

TÀI LIỆU HƯỚNG DẪN

# CẨM NANG CÀI ĐẶT & SỬ DỤNG

## RETAINPDF BY WXYHGK

Dịch PDF Giữ Nguyên Bố Cục: PDF Quét/Ảnh • Công Thức Nội Dòng • Bảo Vệ Mã Nguồn • Thư Viện & Hỏi Đáp AI

★ **GitHub Repository:** <https://github.com/wxyhgk/retain-pdf> (~2.300 Stars)  
👑 **Tác giả:** [wxyhgk](#) — giấy phép MIT, mã nguồn/API/cách triển khai hoàn toàn mở  
📄 **README có bản Tiếng Việt:** [README.vi.md](#) 💬 Nhóm QQ: **1101779791**

## 1. RETAINPDF LÀ GÌ & TẠI SAO ĐÁNG CHÚ Ý?

RetainPDF hiện là **dự án mã nguồn mở duy nhất** dành cho việc dịch PDF dạng ảnh và bản quét mà vẫn giữ nguyên bố cục, với chất lượng dịch thuật và sắp chữ được tác giả mô tả là ngang bằng, thậm chí vượt qua các sản phẩm thương mại cùng loại. Dự án dùng thuật toán sắp chữ tự phát triển để khôi phục công thức phức tạp và bố cục bài báo nhiều cột.

- **Công thức nội dòng — điểm mạnh nhất:** theo chính tác giả, đây là lợi thế vượt trội của RetainPDF — công thức, văn bản xung quanh và bố cục nội dòng vẫn được giữ ổn định sau khi dịch, điều mà các dự án dịch PDF mã nguồn mở khác hiện chưa làm được.
- **Hệ thống dịch tự phát triển:** xử lý chuyên biệt cho nội dung chạy qua nhiều cột hoặc nhiều trang, câu bị ngắt và đoạn văn nối tiếp — khôi phục đơn vị ngữ nghĩa hoàn chỉnh trước khi dịch, thay vì dịch từng khung OCR rời rạc gây mất ngữ cảnh.
- **Bảo vệ mã nguồn & công thức:** có cơ chế tránh dịch nhầm code block, kiểm soát bảng có thể tùy chọn bật/tắt, và chiến lược dịch tùy chỉnh theo quy tắc.
- **Thư viện tài liệu:** quản lý bài báo, sách và tài liệu quét trong cùng một thư viện, theo dõi trạng thái dịch, bộ sưu tập và mục yêu thích.
- **Đọc đối chiếu & đọc Markdown:** PDF gốc và bản dịch hiển thị cạnh nhau để kiểm tra công thức, hình ảnh, trích dẫn; hoặc đọc PDF cùng Markdown đã trích xuất, vẫn giữ công thức và hình ảnh.
- **Hỏi đáp AI theo tài liệu:** đặt câu hỏi trực tiếp về tài liệu hiện tại, nhận câu trả lời kèm trích dẫn theo trang; chuyển rõ ràng sang PDF Agent khi cần chỉnh sửa PDF.
- **Mã nguồn, API và cách triển khai đều mở:** hỗ trợ tự triển khai (self-host) và phát triển mở rộng, giấy phép MIT.

## 2. SO SÁNH NĂNG LỰC CỐT LÕI (THEO README CHÍNH THỨC)

| Dự án            | PDF quét | Công thức nội dòng phức tạp | Không dịch nhầm mã nguồn | Giữ bố cục / Tối ưu PDF |
|------------------|----------|-----------------------------|--------------------------|-------------------------|
| PDFMathTranslate | ✗        | ✗                           | ✗                        | Trung bình              |
| PolyglotPDF      | ✗        | ✗                           | ✗                        | Trung bình              |

| Dự án              | PDF quét | Công thức nội dòng phức tạp | Không dịch nhảm mã nguồn | Giữ bố cục / Tối ưu PDF       |
|--------------------|----------|-----------------------------|--------------------------|-------------------------------|
| Doc2X (đóng nguồn) | ✓        | ✓                           | ✗                        | Tốt / API không mở            |
| RetainPDF          | ✓        | ✓ Bảo toàn vượt trội        | ✓                        | Tốt, tối ưu liên tục + API mở |

### 3. YÊU CẦU CHUẨN BỊ TRƯỚC KHI CÀI

- **Ứng dụng desktop:** không cần cài thêm gì khác — chỉ cần tải đúng bản theo hệ điều hành từ GitHub Releases.
- **Docker & Docker Compose:** cần thiết cho phương án triển khai server/self-host.
- **Git:** để clone repo khi triển khai Docker hoặc build từ mã nguồn.
- **Một model dịch (LLM) và (tuỳ chọn) một OCR provider:** để chạy tác vụ dịch — cấu hình trong lúc dùng ứng dụng hoặc qua API (xem mục 5).

