TECHNICAL REPORT • JULY 2026

Seed Audio 1.0

The Complete Model Report

A practical, evidence-aware guide to unified AI audio generation, prompting, editing, access, and audio-first video workflows.

Visit https://seedaud.io/

Contents

What Seed Audio 1.0 Actually Is **2**

What It Can Actually Do **3**

Which Mode to Use **6**

How It Works **7**

Technical Specs at a Glance **8**

- Inputs 9
- Outputs 9
- Generation Controls 9
- Extended Capabilities 9
- Access Routes 10

How to Write Prompts **10**

Working with Seedance **12**

Where You Can Access It **13**

What It Costs **14**

Is It Worth Using **14**

What Seed Audio 1.0 Actually Is

Let's start with an old problem.

AI video in 2026 is already seriously good. You can throw in a still image and, a few seconds later, get back a moving clip with camera motion, lighting changes, and character animation. Short-form creators use it. Ad teams use it. Even film previs teams are getting in on it.

But one problem has remained stubbornly unsolved: **sound**.

A video can look amazing, but if the background is dead silent, if a character's mouth moves with no dialogue, or if someone just slaps on a generic canned music bed and calls it a day, viewers swipe away in three seconds. If the sound is wrong, even gorgeous visuals still feel unfinished.

So creators have been forced into patchwork workflows: one tool for voice, another for sound effects, another for music, then everything dumped into a DAW for manual alignment, level balancing, and mixing. On a 30-second short, the audio post alone can still eat up one or two hours.

That is the workflow step Seed Audio 1.0 is trying to eliminate.

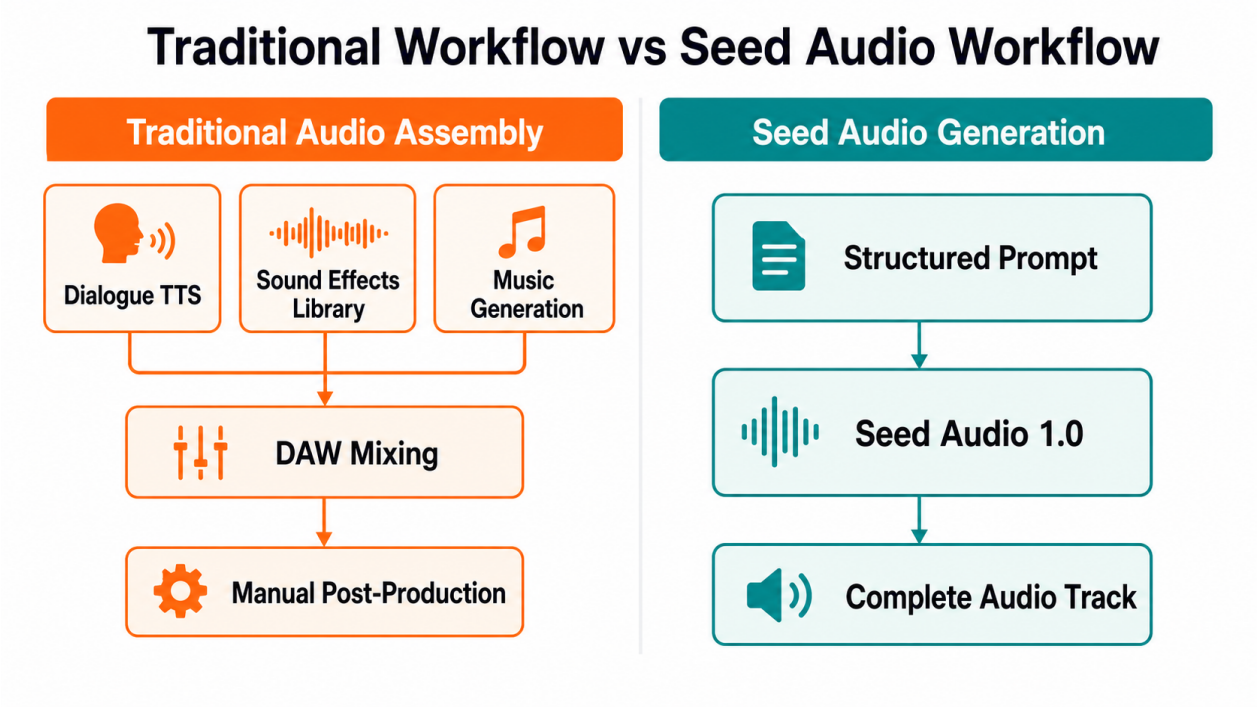

Figure 1: Traditional multitrack assembly versus Seed Audio unified generation

Here is the simplest way to say it.

