FIBRETTS: A FIBRE-BUNDLE VIEW ON CONTENT-DEPENDENT CONTROL IN TEXT-TO-SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable text-to-speech (TTS) aims to preserve target words while expressing requested attributes such as emotion, speaking rate, and emphasis. Yet language-model-based systems often produce different effects for the same instruction across texts, and stronger control can introduce content errors. To explain this interaction, we propose FibreTTS, drawing on fibre-bundle theory to distinguish content from reading while allowing their relationship to vary. Content representations form the *base space*, and the possible readings of each representation form its *fibre*. We construct text- and reference-associated coordinates that explain most of the observed representation variance, providing empirical support for this structured description. Building on this model, we examine two complementary properties and their implications for control. *Local triviality* provides shared coordinates for what is spoken and how it is read within a content neighborhood, while allowing the same reading change to have different acoustic realizations across texts. *Fibre transport*, defined through a chosen connection, guides adaptation of a source reading change to target content while retaining the target's own baseline reading. Experiments on two TTS systems, using public emotional speech data and private financial outbound-call data, support the coordinates' behavioral relevance and reveal the effects of direct cross-text control reuse. Transport-inspired residual updates improve the instruction-following scores on outputs whose original control effect was weak, and the improvements are statistically significant on both systems and datasets.

1 INTRODUCTION

Controllable text-to-speech (TTS) generation aims to express requested speaking attributes while preserving the target words. Language-model-based systems support control through instructions, reference recordings, and internal interventions (Xie et al., 2025a; Ji et al., 2025; Wang et al., 2026). Yet the same request can have different effects across texts, and stronger control can introduce content errors (Wang et al., 2026; Xing et al., 2025). Although text and control can be specified independently, they interact during inference. Understanding this interaction requires a description that distinguishes content from reading while allowing their relationship to vary across samples.

A suitable model should distinguish content from reading while capturing how shared speaking attributes are realized differently across texts. Fibre bundles provide such a structure by organizing a space into families associated with different base points, and have been used to describe structured variation in representation learning (Courts & Kvinge, 2022; Scoccola & Perea, 2023). We adapt this perspective to speech generation and introduce **FibreTTS**. Content representations form the *base space*, and the possible readings associated with each representation form its *fibre*. An audio output is thus described locally by what is spoken and how it is read. To ground this model empirically, we decompose model logits into text- and reference-associated main effects and extract candidate coordinates for the two components. Across the evaluated model–dataset combinations, 20 dimensions per component explain approximately 75–81% of the observed variance. These results support a compact structured description, without by themselves establishing that the coordinates fully separate content and reading.

Building on this model, we characterize local content–reading variation through *local triviality*, providing a geometric description of content-dependent control effects. Within a content neighborhood,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shared coordinates distinguish what is spoken from how it is read, while their joint mapping to audio allows the same reading change to have different acoustic realizations across texts. Different neighborhoods may require different coordinates, so a description valid locally may not extend across the entire space. We use this structure to describe control mismatch through content and reading deviations. The resulting reading may fail to express the requested attribute, or the output may leave the content region associated with the target text and produce incorrect words. This formulation distinguishes the two failures without assuming that the same adjustment is suitable for every text.

To address this mismatch, we formulate a *fibre transport* construction that preserves specified relative reading coordinates, guiding adaptation of source changes to the target baseline. We choose a baseline reading for each content and, given compatible relative reading coordinates, construct a connection whose parallel transport preserves those coordinates across contents. This construction maps the source baseline to the target baseline and transfers the change between baseline and controlled readings. When the preserved coordinates capture the requested attributes, the transferred change provides a suitable adjustment for the target content. This motivates retaining the target's own baseline and applying an adapted source change, rather than directly copying the source adjustment. The required correspondence must be identified from the task and data, and its effectiveness is evaluated through interventions in the TTS models.

We assess the practical implications of these properties on CosyVoice2 (Du et al., 2024) and IndexTTS-2.5 (Li et al., 2026), using the public emotional speech dataset ESD (Zhou et al., 2022) and a private dataset from real-world financial outbound calling operations. The latter covers diverse call scripts and human-annotated speaking attributes, including pauses, emphasis, speaking rate, noise, and audio quality. Coordinate interventions examine whether the candidate content and reading coordinates have the expected behavioral roles, while cross-text exchange experiments test the limitations of directly reusing control effects. We further evaluate transport-inspired adaptation on receivers with low original instruction-following scores. Together, these experiments assess the framework's explanatory and practical value across both emotional speech and real-world business conversations from financial outbound calls.

Our contributions are threefold: First, we introduce FibreTTS, a fibre-bundle model of content and reading, supported empirically by a low-dimensional decomposition of model representations in Section 2. Second, we characterize content-dependent control through local triviality, using coordinate interventions and representation analysis to examine the limitations of direct cross-text transfer in Section 3. Third, we formulate fibre transport for cross-content control adaptation and show that a representation-level approximation improves instruction following on weakly controlled outputs, as we report in Section 4.

2 FibreTTS: Formulation and Empirical Grounding

2.1 Preliminaries

Controllable TTS. We study language-model-based TTS (Du et al., 2024): $z \sim p_\theta^{\mathrm{LM}}(\cdot \mid x, c), \quad y = D(z; d)$. This model family generates speech tokens z jointly from target text x and control signal c, and the decoder D produces waveform y under fixed downstream conditions d. The signal c specifies speaking attributes through reference inputs, instructions, or internal interventions. The goal is to express these attributes while preserving the target words.

Representation space. Similar to Gao et al. (2026), we describe speech realizations in an idealized output space E, where each point e represents a complete waveform $y = G(e)$. Our analysis assumes *finite-dimensional smooth regions* with continuous content variation, enabling the fibre-bundle formulation below. Appendix B states these assumptions and explains why we use continuous content representations rather than discrete text labels.

2.2 FibreTTS Formulation

2.2.1 Fibre-Bundle Structure

Fibre-bundle theory describes a family of spaces attached to different base points (Daniel & Viallet, 1980). A fibre bundle (E, B, π, F) consists of a total space E, a base space B, a projection π, and

a typical fibre F. The projection $\pi : E \to B$ assigns each point $e \in E$ to a base point $b = \pi(e)$. All points assigned to the same b form its *fibre*, $E_b = \pi^{-1}(b)$, which is nonempty. The *typical fibre* F is a fixed space used to describe points within different fibres. For example, if the fibres are lines, F can be $\mathbb{R}$, so that a point on each line is described by a real number f. The choice of how f corresponds to a point on each fibre is specified by a local coordinate map.

The key requirement is that fibres over a neighborhood can be described together using one such map. Specifically, every base point has an open neighborhood $U \subseteq B$ with a map

$$\varphi_U : \pi^{-1}(U) \longrightarrow U \times F, \qquad \varphi_U(e) = (b, f), \quad b = \pi(e). \tag{1}$$

Here $\pi^{-1}(U)$ contains all points whose base points lie in U. The map φ_U describes each of these points by a pair: b identifies which fibre it belongs to, and f identifies where it lies within that fibre. Every pair $(b, f) \in U \times F$ corresponds to exactly one point in $\pi^{-1}(U)$. This is what we mean by a *local coordinate system*: throughout U, points can be expressed and recovered using these two components. In our setting, the spaces and maps are smooth: small changes in a point lead to small changes in its coordinates, without abrupt jumps or sharp changes (Daniel & Viallet, 1980; Courts & Kvinge, 2022). Write $\psi_U = \varphi_U^{-1}$ for the map that recovers a point from (b, f).

2.2.2 Content–Reading Formulation

FibreTTS uses the idealized audio representation space E introduced in Section 2.1, where each point e represents a complete audio output $G(e)$. Figure 1 illustrates this formulation with two content dimensions and one reading dimension.

Content modelling. Let R be an idealized speech recognizer that assigns a probability distribution over possible transcriptions to each audio output. We use these distributions as content representations and model their variation as a continuous and smooth space B of fixed dimension. The projection is

$$\pi(e) = R(\cdot \mid G(e)). \tag{2}$$

A distribution $b \in B$ can then be easily decoded into textual contents by selecting its most probable transcription. The curved surface represents B, and the arrow $\pi(e)$ maps an audio realization to its content point b, which decodes to "The train leaves at nine." Thus, B retains a continuous representation of content while allowing each point to be interpreted as spoken words through its most probable transcription.

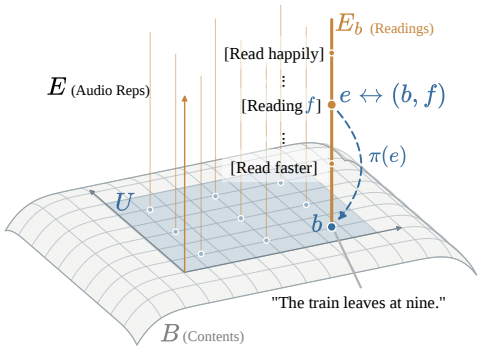


Figure 1: An illustration of FibreTTS. The base space B represents content, and each fibre E_b contains its possible readings. Local coordinates (b, f) specify what is spoken and how it is read.

Reading modelling. The fibre $E_b = \pi^{-1}(b)$ contains all audio realizations with the same content representation b. This fibre is depicted as the vertical line above b. Different points along this line represent different readings of "The train leaves at nine," such as a happy reading or a faster reading. These points share the same content representation but differ in how the sentence sounds.

Within a neighborhood $U \subseteq B$, FibreTTS uses the local coordinate map to represent each realization uniquely by (b, f) and recover it as $e = \psi_U(b, f)$. The figure draws U as a flat patch to illustrate this local description: b selects a vertical line, and f selects a reading along it. The pair therefore specifies a complete audio realization without loss. A control signal c guides the model toward the requested reading while aiming to preserve the target words.

Claim 1. Model representations should support an approximate local decomposition into content and reading coordinates.

2.3 Empirical Grounding

To evaluate Claim 1, we test whether model representations admit an approximate content–reading decomposition and whether the resulting coordinates are locally effective.

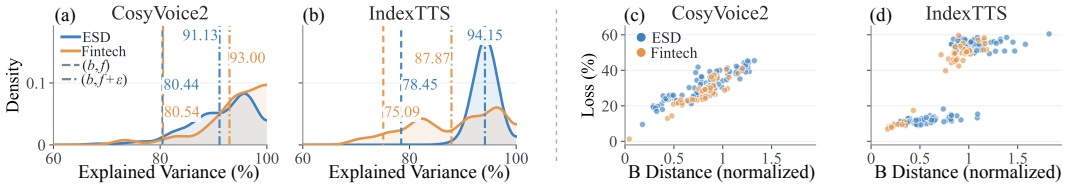


Figure 2: Content–reading reconstruction. (a,b) Explained variance with local residual correction; vertical lines indicate means before and after correction. (c,d) Reconstruction loss versus B distance when reusing the residuals of a fixed source text on other texts.

