

PHAMMINHQUY.COM

Quy trình tạo 1 video AI hoàn chỉnh — A đến Z

Cẩm nang 9 bước: từ ý tưởng đến video, giọng nói và hình ảnh khớp từng giây.

 Đăng ký kênh + comment 'TÀI LIỆU' để nhận bản PDF đầy đủ này.

Tiếng đi trước — Hình bám theo

Bí quyết để giọng và hình không bao giờ lệch

Và đây là lời hứa của video này. <#0.3#> Sự mượt mà bạn vừa thấy không phải phép màu, mà đến từ một nguyên tắc rất đơn giản: tiếng đi trước, hình bám theo. <#0.3#> Giọng đọc quyết định độ dài, còn hình ảnh thì tự xếp mình cho vừa khít. Giữ vững nguyên tắc này, video của bạn sẽ luôn khớp, dù dài hay ngắn. Rồi, giờ đi vào từng bước.

Chín bước của một tập

kịch bản

giọng nói

chia cảnh

mô tả ảnh

bảng nhân vật

sinh ảnh

cảnh động

dựng phim

SEO

Toàn bộ quy trình gồm chín bước nối tiếp nhau, chạy như một dây chuyền. <#0.3#> Đọc kịch bản, tạo giọng, chia cảnh, viết mô tả ảnh, dựng bảng nhân vật, sinh ảnh, thêm cảnh động, dựng phim, và cuối cùng là làm SEO. <#0.3#> Ba bước được tô sáng, tạo giọng, chia cảnh và sinh ảnh, chính là bộ ba giữ cho tiếng và hình khớp nhau. Hãy ghi nhớ ba cái tên đó, ta sẽ quay lại.

Một chủ đề, một dây chuyền

- Đầu vào: một dòng chủ đề
- Máy tự viết kịch bản tiếng Việt
- Mỗi cảnh tự sinh giọng và hình
- Đầu ra: video, thumbnail, mô tả
- Bạn chỉ đạo diễn và tinh chỉnh

Mô hình tư duy rất gọn. Bạn đưa vào một dòng chủ đề, ví dụ mũ sừng của người Viking. <#0.3#> Máy sẽ viết kịch bản, cắt thành các cảnh, mỗi cảnh tự sinh giọng đọc và hình minh họa, rồi ghép lại thành phim kèm thumbnail và mô tả. <#0.3#> Vai trò của bạn không phải làm thủ công từng khâu, mà là làm đạo diễn: đưa chủ đề, duyệt chất lượng, và tinh chỉnh những chỗ chưa ưng ý.

Chọn một niềm tin sai

- Ngách: gỡ điều ai cũng tin nhầm
- Mỗi tập một niềm tin phổ biến
- Bẻ thành năm cú lật nhỏ
- Mỗi cú lật phải có nguồn thật
- Cấm bịa số, cấm nói suông

Bước một, ý tưởng. Kênh này chọn ngách phá vỡ niềm tin sai, tức là gỡ những điều gần như ai cũng tin nhưng thực ra không đúng. <#0.3#> Mỗi tập nhắm một niềm tin phổ biến, rồi bẻ nó thành năm cú lật nhỏ, mỗi cú lật là một bằng chứng. <#0.3#> Nguyên tắc sống còn ở đây là: mỗi con số, mỗi khẳng định đều phải có nguồn thật. Làm nội dung mà bịa, khán giả sẽ rời đi và niềm tin mất sạch.

“[Thứ hằng ngày] Đang Lừa Bạn?” — Công thức tiêu đề có tỉ lệ bấm cao nhất kênh

Tiêu đề quan trọng ngang nội dung, vì nó quyết định người ta có bấm vào hay không. <#0.3#> Công thức thắng nhất của kênh rất đơn giản: lấy một thứ hằng ngày, rồi hỏi thẳng, nó đang lừa bạn phải không. <#0.3#> Có chữ bạn, có cáo buộc, có dấu hỏi mở. Nó chạm thẳng vào người xem, nhưng tuyệt đối không được lộ đáp án. Còn tò mò thì người ta mới bấm.

