Análisis Técnico Integral del Ecosistema Qwen3-TTS: Arquitectura, Rendimiento e Implementación Local

La evolución de la síntesis de voz mediante inteligencia artificial ha experimentado una transformación paradigmática en la última década, transitando desde sistemas concatenativos rudimentarios y modelos paramétricos estadísticos hacia arquitecturas de aprendizaje profundo de extremo a extremo. En este contexto, el lanzamiento de la serie Qwen3-TTS por parte del equipo de Alibaba Cloud en enero de 2026 representa un hito fundamental en la convergencia de la modelización del lenguaje natural y la generación acústica de alta fidelidad.[1] Este sistema no solo se posiciona como una herramienta de generación de habla, sino como un ecosistema robusto de modelos de pesos abiertos (open-weights) que integran capacidades de clonación instantánea, diseño de voces mediante lenguaje natural y un control granular de la prosodia que anteriormente era exclusivo de plataformas comerciales cerradas.[2]

Evolución y Paradigmas de la Síntesis de Voz: El Contexto de Qwen3-TTS

Para comprender la relevancia técnica de Qwen3-TTS, es imperativo analizar el estado previo de la tecnología Text-to-Speech (TTS). Tradicionalmente, los sistemas han lidiado con un compromiso crítico entre la personalización y la naturalidad. Los modelos basados en arquitecturas modulares, que separan el procesamiento de texto, la predicción de la prosodia y la vocoderización, a menudo sufren de errores en cascada y una falta de coherencia emocional.[4] Qwen3-TTS rompe esta tendencia al adoptar un enfoque de modelado de lenguaje discreto que unifica la representación semántica y acústica en un solo proceso autorregresivo.[1]

Este sistema surge tras el éxito de modelos multimodales como Qwen2.5-Omni, compartiendo ADN arquitectónico que le permite heredar capacidades de razonamiento para interpretar instrucciones complejas sobre cómo debe sonar una voz.[6] La serie Qwen3-TTS se distingue por su enfoque en la baja latencia y el control del usuario, ofreciendo una alternativa viable a servicios como ElevenLabs y MiniMax, pero con la ventaja de la soberanía de datos que permite la ejecución local.[2]

Arquitectura de Extremo a Extremo: Modelado Discreto y Multi-Codebook

El pilar fundamental que sustenta el rendimiento de Qwen3-TTS es su arquitectura de modelo de lenguaje (LM) basada en tokens discretos. En lugar de generar espectrogramas de mel continuos que luego deben ser procesados por un vocoder, el sistema utiliza un tokenizer neural que comprime el audio en códigos discretos, permitiendo que el modelo de lenguaje trate el sonido con la misma lógica estructural que el texto.[1]

La Innovación del Tokenizer Qwen3-TTS-12Hz

El Qwen3-TTS-Tokenizer-12Hz constituye el núcleo de la representación acústica del sistema. Este componente logra una compresión extrema de la señal de audio manteniendo la información paralingüística esencial, como la entonación, el ritmo y el estado emocional del hablante.[1] A diferencia de los tokenizers estándar que operan a frecuencias más altas para capturar cada detalle microscópico, la frecuencia de $12.5\,\mathrm{Hz}$ de Qwen3-TTS está optimizada para la eficiencia en la transmisión sin sacrificar la inteligibilidad.[11]

Este tokenizer emplea un esquema de 16 capas de libros de códigos múltiples (multi-codebook) que permite una reconstrucción de alta fidelidad mediante una red convolucional causal ligera.[11] La naturaleza causal de esta arquitectura es lo que permite que el sistema funcione en tiempo real, ya que no requiere el procesamiento de bloques de audio futuros para generar el segmento actual.[11]

Métrica de Calidad del Tokenizer	Valor Qwen3-TTS	Promedio de la Competencia
PESQ (Banda Ancha)	3.21	2.85
PESQ (Banda Estrecha)	3.68	3.42
STOI (Inteligibilidad)	0.96	0.93
UTMOS (Naturalidad)	4.16	3.89
Similitud del Hablante	0.95	0.87

