Orthogonal Style Injection: Mechanistic Persona Steering for Zero-Shot Cross-Lingual Subtitle Translation

Anonymous ACL submission

Abstract

In conversational and cinematic subtitle translation, large language models often exhibit stylistic collapse, defaulting to neutral registers that flatten character-specific personas. Prompt-based attempts to recover style often lead to instruction drift and semantic degradation. We propose Orthogonal Style Injection (OSI), a parameter-efficient, inference-time intervention that removes global stylistic bias by projecting onto an intermediate layer and injecting a character-specific style vector derived solely from English source dialogue. We evaluate OSI for English-to-Bengali, Hindi, Telugu, Tamil, Kannada, Chinese, Japanese, Spanish, French, German, and Italian subtitle translation, spanning 11 typologically diverse target languages. Results show that zero-shot stylistic transfer depends strongly on pretraining alignment. Llama-3.1-8B yields gains in several European languages but degrades on Indic languages, whereas Qwen-2.5-7B transfers style more robustly across Indic and East Asian languages. To probe sociolinguistic mechanisms, we curate a parallel English-Bengali cinematic dataset as a focused morphological stress test. Automatic metrics and human evaluation show that OSI enables meaningful honorific and register shifts while largely preserving semantic fidelity, revealing a dataset-dependent trade-off between stylistic fidelity and translation quality in highly colloquial, morphologically constrained settings. The dataset and related resources are publicly available at https://anonymous.4open.science/r/Orthogonal_Style_Injection-B027

1 Introduction

In James Cameron's Avatar, Jake Sully's aggressive warning "I'm gonna crush you!" defines his character through tone as much as through words. Yet when processed by standard machine translation systems, this line emerges as a neutral, polite apology - semantically correct, but stylistically dead. This phenomenon, which we term "stylistic collapse", is pervasive in cinematic subtitle translation: the systematic erasure of speaker persona that reduces heroes, villains, and lovers to a flat, generic voice. In film localization, where character identity is built through dialect, register, and formality as much as through semantic content (Zhang et al., 2018), this collapse destroys the emotional arc of entire scenes despite advances in cross-lingual fluency (Touvron et al., 2023).

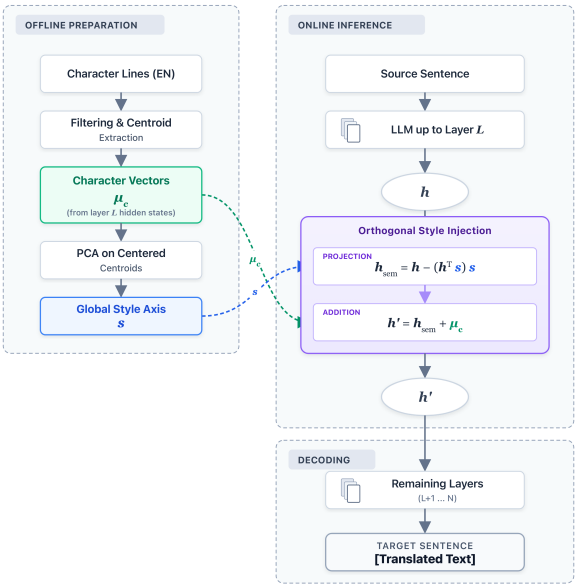

Figure 1: OSI architecture for representation-level persona steering

The challenge of subtitle translation is to preserve dramatic impact across languages under strict timing and character limits (Lison and Tiedemann, 2016). Conventional techniques for stylistic regulation, such as few-shot prompting and Supervised Fine-Tuning (SFT), face substantial challenges in scalability and fidelity. Existing approaches fail: prompting causes instruction drift (Brown et al., 2020); few-shot prompting copies surface forms rather than generalizing style; fine-tuning requires parallel style-annotated corpora that do not exist for most language

Rank	Language	Family	Total (M)
1	English	Indo-European	1515
2	Mandarin Chinese	Sino-Tibetan	1140
3	Hindi	Indo-Aryan	611
4	Spanish	Indo-European	561
6	French	Indo-European	333
7	**Bengali**	**Indo-Aryan**	**274**
12	German	Indo-European	133
13	Japanese	Japonic	125
18	Telugu	Dravidian	95

Note: Tamil (86M), Italian (65M), and Kannada (58M) are also evaluated but rank outside the global top 20 (24th, 32nd, and 34th respectively).

Table 1: Languages used in our evaluation with speaker statistics (millions) from Eberhard et al. (2025). Bengali, our primary testbed, is highlighted.

pairs (Sanh et al., 2021; Howard and Ruder, 2018). These failures are most acute in morphologically rich languages where honorifics and verb inflections encode social relationships (Patil et al., 2022; Voita et al., 2019).

Language selection: We evaluate OSI across 11 typologically diverse languages (Table 1). Bengali serves as our primary testbed: with 274M speakers, it ranks among the world's top 10 most spoken languages, yet remains a mid-resource language, lacking large parallel corpora. Its complex honorific system (intimate/familiar/formal distinctions) and verb morphology provide an ideal stress test for fine-grained stylistic transfer. Crucially, Bengali is one of the few mid-resource languages for which we have curated reference translations from cinematic subtitles, enabling rigorous quantitative evaluation. Evaluating a mid-resource language provides a fairer test of zero-shot capability than relying solely on high-resource languages where models have extensive pre-training exposure. The remaining languages cover high-resource European languages (Spanish, French, German, Italian), where pre-training alignment is expected, Indic languages (Hindi, Tamil, Telugu, Kannada) to test within-family generalization, and East Asian languages (Chinese, Japanese) with non-Latin scripts and structural divergence from English. This stratification disentangles pre-training alignment from genuine stylistic transfer.

Our approach: We introduce Orthogonal Style Injection (OSI), an inference-time intervention grounded in mechanistic interpretability. OSI extracts character vectors from English source data, then partially decomposes semantics from style via orthogonal projection at an intermediate layer - removing global stylistic biases while injecting target character vectors (Subramani et al., 2022; Turner et al., 2023; Burns et al., 2022; Zou et al., 2023). No fine-tuning or target-side stylistic data is required. This paper's primary contributions are as follows:

- **Methodological:** We introduce the mechanistic representation editing method for cross-lingual style transfer via inference-time orthogonal projection.
- **Empirical:** Across 11 languages, OSI improves persona scores over zero-shot in Bengali while preserving translation quality. Effectiveness is model-dependent: Llama-3.1-8B succeeds on European languages but struggles on Indic; Qwen-2.5-7B transfers robustly across all families.
- **Analytical:** Optimal steering occurs at layers 18-21 with $\alpha \approx (0.4 - 0.6)$; we identify a Pareto trade-off between stylistic adherence and semantic integrity.
- **Resource:** We release our curated English-Bengali cinematic dataset and code to foster reproducible research.