Luật viết lời đọc

- Bảy chương, nối bằng câu hỏi mở
- Mỗi câu mười tới hai hai từ
- Một cú lật cho mỗi cảnh
- Sần cứng: tối thiểu mười phút
- Kêu gọi đăng ký ở cuối bài

Khi viết lời đọc, có vài luật giúp giữ chân người xem. <#0.3#> Chia thành bảy chương, mỗi chương kết bằng một câu hỏi mở để người xem tò mò xem tiếp. Câu ngắn, mười tới hai mươi hai từ, đọc lên nghe tự nhiên. <#0.3#> Độ dài có sần cứng là mười phút, vì video quá ngắn thì thuật toán ít đề xuất. Và đừng quên một lời kêu gọi đăng ký ở cuối, viết cho duyên chứ đừng khô khan.

Tạo giọng đọc bằng AI

- Dịch vụ đọc tiếng Việt trên đám mây
- Một khoá API, không cài đặt nặng
- Một giọng cố định xuyên suốt kênh
- Mỗi cảnh xuất một file âm thanh
- Kèm file mốc thời gian từng từ

Bước hai, tạo giọng đọc. Kênh dùng một dịch vụ đọc tiếng Việt trên đám mây, chất lượng cao và ổn định. <#0.3#> Chỉ cần một khoá, không phải cài phần mềm nặng, và giữ nguyên một giọng cho mọi tập để khán giả nhận ra kênh. <#0.3#> Đây là chi tiết quan trọng nhất cho phần sau: mỗi cảnh được xuất ra một file âm thanh riêng, kèm theo một file ghi lại mốc thời gian của từng từ. Nhớ kỹ chi tiết này nhé.

Thông số giọng đang dùng

Thông số	Giá trị
Giọng	Nam · người dẫn podcast
Tốc độ	+10% (nhanh, gọn)
Mô hình	hỗ trợ thẻ ngắt nghỉ
Nhịp thực tế	~271 từ / phút

Đây là bộ thông số giọng đang chạy. Một giọng nam kiểu người dẫn podcast, tốc độ tăng mười phần trăm cho nhịp nhanh và gọn. <#0.3#> Mô hình hỗ trợ thẻ ngắt nghỉ để câu văn có cảm xúc.

<#0.3#> Nhưng con số quan trọng nhất nằm ở dòng dưới cùng: với cấu hình này, giọng đọc khoảng hai trăm bảy mươi mốt từ mỗi phút. Đây là con số bạn phải đo lại mỗi khi đổi giọng.

Ba mẹo giữ giọng tự nhiên

- Dùng thẻ ngắt để tạo nhịp thở
- Tránh thẻ cảm xúc, máy đọc nguyên chữ
- Đổi giọng thì đo lại nhịp đọc
- Độ dài giọng khoá độ dài cảnh

Ba mẹo nhỏ để giọng nghe tự nhiên. <#0.3#> Thứ nhất, dùng thẻ ngắt nghỉ để tạo khoảng lặng, câu văn sẽ có nhịp thở chứ không đều đều như máy. Thứ hai, đừng chèn thẻ cảm xúc kiểu vui hay buồn, vì máy sẽ đọc luôn cả chữ đó. <#0.3#> Thứ ba, mỗi lần đổi giọng hay tốc độ, phải đo lại nhịp đọc. Vì sao quan trọng đến vậy? Bởi chính độ dài của giọng sẽ khoá độ dài của từng cảnh, và đó là cửa ngõ vào phần hay nhất.

Đồng bộ Tiếng và Hình

Phần khán giả quan tâm nhất

Và đây rồi, phần được hỏi nhiều nhất. Làm sao để hình ảnh chuyển đúng lúc giọng nói, không sớm một giây, không trễ một nhịp? <#0.3#> Rất nhiều người dựng video AI bị lệch tiếng và hình, vì họ làm hình trước rồi cố nhồi tiếng vào cho vừa. <#0.3#> Ở đây ta làm ngược lại hoàn toàn, và đó chính là cả bí quyết. Hãy đi thật chậm qua bốn ý sau đây.