Los resultados del análisis en el conjunto de prueba LibriSpeech test-clean demuestran que el tokenizer de Qwen3-TTS logra una preservación de la identidad del hablante casi perfecta (0.95), superando significativamente las soluciones de código abierto previas.[2]

Arquitectura de Doble Vía (Dual-Track) y Generación Híbrida

Una de las innovaciones más significativas para aplicaciones industriales es la arquitectura de doble vía, que permite que un único modelo maneje tanto la generación por transmisión (streaming) como la generación completa.[1] Esto es fundamental para los agentes de IA conversacionales donde la latencia es el factor determinante de la experiencia del usuario.[1]

La capacidad de emitir el primer paquete de audio tras la entrada de un solo carácter permite alcanzar una latencia de extremo a extremo de aproximadamente $97\,\mathrm{ms}$.[1] Esta métrica posiciona a Qwen3-TTS como una solución ideal para traducción en tiempo real, doblaje en vivo y asistentes inteligentes que deben responder de manera humana e instantánea.[2]

El Ecosistema de Modelos: Clasificación y Especialización Funcional

La familia Qwen3-TTS se estructura en dos escalas de parámetros, 1.7B y 0.6B, cada una diseñada para equilibrar el rendimiento y el consumo de recursos de manera diferenciada.[1]

La Serie 1.7B: Rendimiento de Referencia (Flagship)

El modelo de 1.7 mil millones de parámetros es la versión insignia, diseñada para capturar la máxima complejidad de la voz humana. Este modelo no solo ofrece la mayor fidelidad acústica, sino que posee la capacidad de razonamiento necesaria para interpretar instrucciones de diseño de voz sutiles y mantener la estabilidad en narraciones de larga duración.[1]

- **Qwen3-TTS-12Hz-1.7B-VoiceDesign:** Permite la creación de voces totalmente nuevas basadas únicamente en descripciones textuales. El modelo utiliza sus pesos pre-entrenados para interpolar características acústicas y generar una voz que cumpla con los requisitos de edad, género y personalidad solicitados.[1]
- **Qwen3-TTS-12Hz-1.7B-CustomVoice:** Especializado en el control de estilo sobre un conjunto de voces predefinidas de alta calidad. Es ideal para aplicaciones que requieren una identidad vocal consistente pero con una amplia gama emocional.[1]
- **Qwen3-TTS-12Hz-1.7B-Base:** Sirve como el cimiento para la clonación de voz de disparo cero y como punto de partida para que los desarrolladores realicen un ajuste fino (fine-tuning) para dominios específicos.[1]

La Serie 0.6B: Eficiencia y Despliegue en el Borde (Edge)

Para escenarios donde la memoria VRAM es limitada o se requiere una alta concurrencia de usuarios, la serie de 600 millones de parámetros ofrece una alternativa eficiente. Aunque tiene menos parámetros, mantiene la compatibilidad con el tokenizer de $12\,\mathrm{Hz}$ y soporta las funciones esenciales de clonación y generación multilingüe.[1]

Atributo de Selección	Modelo 1.7B	Modelo 0.6B
Tamaño de Almacenamiento	$3.9 -$	$1.8 -$
VRAM Recomendada	$6 -$	$2 -$
Fortaleza Principal	Calidad máxima y control robusto	Velocidad y eficiencia de recursos
Casos de Uso	Producción, audiobooks, cine	Chatbots móviles, dispositivos IoT

El análisis de rendimiento sugiere que mientras el modelo 1.7B destaca en capturar la prosodia sutil, el modelo 0.6B es capaz de operar en hardware de nivel de entrada, como tarjetas NVIDIA con $4\,\mathrm{GB}$ de VRAM, lo que democratiza el acceso a la clonación de voz de alta fidelidad.[3]