Seed Audio 1.0 — also referred to in the Chinese source as **Doubao Audio Generation Model 1.0** — is an AI audio model released by ByteDance's Seed team. Its core capability is straightforward: **you describe a scene in text, and it generates a complete, cinematic-quality audio track that can include dialogue, emotion, ambience, sound effects, and background music, all on a single track.**

No multitrack assembly. No manual post-mix. One prompt in, one finished audio scene out.

This is why it is not the same thing as ordinary TTS.

Traditional text-to-speech systems — Siri-style systems, standard voiceover tools, and similar products — read text aloud. How human they sound depends on the quality of the model, but the basic job is still the same: they are readers.

Seed Audio 1.0 behaves more like an **audio director** than a reader. It is not just speaking lines. It is interpreting a whole scene: a rainy city street at night, two people talking, thunder in the background, distant traffic, low suspense strings under the dialogue, and one character whose voice breaks slightly from nerves. Put all of that into a prompt and the model can try to generate the whole thing at once.

Put bluntly: **TTS solves "turn text into speech." Seed Audio 1.0 tries to solve "turn a scene description into a complete sonic world."**

Its technical lineage also matters.

Seed Audio 1.0 did not appear out of nowhere. To understand it, the source article places it inside ByteDance's broader content-generation stack:

Seedream for images
$rightarrow$ **Seedance for video**
$rightarrow$ **Seed Audio for audio**
$rightarrow$ **Seed 2.1 for script handling**

That exact four-step framing is best read as an explanatory shorthand, not an official ByteDance pipeline label. Public ByteDance Seed materials do confirm that **Seedance 2.0** is a multimodal audio-video model and that the broader Seed organization publicly presents related models such as **Seed2.1** and **Seedream 5.0 Lite** on its official site.

More specifically, the source article argues that Seed Audio sits on top of two mature lines of work:

- **Seed-TTS**, ByteDance's 2024 speech-generation family, which publicly documented zero-shot speech continuation, emotion control, and timbre handling.
- **Seedance**, ByteDance's video-generation line, which publicly documented native joint audio-video generation by the 1.5 Pro and 2.0 generations.

That is a fair interpretation. Public sources do support the idea that ByteDance already had strong speech-generation work through Seed-TTS and audio-video joint generation through Seedance before this audio product surfaced publicly.

Is it usable now?

Yes — the quickest way to try it is through **seedaud.io**, a browser-based platform where you can test Seed Audio 1.0 directly without registering for an API or writing code.

Other access routes listed in the source article include **Volcano Engine Ark**, **BytePlus**, **Higgsfield**, and **fal.ai**. Not all of these are equally well documented in public English-facing materials, so they should be treated as source-level availability claims rather than fully confirmed public product documentation in this report.

One more limitation is worth stating clearly: the Chinese article itself acknowledges that ByteDance had not publicly released a standalone technical paper with parameter count or training-data scale for Seed Audio 1.0 at the time of writing. I did not identify a public standalone technical report in the sources I reviewed for this report either, so I treated unsupported architectural specifics cautiously.

What It Can Actually Do

Once you understand what Seed Audio 1.0 is, the next question is what it can actually do in practice. The source article breaks that down into six core abilities.