Một cảnh, một khối

- Một cảnh = một đoạn giọng đọc
- Đoạn giọng đó = một nhóm ảnh
- Giọng dài bao nhiêu, hình trải bấy nhiêu
- Không tiếng thừa, không hình thiếu

Ý thứ nhất, đơn vị của mọi thứ là một cảnh. <#0.3#> Mỗi cảnh có đúng một đoạn giọng đọc, và đúng một nhóm ảnh minh hoạ đi cùng nó. Giọng đọc của cảnh đó dài mười lăm giây, thì nhóm ảnh của cảnh đó cũng được trải ra đúng mười lăm giây. <#0.3#> Không bao giờ có chuyện tiếng còn mà hình đã hết, hay hình còn mà tiếng đã dứt. Tiếng và hình bị buộc chung vào một khung thời gian duy nhất.

Vì sao đo tiếng trước

- Bước tạo giọng chạy trước tiên
- Máy đo độ dài từng file tiếng
- Rồi mới chia số ảnh cho mỗi cảnh
- Dựa trên số giây thật, không ước lượng

Ý thứ hai, và đây là mấu chốt kỹ thuật. Bước chia cảnh luôn chạy sau bước tạo giọng, không bao giờ chạy trước. <#0.3#> Lý do: sau khi đã có file tiếng, máy đo chính xác cảnh này dài bao nhiêu giây, rồi mới quyết định chia thành mấy tấm ảnh và mỗi tấm giữ trong bao lâu. <#0.3#> Nó dựa trên số

giây có thật của giọng đọc, chứ không phải một con số đoán mò. Đó là lý do vì sao hình không bao giờ bị lệch.

Nhân vật nhép môi theo giọng

- File mốc thời gian ghi từng từ
- Máy suy ra miệng đóng hay mở
- Dựng khẩu hình khớp âm thanh
- Thêm chớp mắt, nhún thở cho sống

Ý thứ tư, dành cho nhân vật dẫn chuyện. Còn nhớ file mốc thời gian từng từ mình dặn ghi nhớ không? Chính nó cho máy biết tại mỗi khoảnh khắc, miệng nhân vật nên đóng hay mở, tròn hay dẹt. Từ đó máy dựng ra khẩu hình nhép đúng theo âm thanh. Cộng thêm vài cái chớp mắt và nhịp thở nhẹ, một nhân vật đứng yên bỗng như đang thật sự nói. Đó là nhép môi, và nó cũng bắt nguồn từ giọng.

Ghép hình, tiếng và hiệu ứng

- Đặt ảnh lên đúng đường tiếng
- Zoom, rê nhẹ cho ảnh tĩnh sống dậy
- Chuyển cảnh mờ dần, không giật
- Cân âm lượng đều toàn video
- Cảnh mở đầu có thể thành clip động

Bước bốn, dựng phim. Đây là lúc mọi thứ ghép lại với nhau. Ảnh được đặt lên đúng đường tiếng như ta vừa nói. Rồi thêm hiệu ứng zoom và rê nhẹ trên ảnh tĩnh, để khung hình lúc nào cũng chuyển động chứ không chết cứng. Các cảnh nối nhau bằng chuyển cảnh mờ dần cho mượt, âm lượng được cân đều để không chỗ to chỗ nhỏ. Tất cả chạy bằng một công cụ dựng nhẹ, không ngốn bộ nhớ dù video có dài.

0.8 giây độ dài mỗi lần chuyển cảnh — mượt mà nhưng không lê thê

Một chi tiết nhỏ nhưng tạo cảm giác chuyên nghiệp: mỗi lần chuyển cảnh chỉ kéo dài khoảng tám phần mười giây. Đủ dài để mắt thấy mượt, đủ ngắn để không làm chậm nhịp kể. Cộng với cú zoom nhẹ trên từng ảnh, video giữ được năng lượng đều đặn từ đầu tới cuối. Nhớ nhé, hiệu ứng không phải để khoe, mà để giữ mắt người xem dán chặt vào màn hình.