Análisis de Capacidades de Voz: Clonación, Diseño y Control

El ecosistema Qwen3-TTS redefine la interacción usuario-voz a través de tres pilares operativos: la clonación rápida, el diseño creativo y el control de estilo instruccional.[1]

Clonación de Voz de 3 Segundos (Zero-Shot Voice Cloning)

La capacidad de clonar una voz con solo 3 segundos de audio de referencia representa el estado del arte en aprendizaje zero-shot.[1] El mecanismo subyacente implica la extracción de un "speaker embedding" o x-vector del audio de referencia, que luego se inyecta en el modelo de lenguaje para condicionar la generación.[6]

A diferencia de los sistemas de clonación tradicionales que requieren horas de datos y procesos de entrenamiento costosos, Qwen3-TTS realiza esta adaptación durante la inferencia.[6] Una implicación crítica de esta tecnología es la preservación de las características del hablante a través de idiomas (cross-lingual cloning). Un usuario puede proporcionar una muestra en español y el modelo generará habla en japonés o alemán manteniendo el timbre y las peculiaridades vocales originales del usuario.[1]

Diseño de Voz mediante Descripciones (Voice Design)

El modelo VoiceDesign introduce una capacidad creativa sin precedentes. Los usuarios pueden crear voces "copyright-free" describiendo personajes ficticios.[5] Por ejemplo, una instrucción como "Utiliza una voz masculina de unos 70 años, un científico estratégico con un tono de autoridad y gravedad" permite al modelo sintetizar una voz coherente con ese arquetipo.[5]

Este enfoque es particularmente valioso para la industria del videojuego y la producción de audiodramas, donde se requieren múltiples personajes con identidades vocales distintas sin necesidad de contratar a decenas de actores de doblaje.[5] El modelo no se limita a cambiar el tono (pitch); altera la cadencia, la textura de la voz (rasposa, suave, brillante) y la forma en que se estructuran las pausas basándose en la personalidad descrita.[4]

Control Instruccional de Emociones y Prosodia

Qwen3-TTS soporta el control de la voz mediante instrucciones naturales que se procesan junto con el texto a sintetizar.[1] Esta capacidad permite ajustar la entrega dinámica de una oración sin cambiar el modelo.[6]

- **Emoción:** Instrucciones como "Habla con gran entusiasmo" o "Tono triste y melancólico" modifican los tokens acústicos para reflejar el estado afectivo.[2]
- **Velocidad y Ritmo:** Se puede ordenar al modelo que hable más despacio para explicar conceptos complejos o que acelere para diálogos casuales.[15]
- **Prosodia Adaptativa:** El sistema tiene una comprensión intrínseca de la puntuación y el contexto, ajustando automáticamente las pausas para sonar más humano.[1]

Evaluación de Rendimiento y Comparativas Técnicas

El rendimiento de Qwen3-TTS ha sido validado mediante benchmarks objetivos y subjetivos, comparándolo con los líderes actuales del mercado tanto en el ámbito comercial como en el de código abierto.[2]

Comparativa de Word Error Rate (WER) Multilingüe

En la evaluación de inteligibilidad, Qwen3-TTS demuestra una superioridad notable en la mayoría de los 10 idiomas soportados. En el benchmark Seed-TTS-Eval, el modelo 1.7B logra el WER más bajo, superando a MiniMax y ElevenLabs.[2]

Modelo	WER Promedio (10 idiomas)	Similitud del Hablante (Avg)
Qwen3-TTS-1.7B	1.835%	0.789

MiniMax	2.14%	0.74
ElevenLabs	2.31%	0.72
OpenAI GPT-4o Audio	2.45%	0.71

El análisis de los datos revela que Qwen3-TTS es especialmente fuerte en chino, logrando un WER de 0.777.[23] En idiomas como el italiano y el francés, el rendimiento se califica como el mejor de su clase, mientras que en alemán y español es altamente competitivo, con ElevenLabs logrando ocasionalmente WERs ligeramente inferiores en casos muy específicos de pronunciación de nombres propios.[15]

