

PLIEGO DE PRESCRIPCIONES TÉCNICAS QUE HA DE REGIR EL CONTRATO DE SERVICIO DE PLATAFORMA PARA LA ELABORACIÓN DE MATERIAL AUDIOVISUAL PERSONALIZADO MEDIANTE TÉCNICAS Y MODELOS DE INTELIGENCIA ARTIFICIAL, PARA LA FUNDACIÓN PARA LA INVESTIGACIÓN BIOMÉDICA DEL HOSPITAL UNIVERSITARIO LA PAZ A ADJUDICAR POR PROCEDIMIENTO ABIERTO SIMPLIFICADO MEDIANTE PLURALIDAD DE CRITERIOS. El Proyecto PI25/01147 ha sido financiado por el Instituto de Salud Carlos III y cofinanciado por la Unión Europea. EXPEDIENTE PAS 45-2026.

ÍNDICE

1. CARACTERÍSTICAS GENERALES

1.1. Objeto del contrato.

1.2. Legislación.....

2. ESPECIFICACIONES TÉCNICAS DEL SUMINISTRO

PLIEGO DE PRESCRIPCIONES TÉCNICAS QUE HA DE REGIR EL CONTRATO DE SERVICIO DE PLATAFORMA PARA LA ELABORACIÓN DE MATERIAL AUDIOVISUAL PERSONALIZADO MEDIANTE TÉCNICAS Y MODELOS DE INTELIGENCIA ARTIFICIAL, PARA LA FUNDACIÓN PARA LA INVESTIGACIÓN BIOMÉDICA DEL HOSPITAL UNIVERSITARIO LA PAZ A ADJUDICAR POR PROCEDIMIENTO ABIERTO SIMPLIFICADO MEDIANTE PLURALIDAD DE CRITERIOS. El Proyecto PI25/01147 ha sido financiado por el Instituto de Salud Carlos III y cofinanciado por la Unión Europea. EXPEDIENTE PAS 45-2026.

1. CARACTERÍSTICAS GENERALES

El servicio tendrá por objeto proporcional la plataforma tecnológica para la elaboración de material audiovisual personalizado mediante técnicas y modelos de inteligencia artificial, destinado a su utilización posterior en el tratamiento de pacientes con afasia postictus en el marco del ensayo clínico PI25/01147.

2. ESPECIFICACIONES TÉCNICAS DEL SUMINISTRO

El proveedor deberá proporcionar la infraestructura para la ejecución del proceso completo de producción de los videoclips, desde la generación de los diferentes elementos audiovisuales hasta la obtención del vídeo final, de acuerdo con las indicaciones proporcionadas por el equipo investigador. La personalización de los contenidos se realizará en función de las características, preferencias e indicaciones definidas para cada paciente o grupo de pacientes.

En particular, el servicio deberá proporcionar los recursos tecnológicos y la capacidad de procesamiento necesarios para permitir la ejecución de, al menos, las siguientes actividades:

- Diseño y preparación del contenido audiovisual. El servicio deberá permitir configurar y preparar los parámetros necesarios para la generación de cada videoclip, incluyendo la definición de personajes, escenarios, acciones, diálogos, duración, secuencia narrativa y demás elementos que determinen el contenido audiovisual a generar.
- Generación de personajes y escenarios mediante inteligencia artificial. El servicio deberá proporcionar acceso a modelos y herramientas de inteligencia artificial capaces de generar personajes, fondos, escenarios y otros elementos visuales a partir de las especificaciones establecidas por el equipo investigador. Dichas herramientas deberán permitir mantener la coherencia visual entre las distintas escenas, incluyendo, cuando resulte necesario, la apariencia, vestimenta y características de los personajes a lo largo de diferentes secuencias.