Tối ưu để được đề xuất

- Tiêu đề chạm thẳng người xem
- Mô tả kèm mốc thời gian từng chương
- Thẻ khoá đúng chủ đề
- Bình luận ghim tặng tài liệu
- Cắt Reel dọc làm phễu

Bước năm, tối ưu SEO, để video có cơ hội được đề xuất. <#0.3#> Tiêu đề viết theo công thức chạm thẳng người xem. Mô tả có kèm mốc thời gian từng chương, giúp cả khán giả lẫn thuật toán hiểu video nói gì. <#0.3#> Thêm thẻ khoá đúng chủ đề, và một bình luận ghim mời khán giả nhận tài liệu tặng kèm. Cuối cùng, cắt vài đoạn ngắn theo chiều dọc để làm phễu kéo người lạ về kênh.

BONUS — Cẩm nang thao tác: tự làm 1 video AI từ A đến Z

Phần này viết để bạn **in ra và làm theo** mà không cần tua lại video. Gồm: checklist 9 bước theo đúng thứ tự, các lệnh và prompt copy-paste được, danh sách lỗi thường gặp, và bảng thông số tham chiếu.

Checklist 9 bước (làm lần lượt từ trên xuống)

#	Bước	Bạn làm gì	Xong khi
1	Ý tưởng	Chọn 1 niềm tin sai phổ biến, bẻ thành 5 cú lật có nguồn	Có 5 cú lật + nguồn thật
2	Tiêu đề	Áp công thức "[thứ hằng ngày] Đang Lừa Bạn?" — không lộ đáp án	Có tiêu đề chạm người xem
3	Kịch bản	Viết 7 chương, câu 10–22 từ, sần cứng 10 phút, có CTA cuối	Script đạt số từ tối thiểu
4	Giọng (TTS)	Tạo giọng — mỗi cảnh 1 file tiếng + 1 file mốc thời gian từng từ	Có đủ file mp3 mỗi cảnh
5	Chia cảnh	ĐO độ dài mp3 → mới chia số ảnh mỗi cảnh (chạy SAU bước 4)	Mỗi cảnh biết cần mấy ảnh
6	Sinh ảnh	Khai báo DNA nhân vật → sinh ảnh từng shot theo cảnh	Ảnh nhất quán, đủ shot
7	Dựng phim	Ghép ảnh lên đường tiếng + Ken Burns + chuyển cảnh + cân âm	Có file MP4
8	SEO	Tiêu đề, mô tả + mốc chương, tags, bình luận ghim tặng tài liệu	meta đầy đủ
9	Thumbnail	Nhân vật to + chữ vàng 2–3 từ + nền phẳng đậm	Đọc được trong 0.5s

3 dòng tô vàng (bước 4–5–6) là "bộ ba đồng bộ". Giữ đúng thứ tự Giọng → Chia cảnh → Ảnh thì tiếng và hình không bao giờ lệch.

Vì sao tiếng và hình luôn khớp (nguyên tắc cốt lõi)

1 cảnh = 1 đoạn tiếng = 1 nhóm hình. Giọng cảnh đó dài bao nhiêu giây, nhóm ảnh được trải ra đúng bấy nhiêu giây.

Tiếng đi trước. Bước chia cảnh chạy SAU khi có file giọng — máy đo số giây THẬT rồi mới chia ảnh, không đoán mò.

Ảnh bám trục thời gian. Cả nhóm ảnh trải lên đường tiếng, người kể đối ý thì hình cũng vừa kịp đổi khung.

Nhân vật nhép môi. File mốc thời gian từng từ → suy ra miệng đóng/mở → dựng khẩu hình khớp âm thanh.

 Lệnh copy-paste (pipeline mẫu — kênh "Phá Vỡ Niềm Tin Sai")

Mục đích	Lệnh
Khởi tạo tập mới	<code>python tools/new_episode.py "Chủ đề" --nnn vi-003 --chapters 7</code>
Parse → TTS → chia cảnh	<code>python tools/autobuild.py vi-003 --until plan</code>
Điền mô tả ảnh (validate)	<code>python build.py vi-003 --step prompts</code>
Prompts → ảnh	<code>python tools/autobuild.py vi-003 --from prompts --until images</code>
Dựng MP4	<code>python build.py vi-003 --step compose</code>
SEO metadata	<code>python build.py vi-003 --step meta</code>
Soi thử 2 khung ảnh	<code>python build.py vi-003 --step images --scenes 1,4 --force</code>
Re-TTS 1 cảnh sau khi sửa	<code>python build.py vi-003 --force --step tts</code>

Với video hướng dẫn dạng slide (như video này) engine khác: `python "<ENGINE>/run.py" --project "<PROJECT>"` rồi `--bumper` (host nhép môi), `--thumbnail`, `--handout`.