Estabilidad y Generación de Larga Duración

Una deficiencia común en los modelos TTS autorregresivos es la acumulación de errores en secuencias largas, lo que puede llevar a que la voz pierda coherencia o empiece a balbucear después de unos minutos.[11] Qwen3-TTS soluciona esto mediante un entrenamiento específico en contextos largos, ampliando la ventana de tokens de 8,192 a 32,768 durante la fase final de pre-entrenamiento.[11]

En pruebas de síntesis continua de 10 minutos, Qwen3-TTS mantiene un WER estable de 2.36% en chino y 2.81% en inglés, lo que lo hace apto para la narración de capítulos completos de libros o podcasts extensos sin intervención humana para corregir la deriva acústica.[2]

Diversidad Lingüística y Biblioteca de Voces

El sistema ofrece una cobertura lingüística global que incluye chino, inglés, japonés, coreano, alemán, francés, ruso, portugués, español e italiano.[1] Esta diversidad se complementa con un soporte sin precedentes para dialectos regionales, especialmente en el contexto del idioma chino.[22]

Soporte Dialectal y Timbres Premium

Qwen3-TTS incluye 49 timbres de alta calidad listos para usar, que abarcan una amplia gama de perfiles de edad y género.[22] El modelo CustomVoice permite seleccionar entre 9 voces premium diseñadas para sonar lo más naturales posible en sus idiomas nativos.[1]

Nombre del Perfil	Características Vocales	Idioma/Dialecto
Vivian	Joven femenina, brillante y orgullosa.	Chino / Inglés [21]
Ryan	Masculina dinámica, ritmo marcado.	Inglés [22]
Aiden	Masculina joven, acento estadounidense soleado.	Inglés [22]
Ono Anna	Femenina japonesa, juguetona y cálida.	Japonés [22]
Sohee	Femenina coreana, cálida y amable.	Coreano [22]
Dylan	Masculina joven, acento de Beijing.	Chino (Beijing) [22]
Eric	Masculina animada, acento de Chengdu.	Chino (Sichuan) [22]
Marcus	Masculina profunda, acento de Shaanxi.	Chino (Shaanxi) [22]

Esta especialización dialectal permite que las aplicaciones de servicio al cliente y los guías turísticos digitales interactúen con los usuarios de una manera que respeta las normas culturales y lingüísticas locales, aumentando la aceptación y la confianza del usuario final.[4]

Implementación Local y Requisitos de Hardware

La capacidad de ejecutar Qwen3-TTS localmente es uno de sus mayores atractivos para empresas con requisitos de privacidad estrictos o desarrolladores que desean evitar costos recurrentes de API.[1]

Requisitos Técnicos y Optimización

Para un despliegue exitoso, se requiere una infraestructura que soporte el procesamiento paralelo masivo necesario para la arquitectura de transformadores.[3]

- **Hardware NVIDIA:** Se recomienda encarecidamente el uso de GPUs con núcleos Tensor. La serie RTX 3000 y 4000 es la más adecuada gracias al soporte nativo de bfloat16.[16]
- **VRAM por Escala:**
  - **0.6B:** Puede ejecutarse en GPUs con $2-4\,\text{GB}$ de VRAM si se utiliza cuantización, aunque se recomiendan $6\,\text{GB}$ para estabilidad.[3]
  - **1.7B:** Requiere entre $6\,\text{GB}$ y $12\,\text{GB}$ de VRAM. Una RTX 3060 ($12\,\text{GB}$) es la configuración ideal para obtener una calidad óptima con latencia aceptable.[15]
- **Optimización de Software:** El uso de FlashAttention 2 es crítico para reducir el uso de memoria en secuencias largas y mejorar la velocidad de inferencia en un 30%.[2]

Procedimiento de Instalación (Linux y Windows WSL2)

La instalación se realiza mediante el gestor de paquetes de Python y el acceso a los repositorios oficiales en GitHub o Hugging Face.[3]