2.3.1 Setup

For models, we use CosyVoice2 (Du et al., 2024) and IndexTTS (Li et al., 2026). Their LMs jointly process target text and reference information. The latter could be reference text or speech tokens, which provides an example reading that guides generation as a control signal (Du et al., 2024). For data, we use the public **ESD** (Zhou et al., 2022) dataset, with speech in Chinese and English for anger, happiness, sadness, and surprise. We also use a private **Fintech** set of Chinese financial outbound-call speech with human comparisons of pauses, emphasis, speaking rate, noise, and audio quality. This extends evaluation beyond emotion to business conversations. Dataset details and the rating guidelines appear in Appendix C.

2.3.2 Constructing FibreTTS from Representation Space

We use the raw logits before the first speech token is sampled as our investigated space. Similar to the ideal space proposed in Section 2.1, logits are continuous-valued, and changing them directly affects token sampling and the generated audio.

Preparing the logits. For each dataset, we select target texts $\mathcal{T}$ and references $\mathcal{R}$, each consisting of an audio recording and its transcript. We form the full grid $\mathcal{T} \times \mathcal{R}$. Alongside the target text, CosyVoice2 receives both reference text and audio, while IndexTTS receives only reference audio, following their respective input formats. For each $(t, r) \in \mathcal{T} \times \mathcal{R}$, we extract raw logits $h_{t,r} \in \mathbb{R}^d$. Appendix D lists the grid sizes and the logit extraction settings.

Decomposing text and reference effects. To build the content–reading structure in FibreTTS here is to separate two patterns in the logits: variation driven by the target text, which we call B, and variation driven by the reference, which we call F. We use two-factor decomposition to separate the two input effects. For each text, we average its logits over all references to estimate the text's contribution. Likewise, we average over texts to estimate each reference's contribution. Subtracting the overall mean μ gives the marginal text effect u_t and reference effect v_r:

$$\mu = \frac{1}{|\mathcal{T}||\mathcal{R}|} \sum_{t,r} h_{t,r}, \qquad u_t = \frac{1}{|\mathcal{R}|} \sum_r h_{t,r} - \mu, \qquad v_r = \frac{1}{|\mathcal{T}|} \sum_t h_{t,r} - \mu. \tag{3}$$

In our balanced grid, these mean effects minimize the squared additive-fit loss $\sum_{t,r} \|h_{t,r} - \mu - u_t - v_r\|_2^2$ under zero-mean constraints on u and v. Subtracting this fit and the overall mean leaves $I_{t,r} = h_{t,r} - \mu - u_t - v_r$, the text–reference interaction in the observed grid. It records how a particular reference response changes with the text beyond their separate average effects.

To obtain compact coordinates, we stack the text effects $\{u_t\}_{t \in \mathcal{T}}$ and reference effects $\{v_r\}_{r \in \mathcal{R}}$ into two matrices and apply SVD separately. For each matrix, we retain the 20 directions that capture the most variation. These directions form D_B and D_F. The coordinates $b_t = D_B^\top u_t$ and $f_r = D_F^\top v_r$ give the weights of these directions for each text and reference. Thus, $D_B b_t$ approximates the original text effect, and $D_F f_r$ approximates the reference effect:

$$h_{t,r} = \mu + D_B b_t + D_F f_r + \varepsilon_{t,r}. \tag{4}$$

Here $\varepsilon_{t,r}$ is the additive-fit residual after dimensionality reduction. It contains the interaction $I_{t,r}$ and the main-effect variation outside the retained 20 directions. Appendix D explains how the same coordinates are read out from a single representation.

2.3.3 Results and Analysis

Coordinate quality. The extracted content and reading coordinates capture most of the observed logit variation. Figure 2(a,b) shows that simply representing by (b_t, f_r) explains approximately

75–81% of the variance across both models and datasets. If we further include a locally estimated residual correction in the reading representation, written schematically as $(D_B b_t, D_F f_r + \widehat{\varepsilon}_{t,r})$, the explained variance rises to 88–94%. As Appendix F shows, its response to reference changes is partly predictable at fixed content and varies across contents, motivating this expanded local reading coordinate for the analyses that follow.

Locality of reconstruction. The residual correction ε becomes less reliable when reused on texts farther from its source in B. In Figure 2(c,d), reconstruction loss generally increases with distance for CosyVoice2. IndexTTS shows a sharper separation between low- and high-loss groups, suggesting more abrupt changes across content regions. This indicates that the high-quality fibre coordinate is only effective within a local content region.

Together, these findings support Claim 1: the representations admit an approximate content–reading decomposition, with residual corrections that depend on the local content region. Appendix D details the reconstruction and evaluation of the coordinates.

3 Locality and Control Mismatch

Building on the content–reading model of FibreTTS, we examine how locally shared coordinates accommodate control effects that vary across texts. We first describe these effects through local triviality and distinguish content errors from reading mismatch, then use coordinate interventions and cross-text transfer experiments to assess this account.

3.1 Local Triviality of FibreTTS

Local triviality gives each region its own coordinate system. In region U, a realization is written as (b, f_U). In region V, it is written as (b, f_V). The two reading coordinates need not have the same form. The same value of f can describe one reading in U and a different reading in V. On the overlap $U \cap V$, the two descriptions are linked by a transition map $f_V = g_{VU}(b, f_U)$, which Appendix B defines. This map depends on the base point b. So the conversion between the two coordinate systems changes with content.

Now consider reusing a reading coordinate f_U from region U in region V without this conversion. Two cases arise. If $U \cap V \neq \varnothing$, the reuse causes partial side effects. For $b \in U \cap V$, f_U is still a valid coordinate. For other base points in V, the correct value is f_V, and it cannot be obtained from f_U without g_{VU}. The reuse then gives the mismatch $\delta f = f_U - f_V$. If $U \cap V = \varnothing$, no shared chart links the two regions. A displacement that moves along the fibre in U is then only a vector in E. Applied at a base point in V, it need not move along the fibre of V. Part of it can point across fibres and change the content coordinate. Section 3.2 makes this decomposition precise.

3.2 Characterizing Control Mismatch

We now introduce control into the content–reading structure. Consider an externally specified signal, such as a fixed prompt "Read happily." In FibreTTS modelling, we can model it as a fixed displacement c. We first choose a base point in U, where c successfully produces the expected reading while preserving the content. We then apply the same c to another base point in V, with $U \cap V = \varnothing$. We consider the case where it has two nonzero components in the target's local coordinates: one changes content, and the other changes reading. Let the target baseline be $e = \psi_V(b, f^0)$. For a small displacement, a first-order expansion gives

$$\begin{aligned}(\Delta b, \Delta f) &= D\varphi_V(e)c, \\ c &= \underbrace{D\psi_V(b, f^0)(\Delta b, 0)}_{c_B} + \underbrace{D\psi_V(b, f^0)(0, \Delta f)}_{c_F}. \end{aligned} \tag{5}$$

Here D denotes the derivative. The first line expresses the displacement in content and reading coordinates. The second separates it into a content component c_B and a fibre component c_F, using the target's local map at the baseline reading.

To first order, the controlled output has coordinates

$$\varphi_V(e + c) \approx (b + \Delta b,\ f^0 + \Delta f). \tag{6}$$

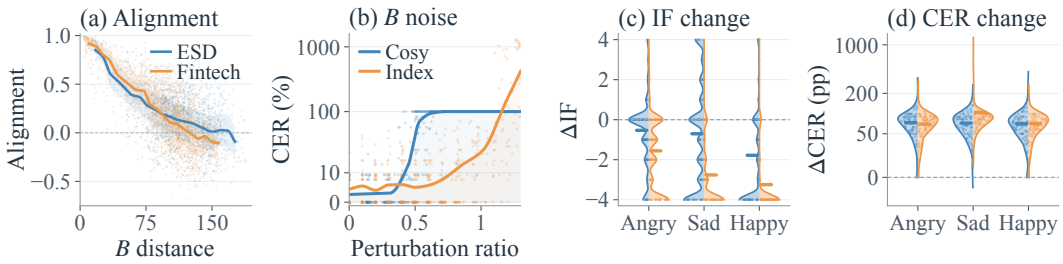


Figure 3: Local variation and direct reuse. (a) Alignment of residual reference responses versus content distance: blue is ESD and orange is Fintech (CosyVoice2). (b) Character error rate under increasing B perturbation: blue is CosyVoice2 and orange is IndexTTS (ESD). (c,d) Paired changes in instruction following and character error rate after direct F reuse; blue is CosyVoice2 and orange is IndexTTS. Each half of the split distributions shows one model.

The output therefore moves toward a reading on the displaced fibre $E_{b+\Delta b}$. If the new base point decodes to different words, a content error occurs. If the resulting reading fails to express the requested attribute, a control mismatch occurs. The same displacement can therefore change both the spoken content and the reading when reused without adaptation.

Claim 2. A control effect that works for one text may fail on another. Direct reuse can weaken the requested attributes or introduce errors in the spoken words.

3.3 Empirical Analysis

We examine Claim 2 through two complementary tests. The first test works on the representation. It asks whether the same reference change produces the same response at different texts. If the responses differ, the reading coordinate depends on content. The second test works on generated speech. It either changes the content coordinate or copies the reading coordinate from another text without adaptation, and it measures the effect on the spoken words and on the requested attribute. The first test shows where the coordinates differ across texts. The second test shows what happens when this difference is ignored.

3.3.1 Experimental Setup

Control and sampling. We examine three control interfaces: a *reference* recording exemplifies an attribute, a natural-language *prompt* describes it, and *steering* shifts a representation along an attribute direction. For the locality tests, residual reference responses are measured on both models with ESD and Fintech; Figure 3(a) presents CosyVoice2, and the appendix presents all four combinations. Direct F reuse uses both models on ESD. The direct reuse comparison has 50 texts and 10 references per condition. Following common practice in activation steering (Wang et al., 2026; Xie et al., 2025b; Zhou et al., 2026), we construct the bases once and apply the interventions at every decoding step. This assumes that the B/F bases remain meaningful after the first speech token. The B sweep uses both models on ESD and Fintech, with 10 texts and 5 references per strength in each dataset. The transport study in Section 4.3 uses both models, both datasets, and all three interfaces. Appendix C describes the data and evaluation, and Appendix E describes the coordinate interventions and reports the full result tables.

Evaluation. *Instruction following (IF)*(Xie et al., 2025a) is an audio-judge score from 1 to 5, with higher scores indicating better fulfillment. We use Gemini-3.1-Pro (Team, 2025) to evaluate IF. *Character error rate (CER)*, computed from Qwen3-ASR (Shi et al., 2026) transcriptions, measures content errors; lower is better and values may exceed 100% when insertions are numerous. Coordinate interventions report both metrics; transport reports IF. Paired changes compare the same receiver before and after the additional intervention. Appendix C gives the scoring procedure.