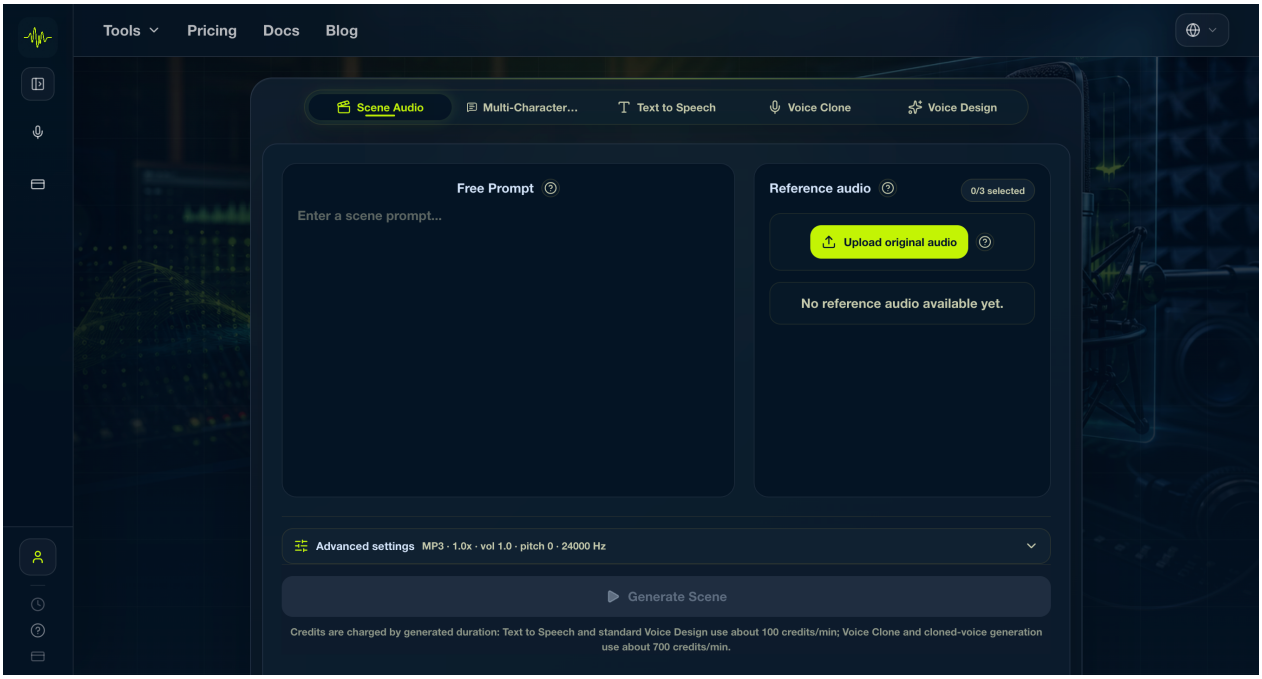

Figure 2: seedaud.io browser-based testing interface

Figure 3: Seed Audio 1.0 Core Capability Matrix

The first, and most important, is this: **one prompt can generate a complete cinematic audio scene**.

In a conventional workflow, a full audio scene gets built piece by piece. You record dialogue. You hunt down effects. You add music. Then you stack everything in a DAW, line it up on the timeline, and keep adjusting the levels until it sounds coherent.

Seed Audio 1.0 aims for something different. You describe the scene in one prompt — who is speaking, what they say, what emotion they are in, what is happening in the background, what kind of music should be playing, which sound cues matter — and the model attempts to generate everything together on a single track. The idea is that volume relationships, timing, and spatial layering are coordinated automatically rather than fixed by hand later. Public fal.ai materials describe this capability in similar terms, presenting the model as a system that can generate dialogue, sound effects, music, and atmosphere "in a single pass."

Take a prompt like this:

> In a convenience store late at night, a man and the clerk are talking. The man sounds tense and keeps his voice low; the clerk sounds distracted. There is a fluorescent hum in the background, along with the low drone of a refrigerator compressor. Every so often, a car passes in the distance. The music is dark, low suspense strings.

The point is that Seed Audio 1.0 is supposed to return a single finished piece of audio: dialogue in the right emotional register, environment built out underneath it, and background score that follows the mood. That cuts audio post time from hours to minutes for some creator workflows.

The second major ability is **multi-character dialogue without voice drift**.

A lot of AI dubbing tools can generate multiple characters. The common failure mode is consistency: Character A starts as a low male voice and gradually mutates into something else. That is easy to miss in a 15-second clip, but it becomes glaring in audiobooks, audio drama, and serialized content.

The source article says Seed Audio 1.0 is specifically designed to address that problem. It says the model can define multiple characters in one prompt, set separate voice traits and emotions for each, and maintain them over long durations. Public fal materials do support the multi-speaker and reference-voice workflow, including the use of `@Audio1`, `@Audio2`, and `@Audio3` for character continuity.

The source article goes further and claims that, in audio longer than 30 minutes, multi-character timbre and emotional consistency stays within **less than 5% error**.

The third major ability is **zero-shot voice cloning**.

That means you do not need to train a dedicated speaker model on a large custom dataset. Public fal.ai documentation says the model can take **up to three reference clips**, each **up to 30 seconds**, and reuse them as prompt-referenced voices. That lines up well with the source article's description. The same public documentation also supports the use of a **single image** as a reference input, though on the public fal endpoint the image-input path cannot be combined with reference-audio input in the same request.

The source article describes three ways to guide generation:

- text description of the desired voice,
- a reference audio sample,
- or image guidance.

That is directionally consistent with publicly visible fal documentation, which supports text, audio

references, and a single reference image.

The fourth ability is easy to overlook but important: Seed Audio 1.0 is presented not just as a generator, but also as an **audio editor**.

Public fal materials say the model supports:

- **audio extending**,
- **inpainting**,
- **stitching**,
- and spoken-line editing.