### 4. HƯỚNG DẪN CÀI ĐẶT CHI TIẾT

#### Cách 1 — Ứng dụng Desktop (khuyến nghị dùng cá nhân)

Tải gói phù hợp từ [GitHub Releases](#):

- Windows: Setup.exe
- macOS: .dmg
- Linux: .deb

**Lỗi "App is damaged" trên macOS** (do chưa có tài khoản Apple Developer để ký notarize): đây không phải lỗi file hỏng thật, mà do cơ chế xác minh chữ ký của hệ thống. Sau khi kéo ứng dụng vào **/Applications**, chạy:

```
sudo xattr -r -d com.apple.quarantine /Applications/RetainPDF.app
```

#### Cách 2 — Triển khai Docker (khuyến nghị cho team & self-hosting)

```
git clone https://github.com/wxyhgk/retain-pdf.git
cd retain-pdf/ops/deployment/docker/delivery
docker compose up -d
```

Sau khi khởi động, mở Web UI tại <http://127.0.0.1:40001>. Cập nhật lên bản mới:

```
docker compose pull
docker compose up -d
```

Chi tiết các tuỳ chọn cấu hình khác (build từ mã nguồn, biến môi trường, cổng dịch vụ...) xem trong [hướng dẫn triển khai Docker](#) của repo.

### 5. TỰ ĐỘNG HOÁ QUA API (THAM KHẢO KỸ THUẬT)

RetainPDF cung cấp REST API cho cả xử lý đồng bộ (một lần gọi, trả file ngay) và bất đồng bộ (tạo job, theo dõi tiến trình). Ví dụ gọi API dịch PDF đồng bộ ("bundle", trả file ZIP kết quả ngay trong một request):

```
curl -X POST "http://127.0.0.1:40001/api/v1/translate/bundle" \
-H "X-API-Key: your-api-key" \
-F "file=@/path/to/paper.pdf" \
-F "ocr_provider=paddle" \
-F "paddle_token=your-paddle-token" \
-F "base_url=https://api.openai.com/v1" \
-F "api_key=your-openai-api-key" \
-F "model=gpt-4o" \
-F "mode=sci" \
-F "workers=100" \
-F "batch_size=1" \
-o translated_bundle.zip
```

Luồng xử lý bất đồng bộ (phù hợp file lớn hoặc tích hợp vào hệ thống khác):

- 1. POST /api/v1/uploads — tải tài liệu lên
- 2. POST /api/v1/jobs — tạo job xử lý
- 3. GET /api/v1/jobs/{job\_id} — theo dõi trạng thái & chẩn đoán
- 4. Tải kết quả: GET /api/v1/jobs/{job\_id}/pdf (PDF đã dịch), /markdown?raw=true (Markdown đã trích xuất & dịch), /download (ZIP đầy đủ kèm dữ liệu debug)

Ba tham số quan trọng trong request: **mode** (chiến lược dịch — **fast**, **sci**, **precise**), **ocr\_provider** (ví dụ PaddleOCR), và bộ **base\_url/api\_key/model** để trở tới bất kỳ model tương thích API OpenAI. Chi tiết đầy đủ tham số, mã lỗi và schema phản hồi nằm trong tài liệu API của repo (mục 6).

## 6. KIẾN TRÚC & TÀI LIỆU CHUYÊN SÂU

Pipeline xử lý gồm 3 tầng chính: (1) Layout & OCR Engine (PaddleOCR / MinerU / Vision model) sinh ra schema thống nhất document.v1.json → (2) LLM Translation & Protection (nối đoạn, bảo vệ token/AST cho LaTeX/code, tiêu thuật ngữ, điều phối đa agent song song) → (3) Typographic Layout & Re-synthesis (tự co giãn font, dựng lại toạ độ vector qua Typst, tối ưu và xuất đa định dạng).

| Thư mục / Tài liệu     | Nội dung                                                                                                                                     |
|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| frontend/              | Giao diện Web, trình đọc PDF, điều phối phía client.                                                                                         |
| backend/               | API gateway (Rust, hiệu năng cao), pipeline xử lý lỗi bằng Python (chuẩn hoá OCR, điều phối dịch, render qua Typst) — xem backend/README.md. |
| ops/deployment/docker/ | Dockerfile và Docker Compose để triển khai production.                                                                                       |
| docs/                  | Tài liệu dự án, hướng dẫn phát triển — xem docs/README.md.                                                                                   |
| CONTRIBUTING.md        | Hướng dẫn đóng góp code cho dự án.                                                                                                           |

## 7. LƯU Ý VỀ NGUỒN

- Chỉ nên cài từ đúng repo chính thức:** [github.com/wxyhgz/retain-pdf](https://github.com/wxyhgz/retain-pdf). Repo có tồn tại dưới các tên hoặc tài khoản khác với nội dung tương tự (bản cũ hơn, ít tính năng và cấu trúc thư mục khác) — nên ưu tiên tài liệu và bản Release từ đúng tài khoản wxyhgz để có bản mới nhất và đầy đủ tính năng nhất (thư viện, đọc đối chiếu, hỏi đáp AI, PDF Agent).

- **Cộng đồng hỗ trợ:** dự án có nhóm QQ chính thức (mã nhóm 1101779791) để trao đổi khi gặp vấn đề trong lúc dùng, triển khai hoặc phát triển thêm.

***Chúc bạn dịch tài liệu khoa học và sách giáo trình giữ nguyên bố cục thành công cùng RetainPDF!***

---

© Kênh AI Mỗi Ngày | Hướng dẫn AI & Linux Tự Động Hóa

Nội dung kỹ thuật tổng hợp từ tài liệu chính thức (bản README.vi.md) của repo [github.com/wxyhgk/retain-pdf](https://github.com/wxyhgk/retain-pdf).