3.3.2 Results

Reference responses depend on content. For the same reference pair, we compare the change in the additive-fit residual across texts. Figure 3(a) shows that these response directions generally

become less aligned as B distance increases. The nearest and farthest fifths have mean cosines of $0.45/0.01$ on ESD and $0.60/-0.06$ on Fintech. Thus, the part of a reference response left beyond the shared additive coordinates varies with content. This is representation-level evidence; it does not by itself establish a corresponding trend in generated audio. Appendix F defines the response and the distance used here, and it reports both models.

B perturbations can disrupt content. We add a fixed random direction in B at every decoding step and increase its magnitude, preserving the other components at each injection. Figure 3(b) shows relatively small CER changes at low strengths, followed by a pronounced increase. For CosyVoice2, CER rises from 2.6% to 70.4% as the perturbation-to-state norm ratio increases from 0.31 to 0.53. IndexTTS shows a later rise, from 13.1% at 0.84 to 165.8% at 1.20. These results show sensitivity along the tested direction, rather than a universal stability radius.

Direct F reuse harms generation. Within the same emotion and control interface, we randomly assign a donor with different target text. At every step, we replace the receiver's entire F coordinate with the donor's, retaining the receiver's current B readout and residual. Figure 3(c,d) shows all three emotions under prompting: IF falls and CER rises in both models. Across all 14 available model–emotion–interface conditions, every mean change has the same adverse sign, and the corresponding nominal intervals exclude zero. This directly demonstrates side effects of unadapted reuse. References are not fixed across donors and receivers, and trajectories are aligned by decoding step; these results do not isolate B distance as the cause.

Together, the response analysis and interventions support content-dependent control and show that preserving a coordinate readout at injection does not guarantee reliable speech. They motivate adapting source effects to the receiver, which we study next.

4 Fibre Transport for Control Adaptation

The preceding analysis reveals limitations of directly reusing reading changes across texts. We now develop a target-dependent adaptation principle through fibre transport, translate it into a representation-level intervention, and evaluate its effect on instruction following.

4.1 Transporting Relative Reading Changes

Transferring a reading change requires a rule for relating readings across contents. We use parallel transport to define this rule. In Figure 4, the base path $\gamma(t)$ connects b_1 and b_2. A chosen connection specifies how a reading moves as the content follows this path. Starting from $e_1 \in E_{b_1}$, it gives the lifted path $\widetilde{\gamma}(t)$ and its endpoint $e_2 \in E_{b_2}$. We write P_γ for the resulting transport:

$$e_2 = P_\gamma(e_1) = \widetilde{\gamma}(1), \widetilde{\gamma}(0) = e_1, \pi(\widetilde{\gamma}(t)) = \gamma(t). \quad (7)$$

Thus, the reading stays on the fibre above the current content throughout the path.

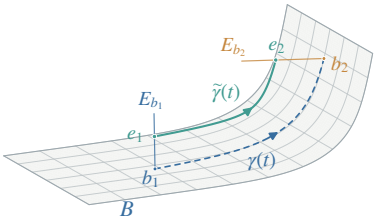


Figure 4: An illustration of parallel transport. The base path $\gamma(t)$ connects contents b_1 and b_2, while its horizontal lift $\widetilde{\gamma}(t)$ transports reading e_1 to e_2 under the chosen connection.

To transport a control effect, we measure each reading relative to its own baseline. For $i \in \{1, 2\}$, let e_i^0 be the baseline at b_i, with reading coordinate f_i^0. Let f_i be the coordinate of e_i. When compatible reading coordinates are available along the path, we choose a connection that preserves the relative change:

$$P_\gamma(e_1^0) = e_2^0, \qquad f_2 - f_2^0 = f_1 - f_1^0. \quad (8)$$

The baseline moves to the target baseline, and the reading retains its relative position. These coordinates must be chosen so that the preserved change expresses the same requested attribute across contents. This requires task information; it is not guaranteed by the bundle structure alone. Appendix G gives the construction of this connection and its conditions.

Now keep the external signal c fixed and suppose it produces a successful source reading $e_1(c)$. Under the task correspondence above, its transported reading $e_2^\star(c) = P_\gamma(e_1(c))$ provides a suitable

target. Write its local coordinates as $(b_2, f_2^\star)$ and those of the actual target output $e_2(c)$ as $(b_2 + \Delta b, f_2^\star + \Delta f)$. We correct the output by reducing these two deviations:

$$\varphi_{U_2}(\widehat{e}_2(k)) = \big(b_2 + (1-k)\Delta b,\ f_2^\star + (1-k)\Delta f\big), \qquad 0 \le k \le 1. \tag{9}$$

At $k = 0$, the output is unchanged. At $k = 1$, it reaches the transported reading. The external signal stays fixed; the correction changes its effect on the target output. Appendix G explains how control deviations separate into content and reading components, and when a content change alters the recognized words of the target text.

4.2 Representation-Level Adaptation

The preceding construction defines adaptation in an idealized output space. We approximate it in model representations using the B/F bases constructed above. For each state h, we jointly read out (b, f) and obtain the projection residual $r = h - \mu - D_B b - D_F f$, following Eq. (14). We call the source sample the *donor* and the target sample to be improved the *receiver*. A fitted predictor $q(b, f)$ estimates this residual at given content and reading coordinates; Appendix D motivates this predictor. We combine the receiver's prediction with the donor's deviation from its prediction:

$$\begin{aligned} \widehat{r}_{\text{receiver}} &= q(b_{\text{receiver}}, f_{\text{receiver}}) + \big[r_{\text{donor}} - q(b_{\text{donor}}, f_{\text{donor}})\big], \\ \widehat{h}_{\text{receiver}}(k) &= h_{\text{receiver}} + k\big(\widehat{r}_{\text{receiver}} - r_{\text{receiver}}\big). \end{aligned} \tag{10}$$

The predictor is fitted on crossed text–reference states without IF labels. Its outputs lie in the same residual space, so the update preserves the receiver's measured B/F coordinates at injection. We select a donor with high IF that shares the receiver's requested attribute and control method but has a different target text. The receiver keeps its original text and external control. Here $k = 0$ gives the original generation and $k = 0.5$ applies half the proposed residual change.

We apply the update at the first speech-token representation in both models. Each $k = 0.5$ output is paired with a $k = 0$ output under the same model, target text, reference, external control, and generation conditions. The paired difference therefore measures the added correction.

Claim 3. Given a task-compatible relative reading coordinate, transport maps an acceptable source reading to an acceptable target reading along an admissible path, and it keeps the spoken words of the target text.

4.3 Empirical Evaluation of Transport Effects

Setup. We use the B/F coordinates from Section 2.3.2 and the control methods in Section 3.3.1. A *donor* has original IF 5, and its *receiver* has original IF 1–3. The two share the control method and the requested attribute but differ in target text. For reference control, they use different reference pairs with the same attribute; for prompt and steering control, they use the same reference input and control setting. This gives 1,216 CosyVoice2 and 661 IndexTTS pairs. We estimate uncertainty with a crossed bootstrap (Owen, 2007) over target texts and references, and report two-sided bootstrap p-values for the mean IF change. Appendix H gives the predictor, pairing, and bootstrap details.

Results. Table 1 shows that mean IF increases after transport in every tested group except IndexTTS with prompt control for sadness on ESD, where it remains 1.00. The symbol "–" indicates that no donor with IF 5 was available. Combining the three control methods within each dataset gives statistically significant improvements: $p < 0.001$ on both datasets for CosyVoice2, and $p = 0.0015$ on ESD and $p = 0.030$ on Fintech for IndexTTS, all below 0.05 (Owen, 2007; Demšar, 2006). Appendix H also reports CER before and after transport for every control method and attribute. CER changes by less than a point in nearly all conditions, so the IF gains do not come at the expense of the spoken words of the target text.

These results support Claim 3: successful donors improve control for receivers with low original IF on average, without harming the spoken words. Our setting is controlled: original IF scores guide donor selection, and both samples' representations are available to estimate the transport function. Practical use requires donor selection with fewer evaluations and transport functions that generalize to new texts and new control conditions.

Table 1: IF before and after transport, classified by control method and attribute. Improvements appear in green; the last column is the averaged score over both datasets. A dash indicates that no valid donor with IF=5 is available under that control condition. The improvements of the two models on both datasets are statistically significant, with maximum $p = 0.03 < 0.05$ (Demšar, 2006).

		ESD				Fintech						All
Control		Anger	Happiness	Sadness	Surprise	Pauses	Emphasis	Speaking rate	Noise	Timbre	Audio quality	**Average**
CosyVoice2												
Reference	Before	1.28	1.82	1.48	1.45	2.02	2.14	–	1.75	2.67	1.69	**1.63**
	After	1.43 (+0.15)	2.20 (+0.38)	2.21 (+0.73)	1.54 (+0.08)	2.53 (+0.52)	2.34 (+0.21)	–	3.50 (+1.75)	3.00 (+0.33)	2.12 (+0.44)	**1.96** (+0.33)
Prompt	Before	1.35	1.77	1.66	1.65	2.45	–	2.56	1.50	2.76	1.76	**1.86**
	After	2.76 (+1.42)	3.32 (+1.55)	3.55 (+1.90)	3.19 (+1.55)	3.27 (+0.82)	–	3.26 (+0.70)	5.00 (+3.50)	3.48 (+0.71)	3.18 (+1.42)	**3.17** (+1.31)
Steering	Before	1.17	1.59	2.00	1.18	1.72	–	–	–	2.00	1.60	**1.44**
	After	1.45 (+0.28)	1.78 (+0.20)	2.46 (+0.46)	1.62 (+0.45)	2.23 (+0.51)	–	–	–	2.43 (+0.43)	2.85 (+1.25)	**1.92** (+0.48)
IndexTTS												
Reference	Before	1.38	2.50	1.67	2.20	2.46	2.12	2.38	1.00	2.71	1.74	**2.05**
	After	1.96 (+0.58)	3.00 (+0.50)	3.13 (+1.47)	3.80 (+1.60)	2.91 (+0.46)	2.33 (+0.20)	2.44 (+0.06)	5.00 (+4.00)	3.62 (+0.90)	2.91 (+1.18)	**2.68** (+0.63)
Prompt	Before	1.64	1.03	1.00	2.33	2.25	2.21	2.31	2.00	2.27	2.08	**2.01**
	After	2.70 (+1.06)	1.17 (+0.14)	1.00 (+0.00)	3.67 (+1.33)	2.66 (+0.41)	2.26 (+0.05)	2.46 (+0.14)	4.33 (+2.33)	2.62 (+0.35)	2.60 (+0.52)	**2.47** (+0.46)
Steering	Before	1.41	3.00	1.78	2.50	2.52	–	2.27	2.00	2.87	1.85	**2.07**
	After	1.75 (+0.33)	4.00 (+1.00)	3.22 (+1.44)	3.50 (+1.00)	2.94 (+0.42)	–	2.36 (+0.09)	5.00 (+3.00)	3.40 (+0.53)	2.76 (+0.91)	**2.59** (+0.52)

5 Related Work

Controllable audio and speech generation. Existing methods control generated audio through text, reference recordings, or changes to model representations. Text descriptions specify the desired sounds or music in general audio generation (Liu et al., 2023; Copet et al., 2023). In TTS, reference recordings demonstrate a speaking style, while natural-language descriptions state the requested attributes (Wang et al., 2018; Guo et al., 2023; Yang et al., 2024; Ji et al., 2025). Other methods modify internal activations to control emotion, both in language-model-based TTS (Wang et al., 2026; Zhou et al., 2026) and in flow-matching TTS (Xie et al., 2025b). FibreTTS studies a question shared by these methods: why does the same control work differently across texts, and how can a successful control effect be adapted to another text?