The Chinese article describes the same general family of tasks in more consumer-facing language: continuing a clip, repairing a section, merging separate clips, swapping out a word or sentence, or generating an alternate ending while keeping the original voice and environment intact. That is one of the strongest overlaps between the source article and publicly accessible product documentation.

The fifth ability is **fine-grained control over sound details**.

Public fal docs verify controls for **speed**, **volume**, **pitch**, **output format**, and **sample rate**. Specifically, the public schema lists output formats including **wav, mp3, pcm, and ogg_opus**, sample rates from **8,000 Hz to 48,000 Hz**, and pitch control measured in semitones. That makes the source article's general description of controllable speaking speed, loudness, pitch, and output format well supported.

The sixth ability, according to the source article, is **3D spatial audio / soundstage simulation**.

This is where the article becomes more speculative. It says the model includes a physical acoustics simulation module that can infer directionality, reverberation, and masking effects from scene descriptions. It also says the model can automate foley from text cues like "door opens," "pouring water," or "footsteps."

The source article also says the model includes **20+ built-in dialect models**, including Cantonese and Southern Min, and can respond to hybrid instructions such as "heavy Sichuan accent + hushed, choking-back-tears delivery."

Which Mode to Use

The source article says Seed Audio 1.0 has two primary operating modes, plus one combined workflow. Public fal documentation supports the same distinction, using the terms **TTS**, **T2A**, and **TA2A**.

TTS mode is the simplest. You give it text, and it reads the text aloud. No environmental effects. No added score. Just clean spoken output.

That makes it suitable for narration, audiobook reading, product explainers, and podcast-style voice work — any case where you want a strong voice without a full scene mix. In this mode, you can either describe the voice you want in text or supply a reference voice clip.

T2A, or **text-to-audio**, is the flagship mode.

This is the workflow where you type text and get back a full audio scene: dialogue, effects, music, ambience, everything together. Public fal documentation explicitly describes T2A as full-scene generation from text only, with no reference voice. That matches the source article's explanation closely.

TA2A, or **text-and-audio-to-audio**, is the more advanced version.

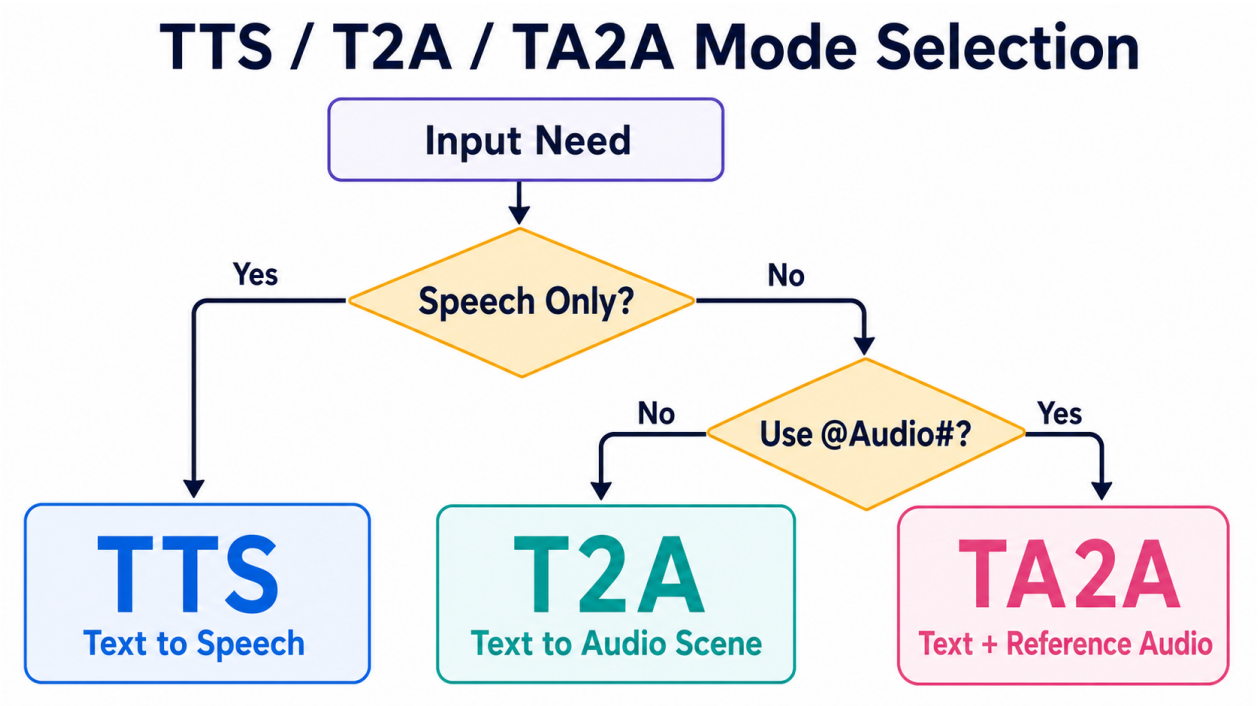

Figure 4: TTS, T2A, TA2A Mode Selection

Here you do not just describe the scene; you also tell the model to use specific previously saved or uploaded voices. Public fal docs confirm that reference voices can be tagged in the prompt with `@Audio1`, `@Audio2`, and `@Audio3`, which aligns neatly with the article's explanation.