1. **Entorno Virtual:** Se recomienda crear un entorno aislado con conda o venv para evitar conflictos de dependencias, especialmente con versiones específicas de transformers y torch.[3]
2. **Dependencias Críticas:**
   ```
   Bash
   pip install -U qwen-tts torch soundfile transformers
   pip install flash-attn --no-build-isolation
   ```
3. **Configuración de Audio:** Para sistemas Linux, puede ser necesario instalar bibliotecas de procesamiento de audio adicionales como ffmpeg o SoX para manejar diferentes formatos de entrada de referencia.[8]

Ecosistema de ComfyUI para Usuarios Visuales

Para usuarios que prefieren flujos de trabajo basados en nodos, la integración de Qwen3-TTS en ComfyUI ha sido muy exitosa. Existen nodos personalizados que permiten orquestar procesos complejos de clonación y generación.[28]

El repositorio ComfyUI-Qwen3-TTS de DarioFT es uno de los más completos, permitiendo:

- **Dataset Maker:** Automatizar la creación de conjuntos de datos para fine-tuning a partir de una carpeta de audios y transcripciones.[29]
- **Save/Load Prompt:** Guardar las "huellas vocales" (prompts acústicos) en archivos .safetensors para reutilizar voces clonadas sin necesidad de procesar el audio de referencia cada vez.[29]
- **Finetune Node:** Realizar un ajuste fino del modelo 1.7B directamente en la interfaz, con soporte para optimizadores de 8 bits que permiten entrenar en GPUs de consumo de

$12 - 16\,\mathrm{GB}$.[29]

Marco de Seguridad y Ética: Qwen3Guard

Dada la potencia de la clonación de voz para crear "deepfakes" de audio convincentes, el equipo de Qwen ha lanzado un sistema de salvaguardas denominado Qwen3Guard.[31] Este marco está diseñado para garantizar que el modelo se utilice de manera responsable y ética.[31]

Detección y Moderación en Tiempo Real

Qwen3Guard funciona en dos modalidades que pueden integrarse en las aplicaciones finales para mitigar riesgos.[31]

1. **Qwen3Guard-Gen:** Un modelo generativo para la clasificación de seguridad fuera de línea. Se utiliza para anotar grandes conjuntos de datos o filtrar prompts antes de la síntesis.[31]
2. **Qwen3Guard-Stream:** Una innovación técnica que adjunta "cabezales de clasificación" ligeros a la capa final del transformador. Esto permite evaluar la seguridad del audio token por token mientras se genera, permitiendo una intervención inmediata si el modelo empieza a producir contenido prohibido.[31]

Este sistema soporta 119 idiomas y dialectos, asegurando una cobertura global contra el discurso de odio, la suplantación de identidad no autorizada y la generación de contenido sensible.[31] Además, el uso de marcas de agua digitales (watermarking) discretas permite rastrear el origen de los audios sintéticos, proporcionando una capa de protección adicional para la integridad de la identidad vocal.[5]

Comparativa con Sistemas Alternativos: Qwen3-TTS vs. El Mercado

Para un analista de IA, es crucial entender dónde se sitúa Qwen3-TTS frente a competidores como F5-TTS, Fish Speech y VibeVoice.[2]

Qwen3-TTS frente a F5-TTS y Fish Speech

F5-TTS ha sido aclamado por su capacidad de clonación fiel basada en "flow matching". Sin embargo, el análisis comparativo indica que Qwen3-TTS ofrece una solución más integrada para el soporte multilingüe y el despliegue en tiempo real.[2] Mientras que F5-TTS destaca en la captura del timbre puro, Qwen3-TTS proporciona un control instruccional mucho más avanzado, permitiendo al usuario dictar la emoción y el ritmo mediante lenguaje natural, algo que en F5-TTS suele requerir una manipulación manual más compleja.[2]