Geometric modeling. Geometric studies examine how concepts and their relations are organized in model representations, from linear directions to multidimensional structures (Park et al., 2024; 2025; Engels et al., 2025). Other work uses geometry to design representations, such as hyperbolic spaces for hierarchical relations (Nickel & Kiela, 2017; Ganea et al., 2018). Closest to our work, sample-dependent geometric models explain why steering effects vary across inputs (Gao et al., 2026). FibreTTS applies this perspective to speech control: it models the relationship between content and reading, derives the implications for control, and tests them through representation analysis and generation experiments on two TTS systems.

Fibre bundles in representation learning. Fibre bundles organize a space into families that share a coordinate system locally. Prior work uses them to model the inputs behind each output of a many-to-one map (Courts & Kvinge, 2022), to reduce dimensionality while preserving topology (Scoccola & Perea, 2023), and to define features that transform consistently across local frames (Cohen et al., 2019). FibreTTS instead uses fibre bundles to describe the possible readings of each content, and a chosen connection defines how a reading moves between contents.

6 Conclusion

We presented FibreTTS, a fibre-bundle framework that organizes TTS representations into content coordinates and content-dependent reading families. A low-dimensional logit decomposition explains 75–81% of observed variance, and the tested direct reuse of reading coordinates across texts reduces instruction following and content accuracy. The proposed residual update raises mean IF among low-scoring receivers, and the improvements are statistically significant on both models and datasets. Future work can extend the transport function to new texts and assess content accuracy alongside instruction following.

AI USE STATEMENT

In this work, we used generative AI tools to edit the paper to improve readability, including grammar, clarity, and presentation, and to help developing software code for the experiment scripts. We have not used generative AI tools to develop the theoretical model, to propose or refine hypotheses, to design research methodology or experiments, or to interpret results. All theoretical modelling, experiments, analyses, visualizations, and scientific conclusions were produced by the authors. Generating synthetic datasets, assisting with translation, and supporting qualitative or thematic data analysis are not applicable to this work.

We have reviewed all AI-assisted work. AI-edited text was checked by the authors for accuracy of meaning, and AI-generated code was reviewed and tested for correctness by the authors before use. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses a private dataset built from an industry partner's customer-service recordings and human ratings of synthesized speech. The private dataset is used only for evaluation and is not released; the paper reports aggregate statistics and contains no transcripts, recordings, personal information, or business information. All human raters took part voluntarily, were informed of the task, and were compensated for their work. The proposed model is a general theoretical account of controllable speech generation and applies to both public and private settings. Our experiments analyze existing open TTS systems and do not release new synthesis models, so the work adds no misuse risk beyond that of the systems it studies. We are not aware of any conflicts of interest or other ethical concerns.

REPRODUCIBILITY STATEMENT

We have taken several steps to make this work reproducible. Section 2.3.1 and Appendix C describe the public ESD dataset with its original reference, the construction of the private Fintech dataset, the human rating guidelines, and the control and evaluation procedures, including the exact prompts for the instruction-following judge. The theoretical assumptions behind FibreTTS are stated in Appendix B, and the derivations for coordinate construction and relative reading transport are given in Appendices D and G. The experimental designs for the coordinate interventions and the transport study are described in Section 3.3.1, Section 4.3, and Appendices E and H. In addition, the supplementary material contains the source code for B/F coordinate construction, the perturbation experiments, and the fibre transport experiments, which can be run directly to reproduce our results.

REFERENCES

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2019. URL https://proceedings.mlr.press/v97/cohen19d.html.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/94b472a1842cd7c56dcb125fb2765fbd-Abstract-Conference.html.

Nico Courts and Henry Kvinge. Bundle networks: Fiber bundles, local trivializations, and a generative approach to exploring many-to-one maps. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBXzcPPOuX.

M. Daniel and C. M. Viallet. The geometrical setting of gauge theories of the Yang-Mills type. *Reviews of Modern Physics*, 52(1):175–197, 1980. doi: 10.1103/RevModPhys.52.175. URL https://doi.org/10.1103/RevModPhys.52.175.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL http://jmlr.org/papers/v7/demsar06a.html.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024. doi: 10.48550/arXiv.2412.10117. URL https://arxiv.org/abs/2412.10117.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d3221cdb27e49d9c1cd35ad254feccfe-Abstract-Conference.html.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5345–5355, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/dbab2adc8f9d078009ee3fa810bea142-Abstract.html.

Lang Gao, Jinghui Zhang, Wei Liu, Fengxian Ji, Chenxi Wang, Zirui Song, Akash Ghosh, Youssef Mohamed, Preslav Nakov, and Xiuying Chen. The cylindrical representation hypothesis for language model steering. In *Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.01844.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. PromptTTS: Controllable text-to-speech with text descriptions. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023. URL https://arxiv.org/abs/2211.12171.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6966–6981, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.346. URL https://aclanthology.org/2025.acl-long.346/.

Yunpei Li, Xun Zhou, Jinchao Wang, Lu Wang, Yong Wu, Siyi Zhou, Yiquan Zhou, Yining Wang, Yaogen Yang, Zhetao Hu, Shiyao Duan, Jiacheng Xu, Jingchen Shu, and Bin Xia. IndexTTS 2.5 technical report. *arXiv preprint arXiv:2601.03888*, 2026. doi: 10.48550/arXiv.2601.03888. URL https://arxiv.org/abs/2601.03888.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23f.html.

Peter W. Michor. *Topics in Differential Geometry*, volume 93 of *Graduate Studies in Mathematics*. American Mathematical Society, 2008. doi: 10.1090/gsm/093. URL https://www.mat.univie.ac.at/~michor/dgbook.pdf.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6338–6347, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Luis Scoccola and Jose A. Perea. FibeRed: Fiberwise dimensionality reduction of topologically complex data with vector bundles. In Erin W. Chambers and Joachim Gudmundsson (eds.), *39th International Symposium on Computational Geometry (SoCG 2023)*, volume 258 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 56:1–56:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi: 10.4230/LIPIcs.SoCG.2023.56. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.SoCG.2023.56.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report, 2026. URL https://arxiv.org/abs/2601.21337.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Siyi Wang, Shihong Tan, Siyi Liu, Hong Jia, Gongping Huang, James Bailey, and Ting Dang. CoCoEmo: Composable and controllable human-like emotional TTS via activation steering. In *Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.03420.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5180–5189. PMLR, 2018. URL https://proceedings.mlr.press/v80/wang18h.html.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. Towards controllable speech synthesis in the era of large language models: A systematic survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 764–791, Suzhou, China, November 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.40. URL https://aclanthology.org/2025.emnlp-main.40/.

Tianxin Xie, Shan Yang, Chenxing Li, Dong Yu, and Li Liu. Emosteer-tts: Fine-grained and training-free emotion-controllable text-to-speech via activation steering. *arXiv preprint arXiv:2508.03543*, 2025b.

Jingyuan Xing, Zhipeng Li, Shuaiqi Chen, Xiaofen Xing, and Xiangmin Xu. EATS-Speech: Emotion-adaptive transformation and priority synthesis for zero-shot text-to-speech. In *Interspeech 2025*, pp. 4358–4362, 2025. doi: 10.21437/Interspeech.2025-1638. URL https://www.isca-archive.org/interspeech_2025/xing25_interspeech.html.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024. doi: 10.1109/TASLP.2024.3402088. URL https://doi.org/10.1109/TASLP.2024.3402088.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, 137:1–18, 2022. doi: 10.1016/j.specom.2021.11.006. URL https://www.sciencedirect.com/science/article/pii/S0167639321001308.

Li Zhou, Hao Jiang, Junjie Li, Tianrui Wang, and Haizhou Li. Emoshift: Lightweight activation steering for enhanced emotion-aware speech synthesis. *arXiv preprint arXiv:2601.22873*, 2026.

A LIMITATIONS

First, FibreTTS assumes a continuous and smooth content space. An idealized ASR model motivates this assumption, but real models may not satisfy it. Content representations can jump or drop sharply between nearby texts. In such cases, a discrete description may fit better. A discrete content space needs its own notions of connection and transport, which we leave to future work. Second, we study only three common control interfaces, and each control signal is short and specific. Long or compound prompts, and control methods outside these three types, may not fit the same picture. Whether the theory applies to them remains open. Third, our experiments focus on language-model-based TTS. Non-autoregressive systems such as F5-TTS have different working process, so our coordinate construction is uncertain to apply to them directly. Whether FibreTTS transfers to these architectures has not been tested.

B SMOOTH BUNDLE ASSUMPTIONS

Section 2.1 states that our analysis assumes finite-dimensional smooth regions with continuous content. Section 2.2.1 then introduces the local coordinate map in Eq. (1). This appendix makes both statements precise. It lists the assumptions that make Eq. (1) a well-defined smooth bundle, defines the coordinate change between two regions, and explains why the content space B must be continuous rather than discrete.

Regular region. We first explain the basic terms used below in simple words. A finite-dimensional manifold is a space where each point can be described locally by a finite list of real coordinates. The Hausdorff condition means that any two different points have disjoint open neighborhoods. So we can always tell two points apart. Second-countability means that a countable family of open sets is enough to build every open set. These two conditions rule out strange spaces, and they let the usual tools of calculus work. A diffeomorphism is a smooth one-to-one and onto map whose inverse is also smooth. It shows that two spaces are the same up to a smooth change of coordinates.