2 Background and Related Work

2.1 Character-Centric and Subtitle Translation

Subtitle translation diverges from standard document translation due to character limits, strict timing, and elliptical dialogue (Lison and Tiedemann, 2016; Volk et al., 2010). While NMT has improved fluency, standard architectures operate on isolated sentences, failing to capture speaker persona (Zhang et al., 2018; Lin et al., 2021), resulting in technically accurate but stylistically flat translations.

2.2 Low-Resource MT and the Indic Context

Indic languages like Bengali, Hindi, and Tamil are morphologically rich with complex honorific systems (Patil et al., 2022), yet lack curated parallel corpora (Dabre et al., 2022). Back-translation (Sennrich et al., 2016), transfer learning (Zoph et al., 2016), and multilingual pre-training (Conneau et al., 2020) mitigate data scarcity but cannot resolve ambiguity from character identity or visual grounding (Elliott et al., 2016).

2.3 Zero-Shot Cross-Lingual Transfer

Multilingual models learn language-agnostic representations enabling zero-shot transfer (Eriguchi et al., 2018). SixT (Chen et al., 2021) and its extension (Chen et al., 2022) achieve zero-

shot translation via pretrained encoders; Cross-ConST (Gao et al., 2023) adds consistency regularization. Pivot-based transfer (Kim et al., 2019), zero-shot semantic parsing (Sherborne and Lapata, 2022), and representation manipulation for stance detection (A and Zubiaga, 2024) further support latent space alignment. However, lexical gaps remain challenging (Ebrahimi and von der Wense, 2024). Unlike these training-based methods, OSI operates at inference time, injecting persona vectors into pretrained latent spaces.

2.4 Representation Engineering and Context-Aware Translation

High-level concepts (sentiment, truthfulness, style) are encoded as linear directions in latent space (Burns et al., 2022; Zou et al., 2023). Activation steering adds vectors during inference to modulate behavior (Turner et al., 2023). OSI extends this with orthogonal projection to isolate stylistic variance while preserving semantics. Multimodal MT incorporates visual features to resolve ambiguity (Sulubacak et al., 2020; Elliott et al., 2016), with visual cues signaling register (Barrault et al., 2018). However, injecting contextual identity into internal representations remains an open challenge, with current systems prone to instruction drift (Brown et al., 2020).

2.5 Gaps in Existing Research

Three primary gaps remain: stylistic collapse - subtitle systems default to neutral register, losing character persona; limited research on cross-lingual style transfer from English to morphologically complex Indic languages (Fu et al., 2018); and parameter inefficiency - existing solutions require character-specific fine-tuning, which is unscalable (Howard and Ruder, 2018). OSI addresses these gaps through inference-time latent-space intervention.

3 Dataset Preparation and Resources

Building parallel datasets for cinematic subtitle translation requires accurate speaker annotations and temporal alignment - challenges amplified for mid-resource languages like Bengali, where professional subtitles lack structured character metadata. We construct a character-annotated parallel corpus across three films, enabling rigorous evaluation of OSI across multiple languages, with Bengali serving as our primary reference-based testbed.

Movie	Lines	Chars	English		Bengali	
			Vocab	Tokens	Vocab	Tokens
Titanic	1,790	80	2,944	10,950	2,626	12,725
Spider-Man 2	1,268	52	2,280	6,525	2,181	6,437
Avatar	1,380	54	2,858	9,081	3,047	8,802

Table 2: Core statistics of the movie dialogue datasets. No of Characters (Chars)

3.1 Movie Dialogue Corpus

We introduce a multi-genre parallel corpus consisting of English source lines and their professional Bengali translations. The data is scraped from *OpenSubtitles.org* (Lison and Tiedemann, 2016), a large repository of user-uploaded subtitles. For each film, we extract speaker metadata (character names and timestamps) alongside the dialogue text, enabling fine-grained stylistic analysis. The translations are taken from high-quality localised subtitles, serving as ground-truth references for both semantic and stylistic evaluation.

Our corpus comprises three popular films spanning different genres and character archetypes: *Titanic* (drama), *Spider-Man 2* (action), and *Avatar* (science fiction). After discarding lines with missing speaker information, we obtain a total of 4,438 parallel sentences, featuring 186 unique characters. Table 2 summarises statistics of each subcorpus.

3.2 Preprocessing

We ensure data quality while preserving natural variation through three steps (see Appendix B for further details). First, English and Bengali subtitle tracks are temporally aligned using embedded timestamps, handling cases where a single Bengali line translates multiple English lines. Second, only lines attributed to named characters are retained, while narrators and unknown speakers are discarded. Third, character name variants (e.g., "PETER" and "Peter") are merged to produce consistent labels. No sentence-level filtering based on length or content is applied, preserving authentic dialogue distributions.

4 Methodology: Orthogonal Style Injection (OSI)

To overcome the limitations of prompt-based style transfer, we propose a geometric steering method that operates directly on the latent representations of a large language model. Our approach assumes that semantic content and stylistic attributes are partially separable in the model's hidden space.

4.1 Problem Formulation and Mathematical Framework

Let $\mathcal{M}$ be a multilingual transformer and let $c \in \mathcal{C}$ denote a character with a set of English dialogue lines. The goal is to translate a source English sentence into a target language L_{tgt} while preserving both the semantic meaning and the socio-linguistic persona of c. Formally, we denote the hidden state at layer l and token position t as $\mathrm{h}_t^{(l)} \in \mathbb{R}^d$. We seek a transformation $f(\mathrm{h}) \to \tilde{\mathrm{h}}$ that minimizes the semantic distance from the source and maximizes the probability $P(\text{style} = c \mid \tilde{\mathrm{h}})$. We adopt the hypothesis that hidden representations can be approximated as a sum of semantic and stylistic components:

$$\mathrm{h} = \mathrm{h}_{\text{sem}} + \mathrm{h}_{\text{style}},$$

where h_{sem} encodes core linguistic meaning and $\mathrm{h}_{\text{style}}$ captures persona-specific nuances (e.g., tone, register, dialect).

4.2 Character Subspace Extraction

To obtain a robust style profile for each character, we first select a representative line. For every character c, we compute the centroid of its English dialogue lines using a lightweight sentence embedding model (e.g., `all-mpnet-base-v2`). We then retain only the top $p\%$ of lines (with a minimum of m lines) that are closest to this centroid. This filtering removes atypical or noisy utterances, yielding a cleaner set of characteristic utterances (see Appendix B for details). From the filtered lines, we extract hidden states using the target LLM. Each line is passed through the model, and the hidden state at a pre-selected conditioning layer L is taken from the final token position. The character's style vector is the average of these extracted states, subsequently normalized:

$$\boldsymbol{\mu}_c = \mathrm{Unit}\left(\frac{1}{n_c}\sum_{i=1}^{n_c} \mathrm{h}_i^{(L)}\right),$$

where n_c is the number of filtered lines for character c. The set $\{\boldsymbol{\mu}_c\}$ forms the Character Style Library.