- Generación y animación de secuencias de vídeo. El servicio deberá disponer de capacidad de computación y de modelos de inteligencia artificial que permitan generar secuencias de vídeo a partir de descripciones textuales, imágenes de referencia u otros elementos de entrada. La capacidad proporcionada deberá ser suficiente para generar movimientos naturales de los personajes y preservar, en la medida de lo posible, la continuidad temporal, espacial y visual entre las distintas secuencias.
- Generación y tratamiento de voz y diálogos. El servicio deberá permitir la utilización de modelos de síntesis y generación de voz para producir las pistas de audio asociadas a los diálogos definidos por el equipo investigador. Las herramientas disponibles deberán permitir controlar aspectos tales como vocalización, ritmo, entonación y características de la voz y, cuando resulte necesario, posibilitar su personalización a partir de muestras de audio de referencia.
- Sincronización entre voz e imagen. El servicio deberá proporcionar herramientas y capacidad de procesamiento para realizar la sincronización automática entre las pistas de audio y los movimientos de los labios y del rostro de los personajes presentes en los vídeos generados, permitiendo obtener secuencias audiovisuales suficientemente naturales y adecuadas para su posterior utilización en el contexto del ensayo clínico.
- Procesamiento, edición e integración audiovisual. El servicio deberá permitir ejecutar operaciones de edición y posprocesamiento sobre los diferentes elementos audiovisuales generados, incluyendo selección y montaje de secuencias, ajuste de duración, incorporación de transiciones, tratamiento de imagen y sonido, sincronización de pistas, ralentización o aceleración de segmentos, recorte, concatenación y otras operaciones necesarias para obtener una composición audiovisual homogénea.
- Procesamiento específico de contenidos con finalidad terapéutica. El servicio deberá disponer de herramientas que permitan realizar operaciones adicionales sobre las secuencias generadas, tales como transcripción automática del habla, generación de subtítulos sincronizados a nivel de palabra, detección y seguimiento facial, aplicación de zoom sobre regiones de interés y ralentización de fragmentos concretos, cuando dichas operaciones sean requeridas por el equipo investigador para la elaboración del material experimental.
- Renderizado y generación del material final. El servicio deberá proporcionar capacidad de procesamiento suficiente para realizar la codificación, normalización y renderizado final de los contenidos audiovisuales, permitiendo obtener vídeos en formatos digitales estándar, con resolución, tasa de fotogramas, codificación de vídeo y parámetros de audio adecuados para su utilización posterior.
- Regeneración, modificación y ajuste de contenidos. La capacidad contratada deberá permitir repetir los procesos de generación y procesamiento cuando los resultados obtenidos no satisfagan los requisitos establecidos por el equipo

investigador. En particular, deberá ser posible modificar parámetros, sustituir elementos, regenerar secuencias o volver a ejecutar cualquiera de las fases del flujo de trabajo sin que ello requiera el desarrollo de una nueva solución tecnológica.

El objeto del servicio no consiste, por tanto, en la realización individualizada por parte del adjudicatario de cada una de las actividades anteriormente descritas, sino en proporcionar durante la vigencia del contrato el acceso y supervisión del equipo investigador (ingeniero contratado por IdiPAZ a cargo del proyecto PI25/01147, para el uso de la infraestructura de computación, herramientas software, modelos de inteligencia artificial, licencias, suscripciones y capacidad de procesamiento necesarias para que dichas actividades puedan ser ejecutadas de forma recurrente en el marco del proyecto de investigación. El dimensionamiento del servicio deberá permitir la utilización combinada de las distintas herramientas y la repetición de los procesos de generación, modificación, integración y renderizado que resulten necesarios durante el desarrollo del ensayo clínico.

La realización del servicio requerirá la utilización combinada de diferentes modelos y herramientas de inteligencia artificial generativa y procesamiento audiovisual, dado que ninguna herramienta individual permite abordar con la calidad y grado de personalización requeridos todas las etapas de generación de los videoclips. Como referencia tecnológica, y de acuerdo con los desarrollos previamente realizados y validados en el marco de esta línea de investigación, el flujo de generación deberá incorporar, como mínimo, las siguientes herramientas o soluciones técnicamente equivalentes o superiores:

- Wan2.1, preferentemente en sus variantes de alta capacidad T2V-14B e I2V-14B, para la generación de las secuencias de vídeo a partir de descripciones textuales y/o imágenes de referencia. La variante T2V permitirá generar escenas a partir de la descripción del escenario, personajes, acciones y contexto establecidos para cada paciente, mientras que la variante I2V permitirá generar nuevas secuencias manteniendo la apariencia visual de personajes y escenarios a partir de imágenes de referencia. El modelo Wan2.1-14B permite alcanzar resoluciones de hasta 720p y requiere aproximadamente 48 GB de VRAM para su ejecución, lo que justifica la disponibilidad de infraestructura de computación de altas prestaciones.
- F5-TTS para la síntesis y personalización de las voces de los personajes. Para el personaje correspondiente al paciente, el sistema deberá permitir la clonación del timbre, ritmo, prosodia y entonación de su voz a partir de muestras de audio de referencia, de forma que los diálogos generados conserven, en la medida de lo posible, sus características vocales. El sistema desarrollado previamente permite utilizar muestras breves, del orden de 10-15 segundos, seleccionando además diferentes grabaciones de referencia en función de la entonación requerida para cada escena.
- LatentSync para la sincronización de los movimientos de los labios y del rostro con las pistas de audio generadas. Esta herramienta deberá emplearse para modificar las secuencias generadas de manera que exista una correspondencia precisa entre el contenido fonético del diálogo y el movimiento labial del personaje. En los trabajos previos se seleccionó LatentSync frente a Wav2Lip por

presentar mayor precisión de sincronización y una mejor consistencia temporal y visual de los movimientos faciales.