We assume that E, B, and the typical fibre F are smooth manifolds. Each of them is finite-dimensional, Hausdorff, and second-countable. The fibre F is also nonempty. The content map is $\pi = R \circ G$. This map is smooth and onto. For each point $b \in B$, there is an open neighborhood U of b. There is also a diffeomorphism

$$\varphi_U : \pi^{-1}(U) \longrightarrow U \times F, \qquad \pi \circ \varphi_U^{-1} = \mathrm{pr}_B \,. \tag{11}$$

This is the precise form of Eq. (1). The map sends the part of E above U to the product $U \times F$. The first component of $\varphi_U(e)$ is always $\pi(e)$. So near each content value, the output space looks like a plain product of the content space and one fixed fibre. As a result, we have $\dim E = \dim B + \dim F$. The directions inside a fibre form the vertical space $V_e E = \ker d\pi_e$. These conditions are assumptions about a regular region of the idealized output space. They do not follow from the finite representation samples. Also, a smooth content map is not enough by itself to give local triviality. Singular points and changes in dimension lie outside this region. For the standard definitions of smooth bundles, we follow (Michor, 2008).

Continuous content. Section 2.2.2 uses transcription distributions as content points, so B is a continuous space. One may ask why we do not simply use each text label as a base point. There are two reasons. First, a discrete base makes local triviality empty. Each text label would have a small neighborhood with no other label inside. A product chart over this neighborhood would describe only one fibre. Local triviality would then hold automatically, and it would say nothing about how different texts relate. Second, a discrete base cannot support transport. Every continuous path into a discrete space stays at one point. So the cross-content transport in Section 4.1 would have no path to follow. For these reasons, we model content as a continuous space, and we treat a finite set of texts as a sample of points drawn from this space.

The continuous view also lets us describe content errors. Let $\widehat{x}(b)$ be the most probable transcription at b. Let $B_x = \{b : \widehat{x}(b) = x\}$ be the region where the words x are recognized. A content change that stays inside B_x keeps the recognized words. A change that crosses the boundary of B_x can change them. Appendix G uses this region.

Coordinate changes. Section 3.1 uses a transition map between two regions. Here is its definition. Let U and V be two regions with charts φ_U and φ_V. On the overlap $U \cap V$, the same realization

has two sets of coordinates. They are related by

$$\varphi_V \circ \varphi_U^{-1}(b, f) = \big(b,\, g_{VU}(b, f)\big). \tag{12}$$

The content coordinate b stays the same. Only the reading coordinate changes, through the map g_{VU}. This map is smooth, and it is invertible in f for each fixed b. When three regions overlap, the maps compose consistently. Note that g_{VU} relates two descriptions of the same base point. It does not relate readings at different base points. That relation needs the extra connection in Appendix G.1.

C Experimental Details

Sections 2.3.1 and 3.3.1 give a short description of the datasets, control methods, and metrics. This appendix gives the details of each. Appendix D covers the coordinate construction separately.

C.1 Data and Pairing

ESD dataset (Zhou et al., 2022). The public ESD dataset contains 350 parallel utterance prompts. Ten English and ten Mandarin speakers read each prompt in neutral, happy, angry, sad, and surprised styles (Zhou et al., 2022). The words stay the same across emotions. So we can vary the requested reading while keeping the content fixed. Our coordinate and generation studies use the text and reference subsets described below.

Fintech dataset. The Fintech dataset is a synthetic dataset. We build it from call-center agent data at ByteDance. This source data contains a large amount of Chinese call audio and transcripts from 35 agent voices. We first randomly sample 1,357 common telemarketing scripts. Each script serves as a target text. For each target text, we randomly pick one agent's audio and transcript as the reference. We then generate speech with two CosyVoice models. The first model is the original CosyVoice model. The second model is a CosyVoice model fine-tuned on these agent recordings and transcripts. So each target text gives one pair of recordings, and the dataset has 1,357 pairs in total. Human raters then compare the two recordings in each pair. They follow the guidelines in the box below, translated from the original Chinese.

Pairwise Rating Guidelines, Translated from Chinese

Task. Compare Audio 1 and Audio 2. Choose the one that sounds more comfortable overall and more like a real customer-service recording.

Criteria.

1. **Overall impression.** Does the audio sound natural, comfortable, and pleasant? Does it fit a real customer-service call?
2. **Voice realism.** Does the voice sound like a real human agent? Does it have an obvious AI, robotic, or over-synthesized feel?
3. **Stress naturalness.** Are stress and intonation reasonable? Are there odd emphases, flat delivery, or stiff tones?
4. **Pause naturalness.** Are pauses within and between sentences natural? Are there wrong breaks, stutters, or overly long pauses?
5. **Speaking rate and rhythm.** Is the rate suitable? Is the rhythm steady, or does it suddenly get too fast, too slow, or uneven?
6. **Audio quality.** Slight noise, a telephone-like sound, or mild compression is acceptable if the content stays clear. Serious noise, truncation, or empty audio should lose points.
7. **Extra non-text sounds.** Are there extra noises, breaths, hums, or swallowing sounds at the start, in the middle, or at the end?

Scale.

$+2$ Audio 1 is clearly better	-1 Audio 2 is slightly better
$+1$ Audio 1 is slightly better	-2 Audio 2 is clearly better
0 The two are about the same	

Notes. Raters should report any empty or unplayable audio, truncation, severe noise, clear misreading, missing words, or extra non-text sounds.

Nine human annotators take part in the rating. Each pair goes to annotators at random. Each pair then receives a score for every criterion in the box. The rated recordings and their scores serve

two later uses. First, they help us build the grid used in the main text. We randomly sample 98 generated recordings along with their texts. These samples serve as references, and they guide the grid construction. We also randomly sample 50 other telemarketing texts from the transcripts. These texts differ from the reference texts, and they serve as target texts. Second, the score differences help us build contrastive data pairs to construct steering vectors.

Transport samples. The transport experiment in Section 4.3 uses the following samples. For each model and each dataset, we select 50 different target texts and 10 different reference recordings. We then test every text–reference combination under each control interface. In ESD, the target texts include 25 Chinese texts and 25 English texts. The references include five recordings for each language. Among these five recordings, two are angry, one is happy, one is sad, and one is surprised. In Fintech, all 50 target texts are in Chinese. The ten references cover six attributes. Pauses, emphasis, noise, and audio quality each have two references. Speaking rate and timbre each have one reference. So each dataset gives 500 pairs for each interface. With three interfaces, each dataset gives 1,500 pairs in total. Within each model's experiment, all interfaces use the same reference set. The requested attribute and the instruction language both follow the reference condition. ESD also includes every cross-language combination. This gives 250 cross-language pairs for each interface. All these inputs come from the construction-grid inventory. So this experiment tests interventions inside the fitted representation.

C.2 Control Methods

Reference control. On ESD, the reference recording carries the requested emotion. On Fintech, we use the preferred recording in a pair for the requested attribute. CosyVoice2 uses its zero-shot interface. It takes the target text, the reference audio, and the reference transcript. IndexTTS takes the target text. It uses the selected recording for both speaker and emotion conditioning, and we set the emotion strength to 1.0. So the reference may also carry speaker and acoustic traits.

Prompt control. CosyVoice2 uses its instruction interface. The instruction is followed by the required end-of-prompt marker. IndexTTS passes the instruction through its text-to-emotion interface. Both models keep a reference recording for audio conditioning. On ESD, the instructions have the form "Please say it in an *angry* tone." We use the same form for a *happy*, *cheerful*, *sad*, or *surprised* tone. Chinese reference conditions use the Chinese versions. On Fintech, we use the Chinese instructions translated in Table 2. In all cases, the request is the same attribute, and each model receives it through its own interface.

Steering-vector construction. For each requested attribute, we build a steering vector from a mean difference in activations Wang et al. (2026). Each calibration pair has a positive reference and a comparison reference. On ESD, the two references are recordings of the same utterance by the same speaker. The positive one is emotional, and the comparison one is neutral. On Fintech, the two references are the preferred and less preferred recordings in a preference pair for that attribute. Within each pair, we keep the target text fixed. We then run the model once with each reference. Each time, we record the activation at the last position before the first speech token. We call these two activations u_i^+ and u_i^-. The direction for attribute a is

$$v_a = \frac{1}{N_a} \sum_{i=1}^{N_a} (u_i^+ - u_i^-), \tag{13}$$

where N_a is the number of calibration pairs for attribute a. We estimate a separate direction for each model and each attribute. On ESD, we also estimate separate directions for Chinese and English. These vectors form the steering baseline. The B/F bases are built in a different way. They come from the variation across the text–reference grid.

Steering application. At each decoding step, we add the attribute direction to the current representation: $u'_t = u_t + \alpha v_a$. We set the strength to $\alpha = 3.0$. For a given model and attribute, all paired comparisons use the same direction and the same strength.

C.3 Evaluation

Instruction Following (IF). The IF judge receives one generated recording and a natural-language instruction. The instruction describes the requested attribute. The judge returns three integer scores

Table 2: Fintech instructions, translated from Chinese.

Attribute	Instruction
Pauses	Make the pauses more natural and fluent.
Emphasis	Make sentence stress more natural.
Speaking rate	Make the speaking rate and rhythm more natural.
Noise	Avoid words or speech fragments absent from the text.
Timbre	Make the voice sound more natural and realistic.
Audio quality	Improve overall audio quality and reduce noise.

from 1 to 5, and we report only instruction following, from 1 for poor fulfillment to 5 for strong fulfillment. Within one comparison, the reference, prompt, and steering conditions use the same attribute wording. We use Gemini-3.1-Pro (Team, 2025) as the judge, with the prompt shown below.

Prompt for the Instruction-Following Judge

Input. Each request includes the generated audio as an attachment, along with the text prompt below.

System prompt.

```
You are a strict evaluator for synthesized speech.  Given an audio
file, its transcript, and an instruction describing the intended
delivery, rate the audio.  Output ONLY a JSON dict with integer keys
instruction_following, naturalness, expressiveness, each 1-5.
1 Instruction Following:  1=ignores instruction; 3=partial;
5=perfectly follows.
2 Naturalness:  1=robotic; 5=human-like.
3 Expressiveness:  1=flat; 5=richly expressive.
```

User prompt.

```
The synthesized speech is attached.  Transcript:  "{t}".  Instruction:
"{i}".
```

Here, `{t}` is the target text, and `{i}` is one of the instructions listed below.

Instructions for ESD.

- `Please say it in an angry tone.`
- `Please say it in a happy, cheerful tone.`
- `Please say it in a sad tone.`
- `Please say it in a surprised tone.`

Instructions for Fintech.

- `Make the pauses more natural and smooth.`
- `Make the stress in the sentence more natural.`
- `Make the speaking rate and rhythm more natural.`
- `Avoid extra non-text sounds.`
- `Improve the overall audio quality and reduce noise.`
- `Make the timbre sound more natural and real.`

Note. For Chinese samples, we use the Chinese version of each instruction. This covers the Chinese subset of ESD and all of Fintech.