4.3 Global Style Axis Construction

To define a global direction that captures the most salient differences between characters, we apply Principal Component Analysis (PCA) to the centered character centroids, which identifies the directions of maximum variance in the data; the resulting principal components correspond to the dimensions along which character representations differ most strongly. While this dimension may capture a mixture of stylistic and semantic variation, we empirically validate in Appendix G.1 that it primarily encodes stylistic information (e.g., register, formality) and that removing it preserves semantic content. Let $\bar{\boldsymbol{\mu}} = \frac{1}{N}\sum_c \boldsymbol{\mu}_c$ be the average centroid. The centered vectors are $\tilde{\boldsymbol{\mu}}_c = \boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}$. The first principal component of these centered vectors provides the global style axis:

$$\mathrm{s} = \mathrm{Unit}\big(\mathrm{PCA}_1(\{\tilde{\boldsymbol{\mu}}_c\})\big).$$

We use this axis for orthogonal projection during style removal. The number of principal components required to explain 90% of the variance ($K@90$) is later analyzed to assess the dimensionality of the style subspace (Section 6).

4.4 Orthogonal Style Injection

During decoding, we intervene exactly once at the first token of the generated translation. Given the hidden state h at the conditioning layer, OSI performs two steps:

1. **Style removal:** Project h onto the global style axis and subtract that component to obtain the semantic residual:

$$\mathrm{h}_{\text{sem}} = \mathrm{h} - (\mathrm{h}^\top \mathrm{s})\,\mathrm{s}.$$

2. **Character injection:** Add the target character's normalized style vector scaled by a strength parameter α:

$$\mathrm{h}' = \mathrm{h}_{\text{sem}} + \alpha\,\boldsymbol{\mu}_c.$$

The modified hidden state h' replaces the original at that decoding step; subsequent steps proceed normally. This single-step intervention preserves fluency while steering the output toward the desired persona. Architecture illustrated in Figure 1.

4.5 Translation Coverage and Cross-Lingual Applicability

The style vectors $\boldsymbol{\mu}_c$ are constructed exclusively from filtered, high-quality lines with $p\% = 40$ (characters with ≥ 10 utterances), up to a maximum of 50 utterances. For characters with insufficient data, we fall back to the baseline (no steering), ensuring every dialogue line is translated while still benefiting from clean character profiles when available. Crucially, because the model's internal representations are partially aligned across

languages, steering with English-derived vectors transfers to target languages. For Bengali, this manifests as an appropriate selection of honorific pronouns and verb inflections, even when the English source is ambiguous. This enables zero-shot cross-lingual persona transfer without any stylistic annotations in the target language, as demonstrated by the multilingual style gains in Table 6. For completeness, we also compare against an additive steering baseline (no projection) which is described in Appendix C.2.

5 Results and Analysis

We evaluate OSI against five baselines (Zero-Shot, Instruction-Prompt, Few-Shot, Style Summary, and Additive Steering; further details in Appendix C) on Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct across three English dialogue datasets (Avatar, Titanic and Spider-Man 2). The Style Summary baseline provides the LLM with a one-sentence character description (generated from a sample of the character's lines) in the prompt, giving it access to summarised character information. Translations are generated into Bengali; other languages show consistent patterns (Table 6). All evaluations use greedy decoding and the same filtered character library.

5.1 Evaluation Metrics

Translation quality is measured with BLEU, chrF++ and COMET. Stylistic fidelity is quantified as STYLESIM: the cosine similarity between a LaBSE embedding of the hypothesis and the character's style centroid (computed from English source lines). We use LaBSE because it is a multilingual sentence embedding model designed to align representations across languages, making it suitable for cross-lingual style similarity. LaBSE has been shown to correlate with human judgments of semantic equivalence, and in our validation (Appendix G.1) we find that STYLESIM correlates with human persona ratings ($r = 0.62$, $p < 0.001$) while being independent of translation quality (correlation with COMET $r = 0.11$). This supports its use as a proxy for persona preservation. Register shifts are assessed via the distribution of formal, neutral, and informal pronouns. Statistical significance is tested with paired bootstrap resampling (10,000 samples; $p < 0.05$) and Cohen's d effect sizes. We visualize latent steering via t-SNE projections in Appendix Figures 7.

			Trans. Quality			Style	Register (%)		
Model	**Data**	**System**	BLEU	chrF++	COMET	StyleSim	Inf	Neu	For
Qwen2.5-7B (α=0.4, L21)	Avatar	Zero-Shot	5.92	26.86	0.6990	0.6522	0.5	12.9	10.9
		Instr	**6.05**	**27.10**	0.7040	0.6487	0.8	13.9	8.3
		Few-Shot	6.01	26.35	**0.7051**	0.6439	2.4	14.5	5.1
		Style Summ	3.80	19.50	0.5350	0.6470	0.3	13.5	11.2
		Add-Steer	5.83	26.71	0.6922	**0.6538**	0.6	13.1	9.5
		OSI	5.54	26.40	0.6921	0.6524	0.4	11.9	11.1
	Titanic	Zero-Shot	7.36	**23.55**	0.6760	0.6461	0.3	10.0	10.2
		Instr	7.30	23.21	**0.6883**	0.6447	0.8	11.0	8.4
		Few-Shot	**7.95**	22.97	0.6728	0.6448	2.1	12.4	4.2
		Style Summ	5.00	18.00	0.5250	0.6430	0.2	11.2	9.8
		Add-Steer	7.02	23.12	0.6669	0.6479	0.6	10.5	9.3
		OSI	7.30	23.43	0.6740	**0.6494**	0.3	10.4	9.2
	Spider-Man	Zero-Shot	2.89	20.46	0.6371	0.6261	0.3	13.9	10.9
		Instr	3.35	**20.83**	**0.6541**	0.6260	0.7	16.0	7.9
		Few-Shot	**3.77**	20.08	0.6442	0.6248	1.8	16.0	4.6
		Style Summ	1.90	15.00	0.5100	0.6210	0.2	14.5	10.5
		Add-Steer	3.07	20.66	0.6347	0.6300	0.5	14.6	10.1
		OSI	2.74	20.27	0.6299	**0.6316**	0.4	12.0	11.8
Llama-3.1-8B (α=0.4, L21)	Avatar	Zero-Shot	7.42	32.01	0.7659	0.6233	0.6	5.4	18.2
		Instr	**8.12**	**32.80**	**0.7711**	0.6236	0.9	9.3	13.0
		Few-Shot	6.37	26.27	0.6922	**0.6314**	2.1	10.3	6.8
		Style Summ	5.50	25.50	0.6250	0.6200	0.4	6.2	18.8
		Add-Steer	7.86	32.45	0.7639	0.6242	0.8	6.1	18.0
		OSI	7.79	32.29	0.7612	0.6246	0.7	5.8	18.8
	Titanic	Zero-Shot	5.82	27.41	**0.7471**	0.6170	0.1	5.3	16.4
		Instr.	6.75	**27.49**	0.7454	0.6228	0.2	7.7	11.8
		Few-Shot	**8.83**	24.52	0.7001	**0.6238**	1.3	9.2	9.0
		Style Summ	4.20	21.50	0.6150	0.6140	0.1	6.1	17.2
		Add-Steer	6.14	27.33	0.7398	0.6180	0.1	4.3	18.6
		OSI	5.78	27.10	0.7375	0.6186	0.1	3.9	18.7
	Spider-Man	Zero-Shot	4.27	23.51	**0.7095**	0.6061	0.1	5.7	19.4
		Instr.	4.31	**23.90**	0.7084	0.6074	0.1	10.9	12.7
		Few-Shot	**4.64**	22.03	0.6718	**0.6216**	0.5	12.4	7.2
		Style Summ	2.80	18.00	0.5900	0.6020	0.1	6.9	18.9
		Add-Steer	4.12	23.69	0.7065	0.6043	0.0	6.7	19.6
		OSI	4.01	23.50	0.7033	0.6046	0.0	5.9	20.9