The cleanest rule of thumb is:

Your need	Best mode
You only need spoken text with no background sound	**TTS**
You need a complete audio scene with dialogue, effects, and music	**T2A**
You need a complete scene and want to anchor it to specific saved voices	**TA2A**

A very simple cue is this: if your prompt includes tags like `@Audio1`, you are using the reference-audio workflow rather than plain text-only T2A.

How It Works

The Chinese source includes a caution that still holds: as of the source article's writing, there was no fully public standalone technical paper for Seed Audio 1.0 laying out parameter count, training-data scale, or a dedicated benchmark suite. That means any deep architectural explanation should be separated into what is publicly documented, what is plausible lineage-based inference, and what is just explanatory language.

The article's core argument is that conventional tools treat speech, sound effects, ambience, and music as separate layers, often generated by separate models and mixed later. Seed Audio 1.0,

by contrast, is described as a unified generation system that maps these different sound types into one shared generation space and outputs a finished mixed track directly.

That framing fits well with the public fal product story, which repeatedly emphasizes dialogue, sound design, and music "in one pass," and with the broader trend visible in ByteDance's Seedance 1.5 Pro and 2.0 papers, both of which describe native audio-video joint generation rather than bolt-on synchronization.

The source article then introduces several architecture labels:

- **hierarchical memory network** for long-horizon consistency,
- **cross-modal contrastive learning** for zero-shot cloning,
- **emotion-acoustics mapping matrix** for emotion control,
- **physical acoustics simulation module** for spatial realism.

But the source also explicitly warns that these are not official ByteDance paper terms.

Where public primary evidence is available, it comes mainly from **Seed-TTS**.

The Seed-TTS paper publicly documents a four-stage speech-generation pipeline:

1. a speech tokenizer,
2. an autoregressive token language model,
3. a diffusion transformer,
4. and an acoustic vocoder.

It also documents zero-shot speech continuation, controllability over emotion and other speech attributes, and speech-editing capability in the diffusion-based variant. That publicly documented lineage makes the Chinese article's technical ancestry argument credible even when Seed Audio's own public architecture remains thinly documented.

The article also links Seed Audio conceptually to **Seedance**.

Public primary sources support that connection. The Seedance 1.5 Pro technical report describes a model built specifically for native joint audio-video generation and highlights multilingual and dialect lip-syncing. The Seedance 2.0 model card goes further, describing a unified multimodal audio-video architecture that supports **text, image, audio, and video inputs**. That does not prove the internal architecture of Seed Audio 1.0, but it does support the article's larger point that ByteDance was already working on joint audio-visual generation before this model appeared in public product form.

The article closes this section by being unusually candid about what remains unknown:

- parameter count is not public,
- training data is not public,
- a dedicated public benchmark table is not public,
- and a standalone technical report is not public.

That remains the right way to frame it. We can say a fair amount about what the product appears able to do, and a fair amount about its likely lineage, but not nearly as much as we could say if ByteDance had published a full Seed Audio paper.

Technical Specs at a Glance

Below is the translated and lightly standardized version of the source article's spec summary. Where a row was publicly verifiable, I standardized it to match accessible documentation. Where

it was not, I left the source meaning intact and note the uncertainty afterward.

Inputs

Item	Translated / standardized spec
Text input	Natural-language prompt; public docs support English and Chinese output workflows
Reference audio	Up to 3 audio files; public docs say each can be up to 30 seconds and 10 MB
Image input	Supported as a single reference image
Video input	The source article says yes via API for visual-context analysis

Outputs

Item	Translated / standardized spec
Output content	A complete audio track that may include voice, ambience, sound effects, and music
Maximum duration per generation	2 minutes
Supported languages	English and Chinese
Output formats	wav, mp3, pcm, ogg_opus
Sample-rate options	8,000 / 16,000 / 24,000 / 32,000 / 44,100 / 48,000 Hz
Channels	The source article says stereo