 Khung prompt ảnh giữ nhân vật nhất quán

[DNA nhân vật] + [ACTION] + [SETTING] + [CAMERA] + [LIGHTING] + [STYLE ANCHOR]

- **DNA** = ngoại hình tĩnh (đầu/mặt/mắt/tay chân/trang phục) — đứng ĐẦU prompt, quyết định style.
- **KHÔNG** nhét action/cảm xúc/bối cảnh vào DNA (để phần ACTION lo).
- **Đổi DNA** → phải chạy lại bảng nhân vật (sheet) **TRƯỚC** khi sinh ảnh loạt.
- Guard chống lỗi: "exactly two arms" (thừa tay) · "ONE single full-frame, NOT a grid" (ảnh lưới).

Lỗi thường gặp & cách xử lý

Triệu chứng	Nguyên nhân	Cách xử lý
Video ngắn hụt so với mong đợi	Nhịp TTS thực nhanh hơn ước lượng	Đo lại wpm từ 1 cảnh (từ/giây×60), ghi đè script.wpm. Đừng nhồi chữ cho đủ giờ.
Tiếng và hình bị lệch	Làm hình trước rồi nhồi tiếng	Luôn TTS trước, chia cảnh sau — để máy đo giây thật.
Ảnh ra lưới/collage nhiều ô	Nhiều doodle inject vào 1 shot	Doodle chỉ trong prompt per-scene + guard "NOT a grid".
Ảnh thừa tay/chân	Model sinh thêm chi	Thêm guard "exactly two arms" trong style anchor.
TTS đọc to ký tự lạ / dấu ---	Ký tự điều khiển lọt vào narration	Lọc --- và comment HTML khỏi script; KHÔNG dùng --- trong cảnh.
Máy đọc nguyên chữ {happy}	Dùng emotion tag	Chỉ dùng thẻ ngắt <#0.3#>; tránh emotion tag.
Font tiếng Việt vỡ dấu trên thumbnail	Font serif (Georgia/Times)	Dùng font sans (Segoe/Arial/Cambria); caption đậm dùng Arial Black.
Chữ thumbnail tràn khung	Câu quá dài, tiếng Việt rộng ~20%	Mỗi dòng ≤ ~9–10 ký tự; chia 2–3 từ all-caps; bỏ phụ đề nếu tiêu đề nhiều dòng.

Bảng thông số tham chiếu nhanh

Hạng mục	Giá trị khuyến nghị
Nhịp giọng (rate +10%)	~271 từ / phút (đo lại khi đổi giọng)
Độ dài 11 phút	≈ 2.300 từ
Sàn cứng 10 phút	≈ 2.100 từ
Chuyển cảnh (crossfade)	0.8 giây
Ken Burns zoom	1.04 (nhẹ) → 1.10 (rõ)
Câu trong script	10–22 từ, kết bằng . hoặc ...
Thumbnail chữ	2–3 từ, chữ vàng viền đen, nền phẳng đậm

Quy trình chống rủi ro (khuyên làm)

- Dựng thử **2 cảnh** (1 host + 1 data) chạy trọn TTS+dựng phim **TRƯỚC** khi viết cả kịch bản.
- Soi 2 khung ảnh trước khi sinh cả tập.
- Sau TTS, đo tổng thời lượng thật — thiếu/dư thì sửa lời rồi re-TTS chỉ cảnh đổi.
- Đọc lại thumbnail + vài slide chữ dài nhất để chắc không tràn, dấu không vỡ.

 **Đăng ký kênh + comment 'TÀI LIỆU' để nhận bản PDF đầy đủ này.**

phamminhquy.com · tài liệu đi kèm video "Quy trình tạo 1 video AI hoàn chỉnh — A đến Z" ·  phamminhquy.com · 
quyphamminh@gmail.com