Qwen3-TTS frente a ElevenLabs

ElevenLabs sigue siendo el líder en conveniencia comercial (SaaS), pero Qwen3-TTS lo iguala o supera en métricas de similitud del hablante y latencia de primer paquete.[2] Para empresas que procesan datos sensibles o que requieren miles de horas de síntesis al mes, la transición de ElevenLabs a una solución local basada en Qwen3-TTS representa un ahorro de costos masivo y una mejora en la seguridad de la información.[2]

Característica	Qwen3-TTS (Local)	ElevenLabs (API)	F5-TTS (OSS)
Latencia	$97\,\mathrm{ms}$ (Local)	$\approx 300 -$	$\approx 200 -$
Costo	Basado en hardware propio	Pago por uso (Créditos)	Gratuito
Privacidad	Total (Local)	Nube (Terceros)	Total (Local)
Idiomas	10 (Nativos)	29+	Principalmente EN/CN
Control	Instrucciones de texto	Deslizadores UI / API	Prompting limitado

Casos de Uso Industriales y Perspectivas Futuras

La versatilidad de Qwen3-TTS abre un abanico de aplicaciones prácticas que ya están siendo exploradas por desarrolladores y empresas globales.[1]

Educación y Localización de Contenidos

El doblaje automático de videos educativos es una aplicación inmediata. Un profesor puede grabar una lección en español y, utilizando Qwen3-TTS, generar versiones en otros 9 idiomas manteniendo su propia voz. Esto mejora significativamente el compromiso del estudiante al recibir información en su idioma nativo pero con una voz familiar.[2]

Entretenimiento y Narración

La producción de audiolibros con múltiples personajes se simplifica drásticamente. En lugar de requerir varios actores de voz, un solo productor puede "diseñar" voces para cada personaje del libro y generar diálogos coherentes con emociones dinámicas (sarcasmo,

alegría, miedo) basándose en las etiquetas del texto o instrucciones adicionales.[1]

Accesibilidad: Restauración de Voz

Para personas que han perdido el habla debido a condiciones médicas, Qwen3-TTS ofrece una herramienta de restauración de identidad. Solo se necesitan unos pocos segundos de una grabación antigua para crear un modelo de voz que les permita comunicarse en tiempo real con su timbre original, integrándose en dispositivos de asistencia móviles gracias a la eficiencia de la variante 0.6B.[3]

Conclusiones

El ecosistema Qwen3-TTS representa la culminación de años de investigación en modelos de lenguaje y procesamiento de señales de audio. Al unificar estas disciplinas en una arquitectura de extremo a extremo basada en tokens discretos, Alibaba Cloud ha resuelto los problemas de latencia y control que limitaban el uso industrial del TTS de alta fidelidad.[4]

La combinación de un tokenizer de $12\,\mathrm{Hz}$ altamente eficiente, una arquitectura de doble vía para transmisión instantánea y un marco de seguridad como Qwen3Guard posiciona a esta serie como el estándar de facto para la síntesis de voz de código abierto en 2026.[1] La flexibilidad de despliegue local, sumada a la calidad que compite directamente con los motores comerciales más potentes, asegura que Qwen3-TTS será el motor preferido para la próxima ola de agentes de IA conversacionales y sistemas multimodales inteligentes.[2]

Obras citadas

1. Alibaba Qwen Releases Qwen3-TTS - 97ms Ultra-Low Latency Voice Synthesis Model, fecha de acceso: febrero 5, 2026, https://comfyui-wiki.com/en/news/2026-01-22-alibaba-qwen3-tts-release
2. Qwen3-TTS: The Complete 2026 Guide to Open-Source Voice Cloning and AI Speech Generation | by cheng zhang - Medium, fecha de acceso: febrero 5, 2026, https://medium.com/@zh.milo/qwen3-tts-the-complete-2026-guide-to-open-source-voice-cloning-and-ai-speech-generation-1a2efca05cd6
3. Qwen3-TTS: Complete Guide to Open-Source Text-to-Speech Model - DEV Community, fecha de acceso: febrero 5, 2026, https://dev.to/gary_yan_86eb77d35e0070f5/qwen3-tts-complete-guide-to-open-source-text-to-speech-model-9oe
4. Qwen3-TTS: Multilingual, Real-Time Speech AI - Webkul Blog, fecha de acceso: febrero 5, 2026, https://webkul.com/blog/qwen3-tts/amp/
5. Qwen3-TTS models introduce two powerful capabilities—Voice Design and Voice Cloning, fecha de acceso: febrero 5, 2026, https://dataglobalhub.org/resource/articles/qwen-3-tts-models-introduce-two-powerful-capabilities-voice-design-and
6. Qwen3 TTS: From Zero to Voice Cloning | atal upadhyay - WordPress.com, fecha