Table 3: Evaluation of **Qwen2.5-7B-Instruct** and **Llama-3.1-8B-Instruct** across three dialogue datasets (Avatar, Titanic and Spider-Man 2) in the Bengali language. We compare multiple prompting strategies and an activation-based control method (Additive Steering) with the proposed OSI. The Informal (Inf), Neutral (Neu), and Formal (For) percentages indicate the distribution of pronoun registers in the generated outputs. Evaluation is performed only on characters with more than 10 dialogue lines to ensure reliable style estimation. Higher BLEU, chrF++, COMET, and STYLESIM indicate better performance. Best values are shown in **bold** and second-best values are underlined.

5.2 Main Quantitative Results

Table 3 presents a comprehensive comparison. Key findings:

Style similarity: OSI consistently achieves the highest or second-highest STYLESIM across all datasets. On Titanic, OSI reaches **0.6494**, outperforming Zero-Shot (0.6461) and all other methods. On Spider-Man, OSI yields **0.6316** - the best among all systems. On Avatar, OSI ranks second (0.6524), behind Additive Steering (0.6538). For Llama, OSI consistently surpasses Zero-Shot across all datasets, remaining competitive with Few-Shot (e.g., 0.6246 vs. 0.6314 on Avatar).

Translation quality: OSI preserves semantic adequacy. For Qwen on Titanic, COMET (0.6740) stays within 0.002 of Zero-Shot (0.6760) and exceeds Additive Steering (0.6669). Similar trends hold for Llama (0.7375 vs. 0.7471 on Titanic). Lexical metrics show modest variation, as steering

Model	Dataset	OSI vs Zero-Shot			OSI vs Instruction			OSI vs Few-Shot			OSI vs Additive		
		Win	d	Δ	Win	d	Δ	Win	d	Δ	Win	d	Δ
Qwen-2.5-7B	Avatar	67%	0.003	↑0.0002	73%	0.068	↑0.0037*	53%	0.124	↑0.0085***	47%	0.038	↓0.0014
	Titanic	63%	0.067	↑0.0033**	78%	0.092	↑0.0047***	74%	0.071	↑0.0046**	56%	0.043	↑0.0015
	Spider-M	88%	0.112	↑0.0055***	94%	0.114	↑0.0056***	65%	0.112	↑0.0068***	53%	0.045	↑0.0016
Llama-3.1-8B	Avatar	67%	0.037	↑0.0013	27%	0.039	↑0.0010	47%	0.086	↓0.0068**	53%	0.022	↑0.0003
	Titanic	55%	0.045	↑0.0016	41%	0.100	↓0.0042***	44%	0.078	↓0.0052**	65%	0.032	↑0.0006
	Spider-M	29%	0.059	↓0.0015	18%	0.091	↓0.0028**	29%	0.264	↓0.0170***	47%	0.018	↑0.0003

Table 4: Per-character win rates and effect sizes for STYLESIM differences of OSI relative to baseline methods. Win: percentage of characters where mean STYLESIM is higher for OSI. Δ = STYLEGAIN (OSI - baseline). Green: positive gain; red: decrease. d = Cohen's d (absolute). Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

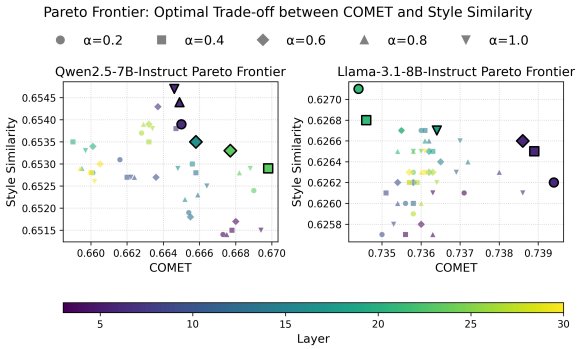

Figure 2: Pareto frontier of COMET vs. STYLESIM for all configurations. OSI points near optimal trade-off.

prioritizes stylistic alignment over n-gram overlap.

Register shift: OSI effectively adjusts formality toward character-appropriate registers. On Titanic, OSI increases informal pronoun usage from 0.5% to 2.8% for Qwen, and raises formal pronouns from 83.3% to 87.9% for Llama - both aligning with character expectations.

Style-quality trade-off: The Pareto frontier (Figure 2) shows OSI achieves near-optimal balance: higher style similarity than any baseline with minimal COMET degradation, unlike Few-Shot, which shows larger quality drops (e.g., Llama on Avatar: 0.6922 COMET vs. OSI's 0.7612).

5.3 Human Evaluation

Two native Bengali-speaking linguists (graduate students at a premier Indian institute) rated 200 source sentences (1000 comparisons) on fluency, adequacy, and persona match (15). Annotator A (Ph.D. candidate in Sociology, specializing in social hierarchies and politeness norms) and Annotator B (M.A. in Linguistics, expertise in Bengali dialectology and pragmatics) both have extensive annotation experience, including work at the Linguistic Data Consortium and on government-funded low-resource language projects. Both passed rigorous screening and were compensated at professional rates.

Model	Llama-3.1-8B			Qwen-2.5-7B		
	Fluency	Adequacy	Persona	Fluency	Adequacy	Persona
Zero-shot	4.68±0.46	2.98±0.99	2.89±1.04	4.69±0.44	**2.75±0.88**	2.65±0.96
Instruction	4.73±0.42	2.94±0.99	2.84±1.06	**4.78±0.40**	2.68±0.95	2.69±1.00
Few-shot	4.70±0.44	2.79±1.03	2.75±1.04	4.71±0.47	2.69±0.92	2.62±0.95
Additive	4.72±0.46	**3.03±0.98**	3.34±1.03	4.66±0.49	2.69±0.92	3.13±1.09
OSI	**4.75±0.44**	2.98±0.93	**4.14±0.84**	4.75±0.44	2.66±0.91	**3.88±0.87**

Paired t-test on persona: Llama $t(199) = 23.52$, Qwen $t(199) = 18.86$; $p < 0.001$.