Character Error Rate (CER). To measure CER, we first use Qwen3-ASR (Shi et al., 2026) to transcribe the generated audio. We then remove common punctuation from both the transcription and the target text. For Chinese text, we also remove spaces. After that, we compute $\mathrm{CER} = (S + D + I)/N$. Here S, D, and I are the numbers of character substitutions, deletions, and insertions. *For Chinese, N is the number of characters in the target text; For English, N is the number of words*. For simplicity, we call both rates CER. CER is never negative, and a value of 0 means an exact match. CER can go above 100% when there are many insertions, typically meaning

Table 3: Text–reference grids used for coordinate construction.

Model	Dataset	Texts	References	Pairs
CosyVoice2	ESD	100	200	20,000
CosyVoice2	Fintech	50	98	4,900
IndexTTS	ESD	100	100	10,000
IndexTTS	Fintech	50	98	4,900

that the audio has broken. We compute CER for each utterance, and the reported value in the main text is the mean over utterances.

For IF and CER, we first compute the change within each matched generation pair. We then average these changes. The transport comparisons measure IF only. We check content preservation in the coordinate-intervention experiments instead. So these results do not show that transport preserves the spoken content of the target text.

D Coordinate Construction and Interpretation

Section 2.3.2 builds the B/F coordinates from a text–reference grid. Section 2.3 then reports their quality in Figure 2. This appendix gives the construction details, defines the explained-variance and reconstruction measures behind Figure 2, and shows that the coordinates together with a residual function form a local product chart of the kind assumed in Section 2.2.

D.1 Construction Details

Crossed input grids. We build a separate coordinate system for each model–dataset combination. Each grid contains every combination of the selected target texts and reference recordings. Each pair gives one representation, and every pair has equal weight. So every text is averaged over the same reference set, and every reference is averaged over the same text set. Table 3 lists the grids. These grids are larger than the subsets used for generation interventions.

The ESD reference set includes emotional and neutral recordings. The Fintech reference set includes both preferred and less-preferred recordings. These labels do not enter the decomposition. We also do not select samples by instruction-following scores. The reference factor can include speaker identity, timbre, and prosody. So it is not limited to emotion.

Representation extraction. For each text–reference pair, we take the raw logits right before the first speech token. They have $6{,}564$ dimensions for CosyVoice2 and $8{,}194$ for IndexTTS. We use this point because all pairs share it, and it captures the joint response to text and reference before the generated histories diverge. CosyVoice2 uses reference conditioning. IndexTTS uses its default reference-conditioned setting. In both models, the reference can affect the representation even without a prompt or a steering signal.

Main effects and compact bases. We call a grid complete and equally weighted when two conditions hold. First, every text–reference combination is present, so the grid has no missing cells. Second, every cell counts equally, so we apply no sample weighting. On such a grid, the centered row means and column means in Eq. (3) give the least-squares additive fit. The interaction term has a zero mean along each row and each column, so it holds the variation that remains after we remove both average input effects.

We keep the top 20 dimensions of the text effects as D_B. We also keep the top 20 dimensions of the reference effects as D_F. The marginal coordinates are then $b_t = D_B^\top u_t$ and $f_r = D_F^\top v_r$. So the residual of the reduced additive fit is

$$\varepsilon_{t,r} = I_{t,r} + (u_t - D_B b_t) + (v_r - D_F f_r).$$

This residual contains three parts: the interaction, the dropped part of the text effects, and the dropped part of the reference effects. On the balanced grid, each text is paired with each reference exactly once. As a result, the centered text effects and reference effects have zero cross-covariance as input factors. However, this does not make their spans in logit space exactly orthogonal. Some overlap may remain, and the joint readout below handles it.

Joint coordinate readout. The coordinates b_t and f_r above describe average effects over the grid. We also need coordinates for a single representation h. Section 4.2 uses this readout. To get the coordinates, we join the two bases into one matrix $D = [D_B\ D_F]$. We then look for the coefficients $\widetilde{b}$ and $\widetilde{f}$ that make $D_B\widetilde{b} + D_F\widetilde{f}$ as close as possible to $h - \mu$:

$$\begin{bmatrix}\widetilde{b}\\ \widetilde{f}\end{bmatrix} = \arg\min_{b,\,f} \left\|h - \mu - D_B b - D_F f\right\|^2 = D^\dagger(h - \mu). \tag{14}$$

This is a least-squares problem, and $D^\dagger$ is the Moore–Penrose pseudoinverse of D. The pseudoinverse turns the problem into one matrix product. We multiply $h-\mu$ by $D^\dagger$, and the result gives $\widetilde{b}$ and $\widetilde{f}$ stacked in one vector. When the columns of D are linearly independent, the pseudoinverse has a closed form, $D^\dagger = (D^\top D)^{-1}D^\top$. In this case, the solution is unique. When some columns are redundant, $D^\top D$ cannot be inverted, and many solutions fit equally well. We then compute $D^\dagger$ from the singular value decomposition $D = U\Sigma V^\top$ as $D^\dagger = V\Sigma^+U^\top$. Here Σ^+ inverts the nonzero singular values and leaves the zero ones at zero. This choice gives the solution with the smallest norm. After the readout, we define the leftover part as

$$r = h - \mu - D_B\widetilde{b} - D_F\widetilde{f}. \tag{15}$$

This term is the part of h that the two bases cannot explain. We read both coordinate sets together because the bases may overlap. If we projected onto each basis separately, the shared part could be counted twice in the two readouts.

The projection residual r differs from $I_{t,r}$ and $\varepsilon_{t,r}$. Its coordinates are fitted to one representation. The other two come from grid averages. In this least-squares readout, $D^\top r = 0$ holds by construction. This orthogonality is a property of the readout, not evidence for a fibre-bundle structure.

Coordinates for generation interventions. The interventions in Section 3.3 change the current representation by

$$h' = h + D_B\Delta b + D_F\Delta f. \tag{16}$$

When D has linearly independent columns, the joint readout of h' gives $(\widetilde{b} + \Delta b, \widetilde{f} + \Delta f)$. The residual r stays the same at that point. So the intervention has a unique description inside the fitted span. What happens afterwards is measured on the generated audio.

D.2 Variance and Reconstruction Measures

Variance and reconstruction. We first define the total variance of the grid as $V = \sum_{t,r} \|h_{t,r} - \mu\|_2^2$. It adds up the squared distance from each representation to the mean. Let $\widehat{h}_{t,r}$ be a reconstruction of $h_{t,r}$. We measure its error as $\sum_{t,r} \|h_{t,r} - \widehat{h}_{t,r}\|_2^2$. We divide this error by V and report the result as the reconstruction loss. This loss is the fraction of variance that the reconstruction misses. The explained-variance score is one minus this loss, so a score of 1 means a perfect reconstruction. We also measure how much variance each basis explains. The share of B is $|\mathcal{R}| \sum_t \|D_B b_t\|_2^2 / V$. The factor $|\mathcal{R}|$ appears because each text effect repeats once for every reference. In the same way, the share of F is $|\mathcal{T}| \sum_r \|D_F f_r\|_2^2 / V$. On the balanced grid, the cross term between B and F sums to zero:

$$\textstyle\sum_{t,r}(D_B b_t)^\top (D_F f_r) = \left(D_B \sum_t b_t\right)^\top \left(D_F \sum_r f_r\right) = 0. \tag{17}$$

Both inner sums are zero because the coordinates are centered. So the variance of the combined fit $D_B b_t + D_F f_r$ equals the sum of the two shares. This result holds even when the two subspaces overlap. All these quantities are based on the additive-fit residual $\varepsilon_{t,r}$. They differ from the residual r in Eq. (15), which is computed for a single representation.

Local residual reconstruction. Figure 2 contains two comparisons. The left panels test whether nearby texts can estimate the residual of a given text. The right panels test whether one fixed text can pass its residual on to other texts.

In the left panels, we treat each text t as a receiver. We measure the distance between two texts by their B coordinates b_t. For each receiver, we find its five nearest other texts and call this set $\mathcal{N}_5(t)$.

The receiver itself is not in this set. At each reference r, we average the residuals of these five texts. We then add this average to the additive reconstruction:

$$\widehat{h}_{t,r} = \mu + D_B b_t + D_F f_r + \frac{1}{5} \sum_{t' \in \mathcal{N}_5(t)} \varepsilon_{t',r}.$$

For each receiver, we average the loss over references. We normalize it by the same complete-grid variance as above. The left panels show the distribution of these explained-variance scores across texts.

In the right panels, we pick one central text s as the source. The source gives its residual at each reference to every other text:

$$\widehat{h}_{t,r} = \mu + D_B b_t + D_F f_r + \varepsilon_{s,r}.$$

For each receiver, we again average the normalized loss over references. Each point plots this loss against the distance between the receiver and s in B space.

D.3 A Local Coordinate Model

Section 2.2 assumes that a local region of the output space is a product of content and reading coordinates. The B/F coordinates only span a linear subspace, and the residual carries the rest. So one may ask whether the two together still fit this product picture. This part gives a positive answer, under one condition on the residual. Section 4.2 uses a predictor $q(b, f)$ for the residual. The construction below explains where this predictor comes from.

Let $D = [D_B \ D_F]$ have linearly independent columns, and let $W = D^\dagger$ be the readout map. Suppose the residual is a smooth function $q(b, f)$ of the coordinates, with values in the kernel of W. Then we can define

$$\widetilde{\psi}(b, f) = \mu + D_B b + D_F f + q(b, f). \tag{18}$$

On an open product domain, this map is a smooth embedding. Applying W to $\widetilde{\psi}(b, f)$ gives back (b, f), because W removes the residual term. The derivative of $\widetilde{\psi}$ has the same left inverse. So the image of $\widetilde{\psi}$ has a product description, even when the residual depends on content. This is a sufficient construction. It does not identify the speech bundle from data. Finite readouts alone cannot show that the residual is a smooth, single-valued function over a region. Also, the fitted content readout on logits need not equal the content projection on complete speech. The generation interventions in Section 3.3 test how the two are related in behavior.

E Coordinate Interventions

Section 3.3 tests Claim 2 with two generation interventions: direct F reuse and B perturbations. Figure 3(b–d) shows the main results. This appendix gives the settings of both interventions, the full result table, and the way we compute the intervals reported in the main text.

Direct F reuse. In this experiment, we copy the F coordinate from one sample to another. We call the sample that gives its coordinate the donor. We call the sample that receives it the receiver. Figure 3(c,d) uses three emotions from the Chinese part of ESD: anger, sadness, and happiness. All samples use prompt control. For each model and each emotion, we cross 50 texts with 10 references. This gives 500 receiver comparisons. Each donor has the same emotion and the same control interface as its receiver. However, the donor has a different target text, and its reference may also differ. At each speech step, we replace the receiver's F coordinate with the donor's:

$$h' = h + D_F\big(f_{\text{donor}} - f_{\text{receiver}}\big).$$

Here f_{receiver} is the receiver's current F readout at that step. The update moves h only along the F basis. So the receiver keeps its current B readout and its projection residual at that step. We measure IF against the attribute that the receiver requests. We measure CER against the receiver's target text. Table 4 reports all available conditions. The figure shows only the prompt subset, since that subset has no missing condition.

