Video Prompt & Audio Pipeline Plan (Replicate)

A compact, end-to-end checklist to: (1) описати референс-відео, (2) перетворити опис у промпт під генератори (Veo/Kling/Pika/PixVerse), (3) зібрати аудіо-доріжку (voiceover, музика, SFX, субтитри), (4) відрендерити, перевірити й експортувати для TikTok/Reels/Shorts.

🎯Цілі

- Отримувати структурований промпт з відео-референсу за 1–2 хв.
- Мати стабільний шаблон промпта для різних T2V/I2V моделей.
- Швидко збирати аудіо: фраза «Nails.», фіксація губ, музичний луп і SFX.
- Вихідні формати для TikTok/Reels (вертикаль), YouTube Shorts.

🧩Інструменти (Replicate)

Video→Text / розуміння сцени - Qwen2-VL-7B Instruct (основний) - MiniCPM-V 4.0 (швидший/дешевший) - VideoLLaMA3 / Apollo 7B / CogVLM2-Video (альтернативи)

LLM для форматування промпта - GPT-4o або Claude Sonnet (будь-який якісний LLM на Replicate)

Генератори відео - Google Veo 3 (Fast/Pro), Kling v1.6 Pro, Minimax director, PixVerse v4, Pika 2.x

Аудіо - TTS: XTTS-v2 (Coqui) - Voice conversion: RVC v2 (за потреби) - Розділення стемів/витяг SFX: Demucs - Музика: MusicGen (лупи), Stable Audio Open (SFX/ембієнт) - ACR+сабтайтли: WhisperX/Whisper (.srt) - Ліпсинк: Wav2Lip або Lipsync-2

🏸Загальний пайплайн

1) **Інгест референсу**
☐ Обрізати до ключової сцени (5–10 с)
☐ Зафіксувати fps (23.976 або 30)
☐ Витягнути 8–16 реперних кадрів (thumbnails)

2) **Опис відео (video→text)**
☐ Подати відео/кадри в Qwen2-VL → отримати: - Об’єкти/персонажі, дії - Камера/крупність/рухи (бажано статична) - Освітлення/кольорова палітра/фон - Атрибути стилю (глянець, неон, хром) - Таймінг подій (секундний план)

3) **Перетворення у промпт (LLM)**
☐ Прогнати опис через LLM із **шаблоном промпта** (див. нижче)

☐ Згенерувати варіанти під: **Veo / Kling / Pika / PixVerse**
☐ Додати **negative cues**

4) **Генерація відео**
☐ t2v / i2v (stable camera, fixed angle)
☐ 5–8 с, 1080×1920 (вертикаль)
☐ seed/temperature зберегти в нотатках

5) **Аудіо-доріжка**
Опція A – зберегти натуру:
☐ Demucs → зняти музику/вокал → залишити потрібні шуми
Опція B – замінити:
☐ XTTS-v2: озвучення «Nails.» (0.6–0.9 с)
☐ Wav2Lip/Lipsync-2: синхрон губ (якщо видимі)
☐ MusicGen: 8-тактовий луп 100–120 BPM (без вокалу)
☐ Stable Audio Open: whoosh/шип/клац/«spray hiss» 1–2 с
☐ WhisperX: SRT (за потреби субтитрів)

6) **Зведення та експорт**
☐ Ducking: −6…−9 dB під voiceover
☐ Лімітинг: −1 dBTP; інтегровано ≈ −14 LUFS
☐ Експорт AAC 192–256 kbps, 48 kHz, стерео
☐ Контрольний перегляд на смартфоні (яскрава/темна теми)

7) **Доставлення**
☐ TikTok/Reels: 1080×1920, 23.976/30 fps, ≤ 30 с
☐ Обкладинка/thumbnail з ключовим кадром
☐ План публікацій: хук на 0–2 с, CTA на 5–7 с

🧩Шаблон системного промпта для LLM (форматування під генератори)

```
You are a Prompt Formatter for video generation models (Veo/Kling/Pika/
PixVerse).
Input: a noisy video description.
Output: a single JSON with these keys:
{
  "scene": "1–2 sentences of the setting/background",
  "subject": "who/what, wardrobe/makeup/props",
  "actions": ["ordered micro-beats with timestamps (s): 0–2, 2–4, ..."],
  "camera": {
    "framing": "e.g., static, locked-off, medium close-up",
```

```
    "lens": "e.g., 50mm look",
    "motion": "none/minimal",
    "transitions": "e.g., cross-dissolve @4.0s"
  },
  "lighting": "3–8 words (key/rim/fill, softbox, specular highlights)",
  "color_palette": ["deep blue", "neon cyan", "chrome silver"],
  "fx": ["glossy skin highlights", "holographic sheen"],
  "duration_sec": 6,
  "aspect_ratio": "9:16",
  "negatives": ["no extra hands", "no shaky cam", "no text overlay"],
  "generator_overrides": {
    "veo": {"style_preset": "beauty-gloss", "motion": "low"},
    "kling": {"camera_lock": true, "detail": "high"},
    "pika": {"guidance": 7.5, "seed": 12345}
  }
}
```