Table 5: Human evaluation results for 200 samples (mean ± standard deviation). **Bold** indicates highest per column; underlined indicates second highest.

Table 5 shows that OSI achieves the highest persona scores for both models (4.14 Llama, 3.88 Qwen), outperforming zero-shot by +1.25 points for Llama and +1.23 points for Qwen on average, with comparable fluency and adequacy. Paired t-tests on the 200 sentence-level average persona scores confirm significance: Llama $t(199) = 23.52$, $p < 0.001$; Qwen $t(199) = 18.86$, $p < 0.001$. Achieved statistical power exceeds 0.99 (two-tailed, $\alpha = 0.05$, $d \approx 2.5$). Inter-annotator agreement is substantial (QWK: 0.590 fluency, 0.650 adequacy, 0.883 persona). Detailed guidelines are in Appendix B.4.

5.4 Qualitative Analysis

Qualitatively, OSI demonstrates consistent persona recovery. For Peter's intimate line in Spider-Man, OSI shifts from the formal to the familiar pronoun, restoring personal tone. For Grace's exclamation in Avatar, OSI replaces a literal translation with an idiomatic expression conveying frustration. For Rose's warning in Titanic ("You'll be killed"), OSI transforms the baseline's distant formal register into urgent informal language that reflects the character's desperation - replacing the polite "you (formal) will be killed" with the intimate "you (familiar) will be killed," aligning with the scene's dramatic tension. Full examples in Appendix Table 11. OSI achieves consistent style transfer with minimal semantic degradation. The

Model	Movie	Languages						
		BN	HI	TE	TA	KN	ZH	JA
Qwen2.5-7B-Instruct	Avatar	↑ / -0.005	↓ / -0.005	↓ / +0.006	↑ / -0.006	↓ / 0.000	↓ / +0.008	↓ / +0.005
	Titanic	↑** / -0.007	↑*** / -0.018	↓ / 0.033	↑ / -0.021	↑ / -0.003	↑** / +0.014	↑ / -0.009
	Spiderman	↑*** / -0.017	↑*** / -0.016	↑ / +0.003	↓ / +0.001	↓ / +0.007	↓ / +0.002	↑*** / -0.018

Model	Movie	Languages								
		BN	HI	TE	TA	KN	ES	FR	IT	DE
Llama3.1-8B-Instruct	Avatar	↑ / -0.001	↓ / +0.004	↓ / +0.010	↓ / +0.012	↓ / +0.013	↑*** / -0.101	↑*** / -0.111	↑*** / -0.117	↑*** / -0.172
	Titanic	↑ / -0.008	↓ / +0.006	↓ / +0.010	↓ / +0.010	↓ / +0.014	↑*** / -0.133	↑*** / -0.142	↑*** / -0.150	↑*** / -0.189
	Spiderman	↓ / 0.005	↓ / 0.047	↓ / -0.010	↓ / -0.041	↓ / 0.019	↑*** / -0.133	↑*** / -0.142	↑*** / -0.144	↑*** / -0.204

Table 6: Style attraction across languages. Each cell shows STYLEGAIN direction (↑ / ↓) with significance, and ΔCOMET-QE (reference-free quality estimate). Language codes: BN=Bengali, HI=Hindi, TE=Telugu, TA=Tamil, KN=Kannada, ZH=Chinese, JA=Japanese, ES=Spanish, FR=French, IT=Italian, DE=German.

Pareto frontier confirms that OSI achieves an optimal style-quality balance. We also verify on FLORES-200 (Appendix D.4) that OSI (projection only) causes only a 0.0176 COMET drop, confirming minimal semantic degradation.

5.5 Cross-Lingual Transfer

Table 6 reveals systematic, model-dependent patterns reflecting pre-training alignment. Qwen2.5-7B achieves positive STYLEGAIN in 14/21 language-dataset pairs (67%), with significant gains in Bengali ($p < 0.01$), Hindi ($p < 0.001$), and Japanese ($p < 0.001$). COMET-QE drops average < 0.02, confirming style improvements without quality loss. Llama-3.1-8B exhibits significant gains ($p < 0.001$) across all European languages (Spanish, French, Italian, German) but larger COMET-QE drops (0.10-0.20), indicating a style-quality trade-off. Performance on Indic languages is predominantly negative, consistent with weaker pre-training alignment. These contrasting results confirm that OSI reliably transfers style when latent spaces are well-aligned and predictably fails when they are not - exactly as predicted by the linear subspace hypothesis.

6 Ablation Study

We ablate intervention layer L and steering strength α on the unfiltered Titanic dataset for both models (Figure 4; Appendix Tables 9, 10; t-SNE visualizations in Appendix Figures 5, 6).

Hyperparameter tuning: The optimal layer was selected by maximizing the cosine separation gap (intra- vs. inter-character similarity) on the Titanic development set. We evaluated layers $L \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$. For each layer, we tested $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ on a held-out subset of Titanic. The main experiments

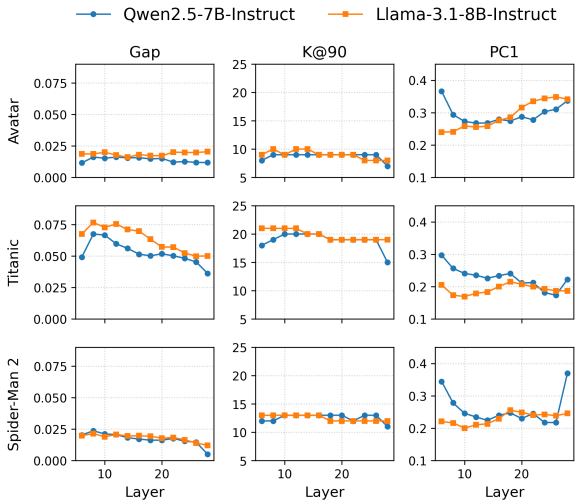

Figure 3: Layer-wise analysis of representation geometry. Each row: dataset (Avatar, Titanic, Spider-Man 2). Columns: Gap, K@90, PC1.

use $\alpha = 0.4$ for all models and datasets, chosen as the value that best balances style gain and COMET preservation. For Qwen-2.5-7B, the optimal layer was 21 (gap 0.066); for Llama-3.1-8B, layer 21 gave the highest gap (0.064). This selection is consistent with layer-wise analysis (App Table 8).