Generation Controls

Item	Translated / standardized spec
Adjustable parameters	Speed, volume, pitch
Voice cloning	Zero-shot, guided by reference audio
Dialect support	The source article says 20+ built-in dialects
Character count	The source article says 3+ characters in a single prompt
Long-form consistency	The source article says voice drift stays under 5% over 30+ minutes

Extended Capabilities

Item	Translated / standardized spec
Audio editing	Extending, repair/inpainting, stitching, line editing, alternate endings
Spatial audio	The source article describes 3D soundstage simulation
Foley automation	Text action descriptions mapped to sound effects

Access Routes

Platform	Endpoint / entry point	Status in this report
seedaud.io	seedaud.io (browser-based)	Live, all users
Volcano Engine Ark	Experience center / API access	Source-text claim; not fully verified on public English pages reviewed here
BytePlus	Enterprise access request	Source-text claim; not fully verified on public English pages reviewed here
Higgsfield	MCP server / Claude integration	Source-text claim; not fully verified on public English pages reviewed here
fal.ai	`bytedance/seed-audio-1.0`	Publicly verifiable

The rows on **reference audio**, **image input**, **2-minute generation length**, **supported languages**, **formats**, **sample rates**, and **control parameters** are supported by public fal documentation. The rows on **stereo output**, **20+ dialects**, **3+ speakers**, and **voice-drift under 5% over 30+ minutes** were not all equally verifiable on the public sources I reviewed, so I retained them as source-level claims rather than elevating them to fully confirmed public specifications.

How to Write Prompts

The next section of the Chinese article is practical rather than theoretical: how to actually prompt the model. Its basic message is correct and still useful. Good Seed Audio prompts are usually not vague requests like "make a conversation." They work better as **mini audio scripts**. Public fal guidance makes essentially the same point, saying the strongest prompts read less like high-level creative briefs and more like compact scene scripts.

A strong prompt typically contains six parts:

1. **Scene setup**: genre, environment, mood.
2. **Continuous sound bed**: the main ambient layer that runs under the whole scene.
3. **First speaker turn**: voice traits, emotion, pacing, and the spoken line.
4. **A concrete effect or transition**: one sound cue that advances the action.
5. **Second speaker turn**: dialogue that escalates or develops the scene.
6. **Ending cue**: silence, a closing sound, a music cue, or a fade.