Приклад запиту до LLM (user-prompt)

```
Take this reference description and format it as the JSON above:
- A beauty macro: female model, glossy makeup, fixed camera, minimal motion.
- She slowly raises her right hand to eye level; at 3.0s says “Nails.” quickly
and clearly.
- Chrome/holographic accents, deep blue gradient backdrop, cinematic rim light.
- Add a cross-dissolve between 3.2–3.6s. Keep identity consistent; no extra
hands.
- Duration 6s, 9:16. Prepare negatives.
```

Шаблон генераторного промпта (текстове ядро)

Цей блок LLM підставляє в API конкретної моделі.

```
[SCENE] ${scene}
[SUBJECT] ${subject}
[ACTIONS] ${actions with timestamps}
[CAMERA] ${camera.framing}; lens ${camera.lens}; motion $
{camera.motion}; transitions ${camera.transitions}
[LIGHTING] ${lighting}
[COLORS] ${color_palette}
[FX] ${fx}
```

```
[DURATION] ${duration_sec}s, AR ${aspect_ratio}
[NEGATIVES] ${negatives}
```

Аудіо-бриф + SFX-лист

- **Voiceover**: «Nails.» (0.6–0.9 s), темп 1.05, жіночий, нейтральний, мінімальна реверберація.
- **Lipsync**: align frame @3.0 s, tolerance ±40 ms.
- **Music loop**: 8 тактів, 112 BPM, upbeat electronic pop, без вокалу.
- **SFX**: short whoosh (0.25 s) на жест; soft spray hiss (0.7 s) під «setting spray» сцени; camera click (0.15 s) на переході.
- **Mix**: VO −3 dBFS peak; music −16 LUFS short-term під VO; SFX −6 dBFS peak.

Готовий текст для MusicGen

```
upbeat electronic pop, 112 BPM, bright synth plucks, sidechain kick, 8-bar
seamless loop, no vocals
```

Готовий список SFX для Stable Audio Open

```
soft whoosh transition 250ms; glossy makeup spray hiss 700ms; subtle camera
shutter click 150ms; studio ambience loop 4s
```

Приклади API-викликів (псевдо)

Замініть `<model_owner/model>` і параметри на актуальні з вашого облікового запису Replicate.

1) Video→Text (Qwen2-VL)

```
curl -s -X POST https://api.replicate.com/v1/predictions
 -H "Authorization: Token $REPLICATE_API_TOKEN"
 -H "Content-Type: application/json"
 -d '{
   "version": "<qwen2-vl:hash>",
   "input": {"video": "https://.../ref.mp4", "prompt": "Describe objects,
actions, camera, lighting, color palette, timeline."}
 }'
```

2) LLM форматування

```
curl -s -X POST https://api.replicate.com/v1/predictions
 -H "Authorization: Token $REPLICATE_API_TOKEN"
 -H "Content-Type: application/json"
 -d '{
   "version": "<gpt-4o-or-claude:hash>",
   "input": {"prompt": "<SYSTEM_PROMPT>\n<USER_PROMPT_WITH_DESCRIPTION>"}
 }'
```

3) Відеогенерація (наприклад, Veo/Kling)

```
curl -s -X POST https://api.replicate.com/v1/predictions
 -H "Authorization: Token $REPLICATE_API_TOKEN"
 -H "Content-Type: application/json"
 -d '{
   "version": "<veo-3-fast:hash>",
   "input": {
     "prompt": "<FORMATTED_GENERATOR_PROMPT>",
     "duration": 6,
     "resolution": "1080x1920",
     "seed": 12345
   }
 }'
```

4) Аудіо: XTTS-v2 → «Nails.»

```
curl -s -X POST https://api.replicate.com/v1/predictions
 -H "Authorization: Token $REPLICATE_API_TOKEN"
 -H "Content-Type: application/json"
 -d '{
   "version": "<xtts-v2:hash>",
   "input": {"text": "Nails.", "speaker": "female_01", "speed": 1.05}
 }'
```

Контрольний список якості

- [ ] Камера нерухома, без «мигань»/плаваючого фону
- [ ] Рука/обличчя зберігають ідентичність між кадрами
- [ ] Хронометраж збігається з аудіо подіями (±40 ms)
- [ ] Немає зайвих кінцівок/артефактів/тексту
- [ ] Кольори: відповідність референсу (глибокий синій + неон/хром)

- [ ] Гучність VO/музики/SFX відповідає LUFS/дБ цілям

Троублшутинг

- **«Дрижить» кадр** → посилити негативи (no shaky cam), lock camera=true; зменшити motion.
- **Додаткові пальці/руки** → додати negatives; скоротити тривалість; підсилити «identity consistency» у тексті.
- **Промах ліпсинку** → повторний рендер VO ровно 0.6–0.9 s; зменшити attack/decay музики під VO.
- **Пересвіт/недосвіт** → уточнити lighting: soft key + rim, додаємо exposure hint.

Експорт і публікація

- Формат: 1080×1920, H.264 High, 20–30 Mbps; AAC 48 kHz 192–256 kbps
- Тривалість: 6–15 s; хук у перші 2 s
- Назва/теги: максимально конкретні; додати CTA у субтитри
- План: Пт—Пн—Ср—Пт—Нд → далі щоденно, якщо стабільна генерація

Нотатки / варіанти

- Варіант B: MiniCPM-V для дешевого опису + Claude для якості формату
- Варіант C: Pika 2.x для швидких ітерацій, після — upscale в інший рушій
- Окремий пресет під «setting spray» сцену (спрей-аудіо + краплі + rim light)