B perturbations. Figure 3(b) uses Chinese anger samples from ESD, with reference control in both models. Figure 5 adds Fintech results with the pause reference. For each model and dataset, we draw

Table 4: Paired mean changes after full F replacement across texts. Each condition has 500 receiver pairs. All nominal 95% IF intervals lie below zero and all CER intervals above zero. CER changes are reported in percentage points rather than relative changes.

		CosyVoice2		IndexTTS	
Emotion	Control	ΔIF	ΔCER	ΔIF	ΔCER
Angry	Reference	−1.53	+73.22	−1.67	+63.51
Angry	Prompt	−0.53	+74.19	−1.56	+70.56
Angry	Steering	−0.17	+24.49	−0.84	+77.61
Sad	Reference	−1.47	+76.66	−2.26	+77.39
Sad	Prompt	−0.70	+73.01	−2.76	+106.19
Happy	Reference	−2.69	+77.32	−3.06	+87.87
Happy	Prompt	−1.78	+71.25	−3.25	+71.77

one random unit direction u in the 20-dimensional B coordinates and keep it fixed. All receivers and strengths share this direction. At each decoding step, we set $h' = h + D_B(su)$, where s is the strength. We test 16 strengths on the same 50 receivers, namely 10 texts crossed with 5 references. To show the size of each change, we plot the ratio

$$\rho_i(s) = \text{average over steps } \tau \text{ of } \frac{\|D_B(su)\|_2}{\|h_{i,\tau}^{(s)} - \mu\|_2}. \tag{19}$$

Here $h_{i,\tau}^{(s)}$ is the logit state at step τ in the run at strength s, just before the change is added. The denominator is the whole centered state, not only its B part. So ρ compares lengths, not signal and noise power as in a signal-to-noise ratio.

In both datasets, CER grows as ρ increases. On ESD, the CER of IndexTTS rises from 3.83% to 24.96% at $\rho \approx 1.0$. The CER of CosyVoice2 rises from 2.23% to 70.43% already near $\rho = 0.53$. On Fintech, both models start below 0.5% CER and exceed 90% at $\rho \approx 1.1$. The two models fail in different ways. CosyVoice2 often produces audio that ASR transcribes as empty, which counts as 100% CER. This happens in about half of its recordings, mostly at large strengths. IndexTTS never produces such audio. Its CER can go above 100% instead, which means it inserts extra content. These results only apply to the tested direction in each model.

Intervals for paired changes. Section 3.3 reports that the nominal intervals of the paired changes exclude zero. We compute these intervals from 2,000 bootstrap resamples. Each resample draws texts and references with replacement and keeps their crossed structure. The intervals are nominal, without correction for multiple comparisons or shared donors. The 500 pairs come from a repeated text–reference grid, so they are not 500 independent samples. Also, the direct reuse here and the transport experiment in Section 4.3 use different generation procedures and donor assignments. So the two experiments are not a matched comparison of copying and adapted transport.

F Content Dependence of Reference Responses

Section 3.3 and Figure 3(a) report that residual reference responses become less aligned as B distance grows. Section 2.3 also states that the residual is partly predictable at fixed content. This appendix supports both statements. It defines the response and the distance behind Figure 3(a), reports the same analysis for both models, and gives the prediction results.

Response and distance. Figure 3(a) uses CosyVoice2 on both construction grids. We start from the rank-20 additive-fit residual $\varepsilon_{t,r}$ in Eq. (4). For a pair of references (r_1, r_2), we define the response at text t as $d_t = \varepsilon_{t,r_1} - \varepsilon_{t,r_2}$. This is the part of the reference change that the additive coordinates do not capture. It is not the full reference effect, and it is not the projection residual r used for transport. For two different texts t and s, we then compare the cosine of d_t and d_s with the distance $\|b_t - b_s\|_2$ between their B coordinates.

Each dataset contributes 50,000 sampled comparisons. The curves in Figure 3(a) average the cosines within successive distance bands. The envelopes show the middle half of the observations in each band. The nearest and farthest fifths have mean cosines of $0.448/0.007$ on ESD and $0.602/-0.057$ on Fintech. The comparisons share texts and references, so they are not independent.

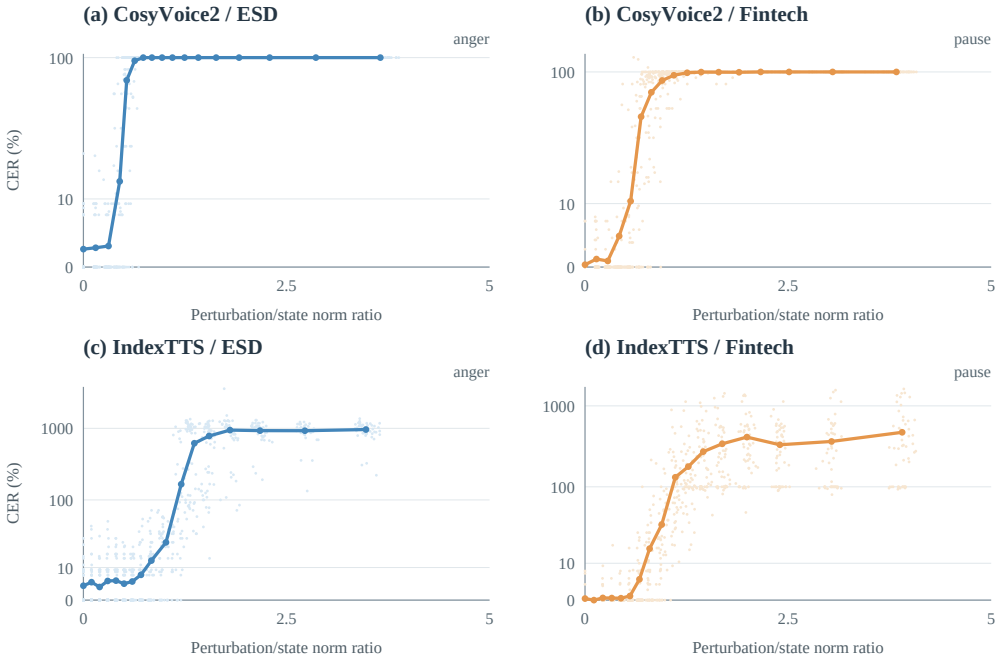


Figure 5: CER across the full B-perturbation sweep. Blue: ESD anger references; orange: Fintech pause references. Faint points are single text–reference pairs, and lines show means. CosyVoice2 levels off at 100% due to empty transcriptions. IndexTTS exceeds 100% due to insertions.

Figure 6 repeats this analysis for both models and both datasets. For IndexTTS, the nearest and farthest fifths have mean cosines of $0.810/-0.029$ on ESD and $0.568/0.021$ on Fintech. Each curve averages 50,000 sampled comparisons in 20 distance bands. Distances are measured in each model's own B coordinates. So their scales are only comparable within a panel.

Interaction structure. Section 2.3 includes a locally estimated residual correction in the reading coordinate. This part explains why. The full interaction $I_{t,r}$ has zero mean across texts for every reference, and zero mean across references for every text. A simple local description is $I(b, f) \approx M(b)f$. At fixed content, the reference-related displacement is partly a linear function of F. The response map $M(b)$ changes with content. Figure 8 illustrates this picture. We test it on CosyVoice2 logits, and Figure 7 shows the results. A linear F model at fixed text predicts held-out references with mean $R^2 = 0.416$ on ESD and 0.245 on Fintech. A model with B/F product terms predicts held-out text rows with $R^2 = 0.447$ and 0.517. When we break the link between the coordinates and the responses, all four R^2 values become negative. So the reference response is regular at fixed content, and it depends on content.

G Relative Reading Transport

Section 4.1 defines transport through a chosen connection in Eq. (7). It then asks the connection to preserve the relative reading in Eq. (8). This appendix explains what a connection is and why the bundle alone does not fix one, builds the connection that preserves the relative reading, states the task-compatibility condition used in Claim 3, and clarifies what the correction in Eq. (9) does and does not guarantee about the generated speech.

G.1 Paths and Connections

A path γ in B joins the source content to the target content. It stays inside one connected piece of B, and it may have corners. This path describes a change of content. It does not describe the passage of audio time. To move a reading along this path, we need a rule. The rule says which directions in

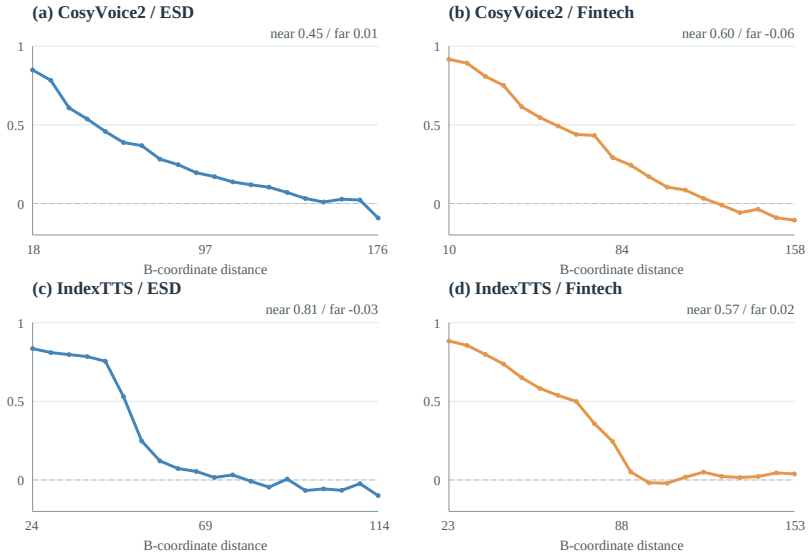


Figure 6: Residual reference-response alignment across content distance for both models and datasets. Blue panels use ESD and orange panels use Fintech. The plotted quantity is the mean cosine between residual responses to the same reference change at two texts; smaller and larger B distances are shown from left to right. Near/far annotations summarize the nearest and farthest fifths of comparisons.

