đơn vị có hệ thống generator dự phòng

7 Kế hoạch vận hành và bảo trì

7.1 Vận hành hàng ngày

Bảng 15: Checklist vận hành hàng ngày

STT	Công việc	Tần suất
1	Kiểm tra Server đang hoạt động (truy cập web)	Mỗi sáng
2	Kiểm tra nhiệt độ GPU ($<$ 80°C khi hoạt động)	1 lần/ngày
3	Kiểm tra dung lượng ổ cứng (thư mục uploads/outputs)	1 lần/ngày
4	Xóa file tạm cũ ($>$ 7 ngày)	Hàng tuần

7.2 Bảo trì định kỳ

Bảng 16: Lịch bảo trì định kỳ

Tần suất	Công việc	Phụ trách
Hàng tuần	Backup cấu hình hệ thống	Quản trị viên
Hàng tháng	Cập nhật OS security patches	Quản trị viên
Hàng tháng	Kiểm tra log lỗi, tối ưu hiệu năng	Quản trị viên
Hàng quý	Vệ sinh Server (bụi, quạt, keo tản nhiệt)	KTV Phần cứng
Hàng năm	Đánh giá nâng cấp phần cứng/phần mềm	Đội CNTT

7.3 Hỗ trợ kỹ thuật

- **Hỗ trợ cấp 1** (Quản trị viên nội bộ): Xử lý các sự cố cơ bản, khởi động lại dịch vụ
- **Hỗ trợ cấp 2** (Đội triển khai): Hỗ trợ từ xa cho các sự cố phức tạp (miễn phí 1 tháng sau nghiệm thu)
- **Hỗ trợ mở rộng**: Hợp đồng bảo trì hàng năm (nếu cần)

8 Kết luận

8.1 Tóm tắt kế hoạch

Bảng 17: Tóm tắt kế hoạch triển khai

Tiêu chí	**Nội dung**
Thời gian triển khai	6 tuần
Tổng ngân sách	745.800.000 VNĐ
GPU Server	NVIDIA A100 80GB – Full Precision BF16
Nhân sự triển khai	8–10 người
Số người dùng	100 người
Mô hình AI	TranslateGemma-27B (Full Precision)

8.2 Các bước tiếp theo

1. Phê duyệt kế hoạch và ngân sách
2. Đặt mua phần cứng GPU Server
3. Thành lập đội dự án và kick-off meeting
4. Bắt đầu Giai đoạn 1: Chuẩn bị

8.3 Thông tin liên hệ

Vai trò	**Thông tin**
Quản lý dự án	[Tên] – [Email] – [SĐT]
Kỹ thuật trưởng	[Tên] – [Email] – [SĐT]

— HẾT —