- FaceFusion para la personalización facial cuando el contenido terapéutico requiera incorporar la apariencia del paciente en el personaje protagonista. La herramienta deberá permitir realizar de forma automatizada la sustitución del rostro en las secuencias generadas, manteniendo la coherencia facial, las expresiones y el movimiento del personaje original. Como referencia, el desarrollo previo empleó FaceFusion con el modelo inswapper_128_fp16, configurado para su ejecución automatizada en modo headless. FaceFusion fue seleccionado frente a alternativas como DeepFaceLab, SimSwap, FaceSwap o HeyGen por su equilibrio entre calidad, automatización, posibilidad de ejecución local y protección de los datos.
- WhisperX para la transcripción del contenido oral y la generación de subtítulos sincronizados a nivel de palabra. Los subtítulos deberán permitir identificar temporalmente cada palabra pronunciada y, cuando resulte necesario para la intervención terapéutica, diferenciar mediante formato o código cromático los distintos interlocutores.
- MediaPipe para la detección y seguimiento automático del rostro del personaje. Esta información se utilizará para aplicar transformaciones sobre regiones de interés del vídeo, incluyendo particularmente el zoom automático sobre el rostro durante la pronunciación de palabras o expresiones objetivo y su combinación con efectos de ralentización destinados a facilitar la observación de los movimientos articulatorios.
- FFmpeg y MoviePy para las etapas de preprocesamiento, edición y posproducción audiovisual. Estas herramientas deberán utilizarse para normalizar formatos, resolución, tasa de fotogramas y parámetros de audio, cortar y concatenar secuencias, aplicar ralentizaciones u otros efectos, sincronizar pistas de audio y producir el videoclip final. En los desarrollos previos, FFmpeg se ha utilizado además para estandarizar los vídeos de entrada a parámetros compatibles con los módulos posteriores de procesamiento.

El proveedor deberá integrar estas herramientas en un flujo automatizado y reproducible, de manera que las diferentes etapas —generación del vídeo, generación o clonación de voz, sincronización labial, personalización facial, transcripción y subtitulado, procesamiento específico de las secuencias y montaje final— puedan ejecutarse de forma coordinada a partir de las especificaciones definidas para cada videoclip. El sistema desarrollado previamente demuestra la viabilidad de esta arquitectura modular y de la ejecución secuencial de estos procesos.

Se permitirá la utilización de versiones posteriores de las herramientas indicadas o de herramientas técnicamente equivalentes o superiores, siempre que el proveedor justifique en su memoria técnica que mantienen o mejoran las prestaciones requeridas en cuanto a calidad visual, naturalidad de la voz, coherencia temporal, sincronización labial, personalización, resolución y protección de los datos.

Asimismo, el proveedor deberá disponer de una infraestructura de computación suficiente para ejecutar los procesos de generación mediante inteligencia artificial, edición, integración y renderizado sin pérdida significativa de calidad y dentro de unos tiempos de procesamiento adecuados (no más de 30 minutos por videoclip).

Como referencia de la capacidad de cálculo necesaria para la ejecución local de los modelos de inteligencia artificial previstos, la infraestructura deberá disponer de recursos equivalentes, al menos, a los siguientes:

- 2 GPU con un mínimo de 48 GB de VRAM cada una, o una GPU con aproximadamente 94 GB de VRAM.
- Capacidad de procesamiento CPU equivalente a dos procesadores de al menos 16 núcleos y frecuencia próxima a 3 GHz.
- Al menos 100 GB de memoria RAM.
- Al menos 1 TB de almacenamiento dedicado disponible para los procesos de generación y tratamiento audiovisual.

Podrán admitirse arquitecturas o soluciones tecnológicas diferentes, incluidas infraestructuras de computación en la nube, siempre que el proveedor acredite que ofrecen prestaciones equivalentes o superiores y permiten satisfacer los requisitos de generación y calidad establecidos en el presente pliego.

Finalmente, el proveedor deberá disponer de los mecanismos técnicos y organizativos necesarios para garantizar la seguridad, integridad, confidencialidad y adecuada gestión de la información utilizada durante la prestación del servicio, evitando el uso o conservación de los datos y materiales proporcionados para finalidades distintas de las expresamente previstas en el contrato.

Madrid, a 13 de octubre de 2026.

POR EL ÓRGANO DE CONTRATACIÓN,

D. Francisco García Río

Presidente de la Comisión Delegada de la Fundación

CONFORME:
EL ADJUDICATARIO
FECHA Y FIRMA