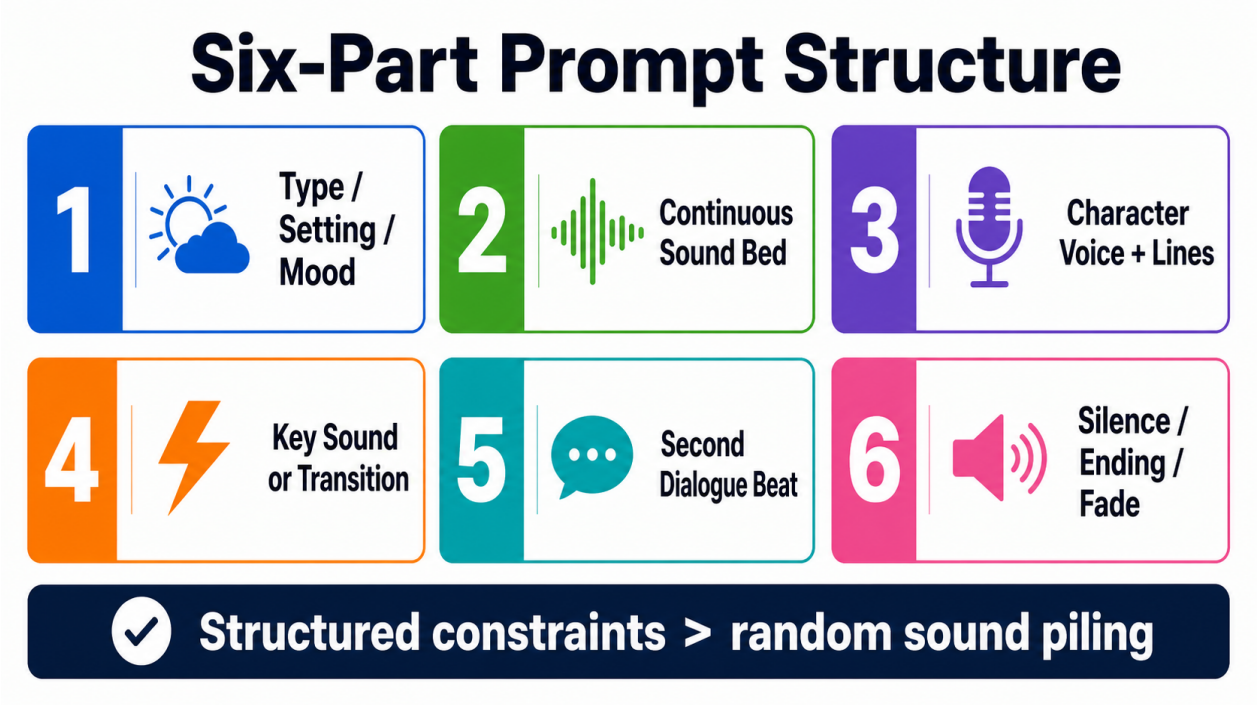

Figure 5: Six-Part Prompt Anatomy

The article's excellent practical advice is that you should **not overload the prompt**. One strong ambient bed plus a few precise cues usually works better than a giant shopping list of random effects. That aligns with publicly accessible fal prompting guidance.

A translated sample prompt from the Chinese article:

> [Fantasy adventure movie style. An ancient underground trial chamber with black stone walls, glowing runes, blue fire ahead and crimson fire behind. Tense, urgent, magical.] [A continuous roar of flames fills the chamber, with crackling embers, low stone echoes, and the hiss of heat.] Tobin (young boy, breathless shaky voice, panicked, speaking fast) blurts while backing away: "We're trapped! Blue fire in front, red fire behind – we're done, we're done!" [A steel beam overhead snaps with a violent metallic crack, sending sparks onto the floor.] Liora (young girl, clear steady voice, calm but quick) replies: "Don't panic. This isn't an attack – it's a logic puzzle. There's a note here… whoever built this place wanted us to think." [A brief silence. Only the flames remain.]

The Chinese article also lists a few recurring prompting mistakes, and they translate cleanly:

- prompts that are too generic,
- incomplete character descriptions,
- effect overload,
- mood/effect contradictions,
- and failure to specify the output language.

That advice still holds. Public fal docs likewise stress clarity around environment, speaker identity, vocal traits, delivery, dialogue, and constraints.

The article then walks through ten representative workflows. In American English, these come across most cleanly as:

sound-design-heavy full scenes, **TTS with saved reference voices**, **TA2A multi-character scenes**, **text-only T2A/TTS voice design**, **using your own voice**, **multi-audio compositing**, **audio extending**, **audio repair/inpainting**, **audio stitching**, and **localized spoken-line editing**.

One especially useful section covers building a reusable reference-voice library. The source says each reference clip should ideally be about 30 seconds, contain only one speaker, keep one stable tone and timbre, avoid music or extra voices, and use a clear, steady recording level. That lines up well with fal's public prompting guide and is worth preserving almost verbatim in substance.

For hands-on testing, visit **seedaud.io** to try Seed Audio 1.0 directly in your browser. After signing up, you can enter prompts, generate audio in real time, and preview results without any local setup.

Working with Seedance

Figure 6: Audio-First Seedance Video Workflow

One of the source article's strongest sections is the one about pairing Seed Audio with **Seedance**.

Its central idea is slightly counterintuitive but compelling: **generate the audio first, then use the audio to guide video generation**. Instead of treating sound as something you patch in after the visuals, you begin with a complete audio scene and build the visuals around it. Public fal materials explicitly market an "audio-first" workflow pairing Seed Audio with Seedance 2.0, and public Seedance documentation confirms that Seedance 2.0 supports **audio among its input modalities**.

The source article's logic is solid.

If a video model can condition on audio, then starting from the audio can improve:

- **lip sync**, because phonetic information is already there;

- **gesture and emotional timing**, because spoken delivery anchors body language;
- **scene atmosphere**, because rain, crowd noise, machinery, and similar cues can inform visual generation;
- and **duration control**, because the audio length already fixes the clip length more precisely than a vague text instruction can.

Public Seedance sources support enough of this claim to make the workflow plausible. The Seedance 1.5 Pro paper describes native audio-visual joint generation and multilingual/dialect lip-syncing, while the Seedance 2.0 model card publicly documents audio as one of the system's four supported input modalities.

The translated six-step workflow from the source article reads like this:

Step one: write a script with performance notes, not just bare dialogue. **Step two:** generate the audio track with Seed Audio, ideally with reference voices if consistency matters. **Step three:** review the audio before you ever touch the video stage. **Step four:** pass that audio into Seedance and write a visual prompt that complements rather than contradicts it. **Step five:** match video duration and other parameters to the finished audio. **Step six:** evaluate lip sync, emotional alignment, environmental coherence, and timing.