de acceso: febrero 5, 2026, https://atalupadhyay.wordpress.com/2026/01/29/qwen3-tts-from-zero-to-voice-cloning/
7. Qwen2.5-Omni: A Real-Time Multimodal AI - LearnOpenCV, fecha de acceso: febrero 5, 2026, https://learnopencv.com/qwen2-5-omni/
8. Qwen2.5-Omni Incoming? Huggingface Transformers PR 36752 : r/LocalLLaMA - Reddit, fecha de acceso: febrero 5, 2026, https://www.reddit.com/r/LocalLLaMA/comments/1jhhsgv/qwen25omni_incoming_huggingface_transformers_pr/
9. Alibaba releases voice cloning models using 3 seconds of audio - Perplexity, fecha de acceso: febrero 5, 2026, https://www.perplexity.ai/page/alibaba-releases-qwen3-tts-voi-CUGC5iB4QCylaAmb3ySI0Q
10. Qwen3-TTS Family is Now Open Sourced: Voice Design, Clone, and Generation!, fecha de acceso: febrero 5, 2026, https://qwen.ai/blog?id=qwen3tts-0115
11. Qwen3-TTS Technical Report - arXiv, fecha de acceso: febrero 5, 2026, https://arxiv.org/html/2601.15621v1
12. [2601.15621] Qwen3-TTS Technical Report - arXiv, fecha de acceso: febrero 5, 2026, https://arxiv.org/abs/2601.15621
13. Daily Papers - Hugging Face, fecha de acceso: febrero 5, 2026, https://huggingface.co/papers?q=low-latency%20streaming
14. Alibaba Releases Qwen2.5-Omni-7B: New Multimodal LLM - Sigma AI Browser, fecha de acceso: febrero 5, 2026, https://www.sigmabrowser.com/blog/alibaba-releases-qwen2-5-omni-7b-new-multimodal-llm
15. Qwen3-TTS: The Open-Source Text-to-Speech Revolution in 2026 - DEV Community, fecha de acceso: febrero 5, 2026, https://dev.to/gary_yan_86eb77d35e0070f5/qwen3-tts-the-open-source-text-to-speech-revolution-in-2026-3466
16. Qwen3-TTS: The Complete 2026 Guide to Open-Source Voice Cloning and AI Speech Generation - DEV Community, fecha de acceso: febrero 5, 2026, https://dev.to/czmilo/qwen3-tts-the-complete-2026-guide-to-open-source-voice-cloning-and-ai-speech-generation-1in6
17. Qwen3-TTS: Design Custom Voices with Text Descriptions: Easy Local Demo, fecha de acceso: febrero 5, 2026, https://www.youtube.com/watch?v=gR5dyKaxpEk
18. Qwen/Qwen3-TTS-12Hz-0.6B-Base - Hugging Face, fecha de acceso: febrero 5, 2026, https://huggingface.co/Qwen/Qwen3-TTS-12Hz-0.6B-Base
19. How to Run Qwen3 Locally - A Practical Guide for AI Enthusiasts, fecha de acceso: febrero 5, 2026, https://onedollarvps.com/blogs/how-to-run-qwen3-locally
20. Pure Rust implementation of Qwen3-TTS speech synthesis - GitHub, fecha de acceso: febrero 5, 2026, https://github.com/TrevorS/qwen3-tts-rs
21. Qwen3-TTS-Flash Review: The Most Realistic Open TTS Model Yet? - Analytics Vidhya, fecha de acceso: febrero 5, 2026,