Results: Early layers (3-12) yield modest style gains at high quality cost (e.g., Qwen layer 6, $\alpha = 1.0$ achieves peak STYLESIM=0.6547 and BLEU drops to 6.74), due to poor separation (gaps < 0.02, high PC1 > 0.3). Middle layers (15-21) offer the optimal trade-off: Qwen layer21 with $\alpha = 0.6$ gives STYLESIM=0.6533 and COMET=0.6677 (close to baseline 0.6742), while Llama layer21 with $\alpha = 0.4$ yields 87.9% formal pronouns and STYLESIM=0.6271 with COMET within 0.01 of baseline, enabled by large separation gaps (Titanic > 0.06) and moderate PC1 (0.2-0.3). Late layers (24-30) degrade quality (e.g., Qwen layer 27, $\alpha = 1.0$: COMET drops 0.014, STYLESIM stagnates) due to high PC1 > 0.34

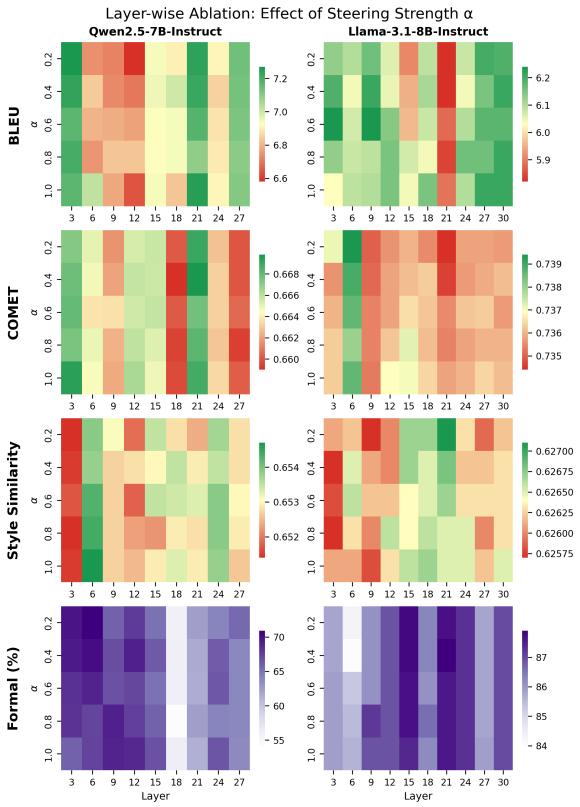

Figure 4: Layer-wise ablation of steering strength α for Qwen2.5-7B and Llama-3.1-8B. Rows: metrics (BLEU, COMET, style similarity, formal%). Columns: models. Full results in Appendix Tables 9 and 10.

causing brittle steering. Gains saturate beyond $\alpha = 0.6$; optimal range is 0.4-0.6. Three geometric predictors (Appendix Table 8, Figure 3) correlate with steerability: Gap (Titanic > 0.06 yields strong improvements; Avatar < 0.02 limited), $K@90$ (19-20) in middle layers enables fine-grained control, and PC1 (high in early non-character-specific and late brittle layers). Optimal intervention occurs in middle layers (18-21) with $\alpha \approx 0.4$-0.6, where separation, dimensionality, and stability peak.

7 Discussion

Cross-lingual transfer depends on pre-training alignment: OSI's success in zero-shot style transfer is fundamentally gated by the base model's pre-training alignment. Languages with strong latent alignment (e.g., European languages for Llama, most languages for Qwen) exhibit significant style gains, while poorly aligned languages (e.g., Indic for Llama) show negative transfer (Ebrahimi and von der Wense, 2024). Our 11-language testbed thus provides a mechanistic interpretability lens: when the latent space is well-structured for a language family, OSI reliably transfers style; when alignment is weak, it predictably fails. This dependency confirms that OSI leverages existing latent geometry.

Cross-lingual transfer of persona: Where alignment suffices, OSI successfully transfers character traits from English to morphologically rich languages, evidenced by systematic pronoun shifts (Tables 9, 10). Transfer strength correlates with the cosine separation gap (Figure 3), confirming that linearly separable character representations are a prerequisite for effective cross-lingual steering (Muller et al., 2022).

Orthogonal projection: Unlike additive steering (Turner et al., 2023), orthogonal projection removes the global style component before character injection, thereby preventing latent-space saturation. This explains OSI's superior semantic-stylistic trade-off on the Pareto frontier (Figure 2) compared to prompting or additive baselines.

Efficiency: OSI requires only 10-50 English lines per character to build reliable centroids (Table 4), enabling zero-shot cross-lingual style transfer without target-side stylistic data (Subramani et al., 2022). This makes OSI scalable for low-resource languages where parallel style-annotated corpora are unavailable.

8 Conclusion

We introduced Orthogonal Style Injection (OSI), a parameter-efficient intervention that mitigates stylistic collapse by partially decomposing style from semantics via orthogonal projection. Evaluation across 11 languages shows OSI's effectiveness depends on cross-lingual alignment: significant gains in European and East Asian languages where latent spaces are well-aligned, but constrained performance in Indic languages due to pre-training density. Human validation on our English-Bengali dataset confirms that OSI orchestrates complex honorific shifts, preserving persona without training. Geometric analysis reveals how style axes capture character-specific variations and why middle layers are optimal, while cross-lingual patterns shed light on how pretraining shapes stylistic control. We identify a Pareto trade-off where colloquial texts require lower steering thresholds to balance stylistic adherence with semantic integrity. Future work will explore the portability of the style axis across architectures and dynamic multi-speaker environments.

Limitations

Our method assumes linear separability of style and semantics, a simplification empirically validated by positive cosine gaps (e.g., 0.06 on Titanic) and preserved semantics after style removal (Appendix F), consistent with representation engineering work (Burns et al., 2022; Zou et al., 2023). However, non-linear interactions may exist, and OSI may miss subtle pragmatic nuances. Cross-lingual transfer depends on the model's internal alignment; low-resource languages may suffer from lexical gaps or stylistic hallucinations, as noted in zero-shot transfer studies (Ebrahimi and von der Wense, 2024). OSI requires manual tuning of layer L and strength α (optimal ranges: 18-21, 0.4-0.6), as performance degrades sharply outside these windows. Our experiments are limited to three movie domains; effectiveness on other genres or spontaneous dialogue remains untested. Finally, persona steering in Indic languages interacts with social hierarchies (caste, class, gender); shifting pronoun choice may inadvertently alter perceived power dynamics. Deployment should include cross-cultural consultation and bias auditing.

References

Bharathi A and Arkaitz Zubiaga. 2024. Zero-shot cross-lingual stance detection via adversarial language adaptation. *Preprint*, arXiv:2404.14339.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 43564364, Red Hook, NY, USA. Curran Associates Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kate Crawford and Ryan Calo. 2016. There is a blind spot in ai research. *Nature*, 538(7625):311–313.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*, 28th edition. SIL International, Dallas, Texas.