The source article is also right about common failure cases:

- using heavily compressed audio,
- pairing quiet intimate audio with explosive visual prompts,
- skipping audio review,
- omitting environmental sound entirely,
- and failing to provide visual identity anchors when character consistency matters.

That is good practical advice even when some of the surrounding ecosystem claims are still evolving.

The article also gives a version progression for Seedance:

Version	Key capability
Seedance 1.5 Pro	Native audio-video sync; multilingual lip sync
Seedance 2.0	Unified audio-video generation with text, image, audio, and video inputs
Seedance 2.5	Preview-stage claims in the source article

The first two rows are supported by public sources. The "2.5 preview" row appears in the source article, but I did not independently verify it in the public sources reviewed here, so I leave it as source-level context rather than a confirmed public spec.

Where You Can Access It

The source article lists several access paths aimed at different audiences.

For **all users**, the quickest option is **seedaud.io** — a browser-based platform with no API registration or coding required. For **domestic Chinese users**, it points to **Volcano Engine Ark**. For **overseas enterprise users**, it points to **BytePlus**. For **Claude/MCP-based workflows**, it points to **Higgsfield**. For **developers** comfortable with APIs, **fal.ai** provides a programmatic endpoint at `bytedance/seed-audio-1.0`.

Here the verification picture is uneven.

I could verify **fal.ai** publicly and directly. Public fal pages clearly list the model, the endpoint, public schema details, and public pricing.

I could also verify that public **ByteDance Seed** materials prominently document **Seedance 2.0** and the broader Seed model family, but I did **not** find a comparable standalone public English product page for Seed Audio 1.0 on the ByteDance Seed site during this review. On the official ByteDance Seed model page I reviewed, the visible model list includes Seed2.1, Seedance 2.0, Seedream 5.0 Lite, Seeduplex, and Seed GR-RL. I did not see Seed Audio 1.0 listed there.

What It Costs

The source article says pricing is simple: generated audio is billed by duration, down to the second. It gives three figures:

Platform	Price
seedaud.io	Free browser-based trial
ByteDance official / Volcano Engine	$0.18 / minute
BytePlus	300 minutes of trial credit after activation
fal.ai	$0.1875 / minute

Only the **fal.ai** figure was directly and publicly verifiable in the sources I reviewed. The fal model page explicitly lists **$0.1875 per minute**.

The source article then translates those prices into practical costs:

- about **$0.09** for 30 seconds,
- about **$0.90** for 5 minutes,
- and **$18.00** for 100 minutes.

Those arithmetic conversions are correct if you use the verified fal.ai rate of $0.1875 per minute.

Is It Worth Using

The article's final section is the strongest analytically, because it stops listing features and actually asks the right question: not "Is the audio quality good?" but "What category of tool is this, really?"

Its answer is that Seed Audio 1.0 is not best understood as just another TTS product.

That is the article's key insight, and it holds up.

The source explicitly argues that Seed Audio 1.0's real competitor is **not** a single TTS system like ElevenLabs. Its real competitor is the **entire traditional chain**:

TTS + sound-effects library + music generation + DAW mixing

That framing makes sense, especially in light of public fal positioning, which markets Seed Audio as a model that can generate dialogue, sound design, and music in one pass while also editing existing audio.

The article says its three most unusual strengths are:

- **complete scene-audio generation**,
- **natural-language-driven audio editing**,
- and **deep integration with Seedance-style audio-first video workflows**.

Those claims are directionally supported by public fal product documentation and by public Seedance primary sources.

The article is also candid about the weaknesses.

It says the language support is still narrow. Public fal docs support that: the currently verified public-language coverage is **English and Chinese**. It says per-generation duration is limited to **2 minutes**, which public fal docs also support. And it says the product's technical transparency is weak, which remains true in practical terms compared with products backed by fuller public technical reporting.

The source article's strongest use-case recommendations are also sensible. It argues that Seed Audio 1.0 is most compelling when you need **not just a voice, but a full sound world**:

- audiobooks and audio drama,
- AI-video creation,
- ad spots and game cutscenes,
- and creator workflows that need a stable brand voice across episodes or campaigns.

That is the best concise statement of the product's value proposition in the whole piece.

Its final judgment, translated cleanly, reads like this:

Seed Audio 1.0 matters less because any single sub-capability is unbeatable, and more because it changes the production assumption. The old assumption was "make the video first, then patch in audio." The new possibility is "build the audio first, then let the audio drive the rest of the workflow."

That is a strong and faithful rendering of the article's closing idea. To verify this for yourself, try running a piece of your existing content through seedaud.io and compare.

Turn a scene description into a complete sonic world.

Explore Seed Audio 1.0 in your browser.

Visit https://seedaud.io/