https://www.analyticsvidhya.com/blog/2025/12/qwen3-tts-flash-review/
22. Qwen3-TTS Update! 49 Timbres + 10 Languages + 9 Dialects, fecha de acceso: febrero 5, 2026, https://qwen.ai/blog?id=qwen3-tts-1128
23. [Quick Review] Qwen3-TTS Technical Report - Liner, fecha de acceso: febrero 5, 2026, https://liner.com/review/qwen3tts-technical-report
24. Alibaba Launches Powerful Text-to-Speech Model Qwen3-TTS, 49 Voices Meet Your Voice Needs! - AI NEWS, fecha de acceso: febrero 5, 2026, https://news.aibase.com/news/23579
25. Qwen/Qwen3-TTS-12Hz-0.6B-CustomVoice - Hugging Face, fecha de acceso: febrero 5, 2026, https://huggingface.co/Qwen/Qwen3-TTS-12Hz-0.6B-CustomVoice
26. GPU System Requirements Guide for Qwen LLM Models (All Variants), fecha de acceso: febrero 5, 2026, https://apxml.com/posts/gpu-system-requirements-qwen-models
27. Qwen3 TTS – install and test in ComfyUI – MyClone Poser and Daz Studio blog - JURN, fecha de acceso: febrero 5, 2026, https://jurn.link/dazposer/index.php/2026/01/24/qwen3-tts-install-and-test-in-comfyui/
28. Qwen3-TTS, a series of powerful speech generation capabilities : r/StableDiffusion - Reddit, fecha de acceso: febrero 5, 2026, https://www.reddit.com/r/StableDiffusion/comments/1qjuebr/qwen3tts_a_series_of_powerful_speech_generation/
29. DarioFT/ComfyUI-Qwen3-TTS: A ComfyUI custom node suite for Qwen3-TTS, supporting 1.7B and 0.6B models, Custom Voice, Voice Design, Voice Cloning and Fine-Tuning. - GitHub, fecha de acceso: febrero 5, 2026, https://github.com/DarioFT/ComfyUI-Qwen3-TTS
30. Full Voice Cloning in ComfyUI with Qwen3-TTS + ASR - Reddit, fecha de acceso: febrero 5, 2026, https://www.reddit.com/r/comfyui/comments/1qqfnro/full_voice_cloning_in_comfyui_with_qwen3tts_asr/
31. Qwen3Guard: Real-time Safety for Your Token Stream | Qwen, fecha de acceso: febrero 5, 2026, https://qwenlm.github.io/blog/qwen3guard/
32. Qwen3Guard, fecha de acceso: febrero 5, 2026, https://qwen.ai/blog?id=qwen3guard
33. Qwen3-TTS: Multilingual, Controllable TTS - Emergent Mind, fecha de acceso: febrero 5, 2026, https://www.emergentmind.com/papers/2601.15621
34. Compare Chatterbox vs. Qwen3-TTS in 2026 - Slashdot, fecha de acceso: febrero 5, 2026, https://slashdot.org/software/comparison/Chatterbox-Voice-Cloning-vs-Qwen3-TTS/
35. Qwen3-TTS 1.7B vs VibeVoice 7B : r/StableDiffusion - Reddit, fecha de acceso: febrero 5, 2026, https://www.reddit.com/r/StableDiffusion/comments/1qne14v/qwen3tts_17b_vs_vibevoice_7b/
36. What's the Highest Quality Open-Source TTS? : r/LocalLLaMA - Reddit, fecha de

acceso: febrero 5, 2026, https://www.reddit.com/r/LocalLLaMA/comments/1qqmmn0/whats_the_highest_quality_opensource_tts/