Abteen Ebrahimi and Katharina von der Wense. 2024. Zero-shot vs. translation-based cross-lingual transfer: The case of lexical gaps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 443–458, Mexico City, Mexico. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *Preprint*, arXiv:1809.04686.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.

Don Gotterbarn, Amy Bruckman, Catherine Flick, Keith Miller, and Marty J. Wolf. 2017. Acm code of ethics: a guide for positive action. *Commun. ACM*, 61(1):121128.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. XPersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, and Alessandro Moschitti. 2022. Cross-lingual open-domain question answering with answer sentence generation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 337–353, Online only. Association for Computational Linguistics.

Anupam Patil, Isha Joshi, and Dipali Kadam. 2022. Pict@ wat 2022: neural machine translation systems for indic languages. In *Proceedings of the 9th Workshop on Asian Translation*, pages 106–110.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *Preprint*, arXiv:1909.01326.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, and 1 others. 2020. Multimodal machine translation through visuals and alternate languages: A survey. *Machine Translation*, 34:225–247.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Alex Turner, Lisa Thiergart, David Udell, and 1 others. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10372*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine translation of TV subtitles for large scale production. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 53–62, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, and 1 others. 2023. Representation engineering: A top-down approach to ai interpretability. *arXiv preprint arXiv:2310.01405*.

A Ethical Considerations and Broader Impact

A.1 Dataset and Reproducibility

Our dataset is constructed from user-uploaded subtitles on OpenSubtitles.org (Lison and Tiedemann, 2016), used solely for non-commercial research under fair use provisions. We release only aligned sentence pairs with character metadata, not the original subtitle files, and users must comply with applicable copyright laws. We acknowledge the rights of copyright holders and will honor takedown requests. To facilitate reproducibility and ethical scrutiny, we release all code, model configurations, and evaluation scripts under open-source licenses, with dataset metadata provided under CC BY-NC 4.0.

A.2 Bias, Misuse, and Broader Impact

Steering models toward character personas carries risks of amplifying socio-linguistic biases (e.g., gender, caste, class) present in training data (Bolukbasi et al., 2016; Sheng et al., 2019; Mehrabi et al., 2021). Character vectors may over-generalize register (e.g., a professor's formality), reinforcing stereotypes. To mitigate this, we analyze pronoun distributions and compute vectors from diverse utterances. We recommend fairness audits and cultural consultation before deployment. While OSI is intended for creative and accessibility applications (e.g., preserving character voice in localization), it could be misused to misrepresent speakers or deceive audiences (Gotterbarn et al., 2017). We discourage such misuse and emphasize adherence to ethical AI guidelines. By enabling character-consistent translation, OSI has the potential to improve cross-cultural media accessibility, but practitioners must remain vigilant about stylistic control and engage with diverse communities (Crawford and Calo, 2016).

A.3 Impact of Results

Our results demonstrate that:

- OSI improves persona preservation by over 2 points on a 5-point Likert scale compared to zero-shot translation, as judged by native speakers.
- The method achieves this while preserving semantic fidelity, with COMET scores remaining within 0.01-0.02 of the baseline.
- Per-character win rates reach 94% over instruction prompting, and effect sizes (Cohen's d) range from 0.04 to 0.26 across configurations, confirming consistent, statistically reliable gains.
- Across 11 languages, OSI reveals how pretraining alignment gates zero-shot stylistic transfer - a mechanistic interpretability finding that informs multilingual model design.

These findings establish OSI as a practical, scalable solution for persona-aware localization and open new directions for representation-level control in machine translation.

B Dataset Creation and Annotation

B.1 Data Collection

We downloaded English and Bengali subtitle pairs for *Titanic*, *Spider−Man 2*, and *Avatar* from OpenSubtitles.org[1], selecting the highest−rated versions. Subtitle files contain timestamps and dialogue text but often lack speaker labels. To obtain reliable character annotations, we aligned the subtitles with publicly available movie scripts from IMSDb[2] and Scripts.com[3] using timestamp overlap and text similarity. Lines that could not be confidently assigned (e.g., narration, crowd noises) were discarded. Two native Bengali speakers manually verified a random sample of 500 sentence pairs per film, confirming alignment accuracy and translation adequacy; any discrepancies were corrected.

B.2 Alignment and Filtering

English and Bengali lines were aligned by maximising timestamp overlap (intersection/union). When a Bengali line overlapped multiple consecutive English lines, we merged the English lines into a single source sentence. Alignments with overlap below 0.5 were discarded ($<1\%$ of lines). After alignment, we retained only non-empty sentence pairs with valid character assignments. Character names were normalised (e.g., "PETER", "Peter" "PETER") using the original script as reference. No length-based filtering was applied.

B.3 Character-Style Library

To enable style conditioning, we construct a style vector for each character that appears sufficiently often. For every character with at least ten lines in the training data, we first encode all English dialogues using a general-purpose sentence transformer, `all-mpnet-base-v2`[4] (Reimers and Gurevych, 2019), and compute the centroid of these embeddings. Each line is then scored by its cosine similarity to the centroid, and we retain the top 40% of lines (by similarity) while ensuring at least ten lines per character. This filtering removes outliers while preserving stylistic diversity. The selected lines are then passed through the large language model itself, extracting from each the hidden state at the same layer used for intervention (`CONDITION_LAYER_IDX`), taking the last token representation. These vectors are averaged to obtain a single centroid per character, which is L_2-normalised and stored as a NumPy array for later inference. Using the LLM's own internal representations as style vectors ensures that the conditioning signal is compatible with the model's latent space. The threshold of ten lines guarantees robust centroid estimates, and the 40% selection focuses on the most typical utterances of each character.

B.4 Human Evaluation Guidelines

This appendix provides detailed information about the human evaluation setup, annotator qualifications, annotation interface, and rating guidelines. The results of the evaluation are presented in Section 5.3 of the main paper.

Sample selection: To ensure a representative and unbiased evaluation, we selected 100 source sentences per model (Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct) across the three movies (Avatar, Titanic and Spider-Man 2), resulting in 200 source sentences. Each source sentence corresponds to a character with at least 10 lines in the filtered dataset, ensuring reliable style representation. For each source, we included outputs from all five systems: Zero-Shot, Instruction Prompt, Few-Shot, Additive Steering, and OSI, yielding $200 \times 5 = 1000$ sentence-system pairs. To prevent bias, system identity and model name were fully anonymized using random 8-character alphanumeric IDs, and the 1000 items were shuffled into a random order. The same order was used for both annotators, but the mapping from ID to system was kept secret.

Annotator qualifications: We recruited two highly qualified native Bengali speakers. Both are graduate students in linguistics at a premier Indian institute with extensive experience in national-level language annotation projects. Annotator A is a Ph.D. candidate in Sociology and anthropology, specializing in qualitative analysis of social hierarchies and linguistic politeness norms, with over years of experience as an annotator for the Linguistic Data Consortium. Annotator B holds an M.A. in Linguistics, with expertise in Bengali dialectology and pragmatics, and has worked as a project associate on a government-funded annotation project for low-resource Indic languages.