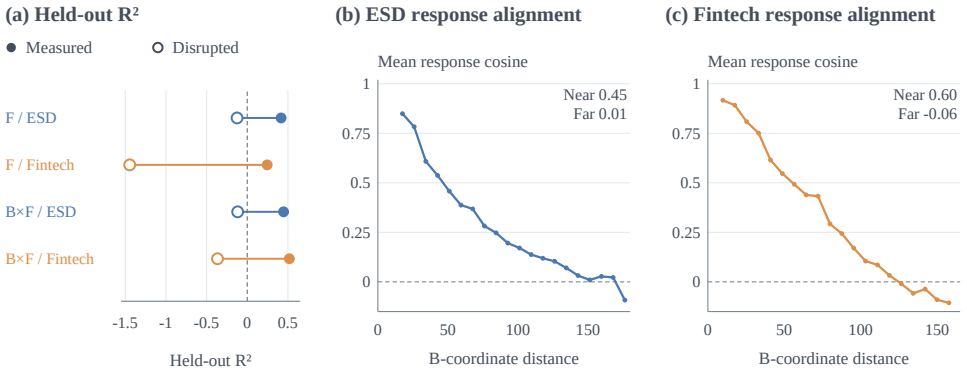


Figure 7: Text–reference interaction in CosyVoice2 logits. (a) Held-out R^2 for measured and disrupted B coordinates. (b,c) Response-direction alignment decreases with B-coordinate distance on ESD and Fintech; Near and Far denote the closest and farthest 20% of comparisons.

E count as "changing content only." An Ehresmann connection gives this rule. At each point e, it picks a horizontal space H_eE that complements the fibre directions:

$$T_eE = \ker d\pi_e \oplus H_eE. \tag{20}$$

The map $d\pi_e$ sends H_eE onto the tangent space of B in a one-to-one way. So each velocity $\dot{\gamma}(t)$ of the base path has exactly one horizontal velocity above it. Following these horizontal velocities from the source reading gives the lifted path in Eq. (7). For a smooth connection, the lift is unique near each point. We assume that it exists along the whole path. Two points need care. Different connections give different endpoints, and one connection can give different endpoints along different paths. Also, local triviality alone does not pick a connection; the task and data must.

G.2 Relative Coordinates

We now build the connection used in Eq. (8). Suppose that a region over $U \subset B$ has a smooth invertible map $\mathcal{R}(e) = (\pi(e), \eta(e))$ onto $U \times \Omega$. Here $\Omega \subset \mathbb{R}^{\dim F}$ is open and contains 0. The map

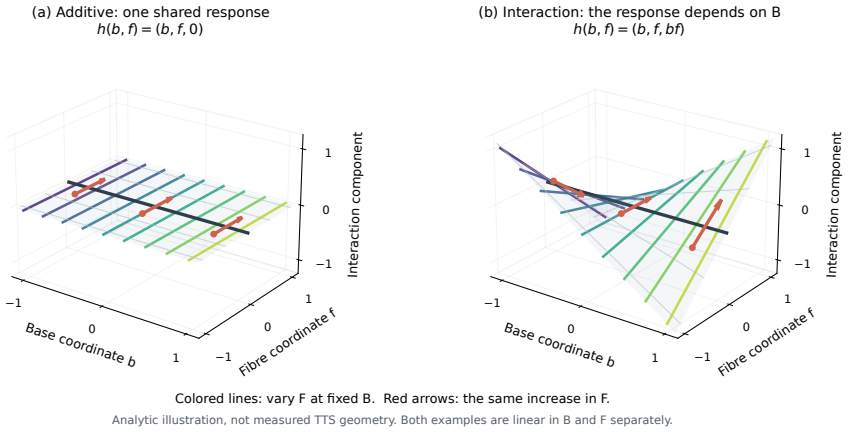


Figure 8: Analytic illustration of a content-dependent reading response. Each colored curve varies the reading coordinate at fixed content; arrows represent equal reading-coordinate changes. An interaction such as bf makes the represented displacement depend on content even when each fixed-content response is locally regular.

η gives a reading coordinate relative to a baseline. The baseline at content b is $\sigma(b) = \mathcal{R}^{-1}(b, 0)$, the reading with $\eta = 0$. The key property is that η has the same meaning at every content in U. This shared meaning is extra information from the task. The bundle structure does not provide it. With this map, we define the connection by keeping η fixed:

$$H_e E = \ker d\eta_e, \qquad P_\gamma(e_s) = \mathcal{R}^{-1}\big(\gamma(1), \eta(e_s)\big). \tag{21}$$

This is a valid connection because $d\mathcal{R} = (d\pi, d\eta)$ is invertible. The fibre directions keep π fixed, and the horizontal directions keep η fixed. The two sets of directions complement each other. The lifted path is $e(t) = \mathcal{R}^{-1}(\gamma(t), \eta(e_s))$. It exists along any path in U, and it keeps η constant. So it carries the source baseline to the target baseline, and it keeps the relative reading. Eq. (8) uses reading coordinates that agree with η.

The relative reading stays fixed, but the original fibre coordinate f can still change. As an example, let $f^0(b)$ be a smooth baseline. Choose an invertible matrix $L(b)$ at each content and set $\eta(b, f) = L(b)(f - f^0(b))$. Then transport gives

$$f_t = f^0(b_t) + L(b_t)^{-1} L(b_s)\big(f_s - f^0(b_s)\big). \tag{22}$$

The adjustment to f depends on the content through L and f^0. This construction needs the shared map $\mathcal{R}$. It works within one region U, and it does not need a global trivialization. Relative coordinates chosen separately in different regions need not combine into one consistent connection.

G.3 Task Compatibility

Claim 3 requires a task-compatible coordinate. This subsection gives a precise sufficient condition. Let $\mathcal{E}_c(b)$ be the set of readings at content b that fulfill request c. Suppose there is one fixed set $\mathcal{S}_c \subseteq \Omega$ such that $\mathcal{E}_c(b) = \eta_b^{-1}(\mathcal{S}_c)$ for every b in the region. In words, whether a reading fulfills the request depends only on its relative coordinate η. Transport keeps η fixed. So it sends every acceptable source reading to an acceptable target reading. If the target content lies in B_x from Appendix B, the endpoint also has transcription x. A coordinate that is preserved but does not meet this condition gives no such guarantee.

G.4 Scope of the Correction

Eq. (5) splits a control displacement into a content part and a reading part. This split uses the chosen product chart and holds to first order. A nonzero content part does not always change the recognized words, because it can stay inside B_x. The reverse direction also has a gap: fixing the estimated B readout in the logits does not guarantee that the decoded audio stays in the target content region.

Eq. (9) moves the output toward the transported reading and scales the coordinate gap by $1 - k$, which holds only if the whole interpolated segment stays inside the chart. It is a statement about

Table 5: CER (%) before and after transport, classified by control method and attribute. The layout, including the conditions without a valid donor, follows Table 1; values cover all receivers of each condition. Decreases of at least 0.5 points appear in green and all other changes in gray; the last column averages over the available conditions of each row.

		ESD				Fintech						All
Control		Anger	Happiness	Sadness	Surprise	Pauses	Emphasis	Speaking rate	Noise	Timbre	Audio quality	**Average**
CosyVoice2												
Reference	Before	2.80	3.08	2.89	3.29	2.47	2.09	–	1.86	1.13	1.90	**2.50**
	After	2.73	4.25	3.15	3.18	2.06	1.81	–	2.04	1.13	1.74	**2.55**
		(-0.06)	(+1.18)	(+0.25)	(-0.11)	(-0.41)	(-0.28)		(+0.18)	(+0.00)	(-0.16)	(+0.06)
Prompt	Before	3.27	2.64	2.89	2.76	1.64	–	1.63	2.17	0.84	3.30	**2.58**
	After	2.96	3.08	2.77	2.66	2.01	–	1.65	1.50	1.97	2.67	**2.49**
		(-0.31)	(+0.44)	(-0.12)	(-0.11)	(+0.37)		(+0.02)	(-0.67)	(+1.13)	(-0.63)	(-0.08)
Steering	Before	63.65	39.80	26.90	61.87	10.20	–	–	–	8.47	6.11	**36.85**
	After	63.82	39.12	28.06	60.97	10.45	–	–	–	8.96	4.74	**36.73**
		(+0.17)	(-0.68)	(+1.16)	(-0.90)	(+0.25)				(+0.49)	(-1.37)	(-0.13)
IndexTTS												
Reference	Before	2.97	3.16	3.15	2.97	1.14	1.08	0.99	0.72	1.10	1.22	**2.04**
	After	3.01	3.35	3.40	2.87	1.04	1.10	1.16	0.74	0.79	1.17	**2.07**
		(+0.04)	(+0.19)	(+0.26)	(-0.10)	(-0.10)	(+0.02)	(+0.17)	(+0.02)	(-0.31)	(-0.05)	(+0.02)
Prompt	Before	3.13	3.36	3.60	2.75	0.92	0.85	0.91	0.96	0.87	0.79	**2.04**
	After	3.05	2.92	2.86	3.00	0.98	0.96	0.62	0.96	0.74	0.72	**1.92**
		(-0.08)	(-0.44)	(-0.75)	(+0.25)	(+0.06)	(+0.11)	(-0.29)	(+0.00)	(-0.13)	(-0.07)	(-0.12)
Steering	Before	3.00	3.25	3.42	2.83	1.13	–	0.91	0.87	0.89	0.93	**2.15**
	After	2.93	3.62	3.70	3.15	0.93	–	0.91	0.98	0.79	1.13	**2.25**
		(-0.06)	(+0.36)	(+0.28)	(+0.32)	(-0.20)		(+0.00)	(+0.11)	(-0.10)	(+0.20)	(+0.10)

coordinate error, not a promise of monotone improvement in IF or CER. The representation-level update in Section 4.2 estimates this correspondence from model states. Whether it improves the generated speech is an empirical question, which Section 4.3 tests.

H Transport Analysis Details

Section 4.2 defines the residual update, and Section 4.3 evaluates it; this appendix gives the residual predictor, the donor–receiver pairing, and the CER results.

Residual predictor. The update in Eq. (10) needs a predictor $q(b, f)$ for the projection residual r in Eq. (15). Appendix D.3 explains why such a predictor is natural. We fit one predictor per model and dataset, with the B/F coordinates and their pairwise products as inputs and the residual as output, using regularized least squares on the crossed text–reference states without any IF labels. At test time, the update adds the donor's deviation from its predicted residual to the receiver's predicted residual and scales the change by $k = 0.5$. The change lies outside the B/F span, so the receiver keeps its measured B/F coordinates at the intervention point. This update is an empirical stand-in for the relative-reading transport in Appendix G.2.

Data and pairing. Each model has 50 target texts and ten references per dataset and control interface, as listed in Appendix C. The baseline and treated outputs of a receiver share the same request, reference, and text. A donor shares the requested attribute and control method with its receiver but has a different target text. Receivers are the samples with original IF 1–3 for which an IF-5 donor exists. This gives the 1,216 CosyVoice2 pairs and 661 IndexTTS pairs in Section 4.3.

CER before and after transport. Table 5 reports CER in the layout of Table 1 over all receivers of each condition. Nearly every cell changes by less than one point, and the averages over both datasets by less than half a point. The largest changes stay below 1.5 points and go both ways. The high CER under steering on CosyVoice2 comes from the steering baseline itself, and transport leaves it essentially where it is.