

Dify audio transcription and synthesis

Use this family to convert one authorized audio file to text or convert authorized text to speech. Both operations use the selected app credential and a stable Dify end-user identity.

SECTION	Connectors
TYPE	Connector
PROTOCOL	AIP 1.0
SOURCE	docs/connectors/dify/capabilities/audio.md

Audio content can contain sensitive personal data. Approval must cover the source material, selected app, end user, model invocation, and intended result handling before either request runs.

Choose an operation

Operation suffix	Method and route	Request	Result	Connector projection
audio.transcribe	POST /v1/audio-to-text	Multipart file and user	JSON transcript	All six app modes
audio.synthesize	POST /v1/text-to-audio	JSON body and user	Bounded binary	All six app modes

Both operations are medium-risk Tool mutations. Each requires tenant-policy approval and one Action idempotency key, publishes no safe retry or cancellation, and declares rollback unsupported.

The capability template is:

```
TEXT
cap:dify:<app_id>:<operation>
```

Use tenant discovery to select the app capability. The Action cannot replace the configured provider origin or Bearer key.

Transcribe one audio file

Before dispatch, verify that the audio belongs to the selected end user and that the chosen Dify app enables speech-to-text. Supply a safe filename, exact MIME type, and base64 content:

```
JSON
{
  "user": "tenant-user-0190",
  "file": {
    "filename": "request.wav",
    "content_type": "audio/wav",
    "content_base64": "UK1GRg=="
  }
}
```

The connector decodes the content and sends multipart fields named `file` and `user`. It does not send Action body as a multipart field for this operation.

The pinned provider accepts these MIME types:

- audio/mp3
- audio/mpga
- audio/m4a

- `audio/wav`
- `audio/amr`

The provider limit is 30 MiB. The connector permits at most 32 MiB of decoded upload data, while its base64 schema and the native Action envelope can impose a lower effective ceiling. The smallest applicable limit governs the request.

A successful result uses the generic JSON envelope:

```
JSON
{
  "http_status": 200,
  "body": {
    "text": "The transcribed speech appears here."
  },
  "content_type": "application/json",
  "provider_request_id": null
}
```

The connector does not normalize the language, confidence, speaker identity, or punctuation. At the pinned provider revision, the public response contains one required `text` value.

Synthesize bounded speech

The provider accepts `message_id`, `voice`, `text`, and `streaming` in the operation body. The connector inserts the validated `user` value.

Use direct text when the caller owns the content:

```
JSON
{
  "user": "tenant-user-0190",
  "body": {
    "text": "Your scheduled review begins at fourteen thirty.",
    "voice": "provider-voice-id"
  }
}
```

When `message_id` is present, Dify gives that message priority over `text`. The message lookup is constrained to the selected app and end user. Without a message ID, `text` is required.

`voice` is optional. Dify uses the app's configured voice when available, or a voice returned by its model provider. The connector does not publish a voice catalog and does not validate provider-specific voice identifiers.

The `streaming` field is reserved by Dify for compatibility. Provider output can arrive as streamed `audio/mpeg`, but this AIP operation buffers the bytes and returns one bounded binary result.

```
JSON
{
  "http_status": 200,
  "content_base64": "SUQz",
  "content_type": "audio/mpeg",
  "size_bytes": 3,
  "provider_request_id": null
}
```

The configured connector response limit applies to all returned audio. Its default is 8 MiB, with an allowed configuration range from 1 to 64 MiB.

Decode the result only after authorization, content-type policy, malware controls, and output-size checks. The connector does not write audio to a caller-selected filesystem path or preserve it as durable provider state.

Apply mutation governance

Each operation requires a non-blank Action idempotency key of at most 512 bytes. AIP retains the tenant-scoped reservation for 24 hours and revalidates the input hash on collision.

The key does not prove that Dify or the selected model provider deduplicates the invocation. Neither operation publishes automatic retry, transaction, reconciliation, cancellation, or compensation support.

Treat the uploaded audio, source text, referenced message, transcript, and generated bytes as confidential PII-bearing data. Retain only the evidence needed to establish authorization and outcome.

Recover safely

Observation	Decision
Filename, base64, user, or body is rejected before dispatch	Correct the same request without claiming a provider effect
Provider rejects the audio MIME type	Convert or select an allowed source under a new reviewed input
Audio exceeds an effective request limit	Reduce the authorized source; do not raise response limits
Speech-to-text or text-to-speech is disabled	Correct the selected app configuration before a new Action
Message lookup returns no usable audio	Verify the same app, end user, message ID, and message state
Voice is absent or unsupported	Select a provider-advertised voice or repair app configuration
Binary output exceeds the response bound	Treat the result as failed and review expected output size
Response is lost after model invocation	Treat the external outcome as uncertain; do not replay automatically
Returned transcript or audio crosses an app or user boundary	Stop traffic and treat the event as an isolation incident

Preserve the Action ID, input digest, approval, idempotency key, provider request identity, and any bounded result used to settle an ambiguous outcome.

Related documentation

- [Dify capability index](#)
- [Dify app metadata and files](#)
- [Dify connector configuration](#)
- [Authentication and credential scopes](#)
- [Errors and retry decisions](#)