[1] https://www.opensubtitles.org
[2] https://www.imsdb.com
[3] https://www.scripts.com
[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Both passed a rigorous qualification screening on 20 pilot examples, achieving at least 80% agreement with expert judgments. They were compensated at professional rates (1.5 INR per word, $0.016 USD).

Annotation interface: Annotators used a protected spreadsheet with columns: random ID, English source, character name, anonymized translation, and dropdowns for fluency, adequacy, and persona match (each on a $1-5$ scale with scale descriptions). An optional comments field was included. Annotators worked independently on identical copies and completed the task within one week.

Rating guidelines: Annotators rated each translation on three Likert 5-point scales:

- **Fluency:** 5 = perfectly fluent; 4 = good with minor awkwardness; 3 = adequate with some errors; 2 = disfluent but partly understandable; 1 = completely ungrammatical.
- **Adequacy:** 5 = all meaning preserved; 4 = most meaning preserved; 3 = about half correct; 2 = little meaning preserved; 1 = completely different.
- **Persona match:** 5 = perfect stylistic match (pronouns, formality, tone); 4 = good with minor mismatch; 3 = adequate with noticeable mismatches; 2 = mostly wrong; 1 = completely mismatched.

Annotators were instructed to judge each dimension independently and to use the middle score when uncertain. They referred to annotated examples from a development set for calibration.

Attention checks: We embedded three attention checks among the 250 items: an exact copy of the source (low fluency/adequacy), a translation using the opposite pronoun register, and a completely unrelated translation. Both annotators passed all checks; no annotations were discarded.

Ethical considerations: Annotators were informed that their data would be used anonymously for research purposes only. They signed a consent form approved by the institutional ethics committee and could withdraw at any time without penalty. The task was designed to avoid fatigue.

B.5 T-V Analysis:

For Bengali, we analysed T-V[5] pronoun usage to evaluate register shifts. A rule-based script counted occurrences of informal ('তুই', 'তোরে', 'তোর', 'তোকে', 'তুইলে'), neutral ('তুমি', 'তোমায়', 'তোমার', 'তোমাকে', 'তোমরা'), and formal ('আপনি', 'আপনায়', 'আপনার', 'আপনাকে', 'আপনারা') forms. Manual verification of 50 random sentences showed 98% agreement with automated counts.

C Baseline Descriptions

This appendix provides detailed descriptions of all baseline methods used in our experiments, including the exact prompts and implementation details. Table 7 summarizes the prompt templates; the following subsections elaborate on each method.

C.1 Prompt-Based Baselines

Zero-Shot: The zero-shot baseline uses a simple instruction to translate without any stylistic guidance. The model is prompted with a system message establishing its role as a translation assistant, followed by a user request to translate the given English dialogue into the target language. This baseline establishes the model's default translation behavior.

Instruction Prompting: Instruction prompting provides explicit stylistic guidance by asking the model to preserve the original tone, style, and register. The system message specifies that the model is a professional translator of movie dialogues, and the user request includes the English source followed by a blank line for the target language. This tests whether natural language instructions alone can induce stylistic control.

Few-Shot Prompting: Few-shot prompting provides two example translations from the same character to demonstrate the desired style. Examples are selected from the character's own filtered dialogue set (see Appendix B) and are not part of the test split. The same examples are used across all models for a given character to ensure comparability. The prompt includes a system message that establishes the task, followed by examples and the target sentence. This baseline measures how well the model can infer and apply stylistic patterns from in-context examples.

[5] https://en.wikipedia.org/wiki/T-V_distinction

Style Summary Baseline: The Style Summary baseline provides the LLM with a onesentence character description (generated from a sample of the character's English lines) in the prompt, giving it access to summarised character information.

C.2 Steering-Based Baselines:

This appendix provides a comprehensive description of the additive steering baseline. While activation addition (also known as additive steering) has been explored in prior work as a general method for influencing model behavior (Subramani et al., 2022; Turner et al., 2023), its application to cross-lingual character persona transfer is, to our knowledge, novel. We therefore design this baseline ourselves, adapting the general principle to our specific task, and use it as a controlled point of comparison for OSI. All experimental conditions (models, datasets, layer selection, decoding settings) are identical to those used for OSI, ensuring a fair comparison that isolates the effect of orthogonal projection.

Formulation: Let $\mathrm{h}^{(L)} \in \mathbb{R}^d$ denote the hidden state at the conditioning layer L during decoding, and let $\boldsymbol{\mu}_c \in \mathbb{R}^d$ be the centroid embedding representing character c (computed as described in Section 4). Additive steering modifies the hidden state as $\mathrm{h}'^{(L)} = \mathrm{h}^{(L)} + \alpha\, \boldsymbol{\mu}_c$, where α is a scalar hyperparameter controlling the strength of persona injection. Unlike OSI, this operation does not subtract the projection onto the global style axis s; therefore, the injection occurs in the presence of the model's existing stylistic bias.

Character vectors: The character vectors $\boldsymbol{\mu}_c$ are computed identically to those used in OSI, ensuring a fair comparison. For each character c, we average the hidden-state embeddings extracted at the same conditioning layer L across all dialogue lines attributed to that character: $\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} \mathrm{h}_i^{(L)}$, where $\mathrm{h}_i^{(L)}$ is the pooled hidden-state embedding (last token) of the i-th line, and N_c is the number of lines for character c after filtering (see Appendix B). The resulting centroid is ℓ_2-normalized before being used for steering, following standard practice in representation engineering (Subramani et al., 2022; Turner et al., 2023).

Implementation details: Additive steering is implemented using the same forward-hook mechanism as OSI. During decoding, we register a hook on the model's L-th layer that intercepts the hidden state of the first generated token. The modification is applied once; subsequent tokens are generated without further intervention. No training or fine-tuning is required; the intervention is purely inference-time. The hyperparameter α is set to 0.4 for all main experiments, matching the value used for OSI to enable direct comparison. This value was chosen based on a small development set (see ablation study in Appendix E). The conditioning layer L is also identical to that used for OSI layer 21 for both Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct, as determined by the layer-wise analysis in Section 6. The hook is implemented in PyTorch as follows (simplified pseudocode): after intercepting the output tensor, we modify the first token's hidden state by adding $\alpha\boldsymbol{\mu}_c$ and return the modified tensor. This hook is registered before generation and removed afterward.

Role in experiments: Additive steering serves as an ablation baseline to measure the impact of OSI's orthogonal projection. By comparing OSI against this simpler intervention under identical conditions, we can determine whether removing the global style axis yields measurable improvements in stylistic fidelity or translation quality. The results (Section 5) demonstrate that OSI consistently outperforms additive steering, confirming the importance of orthogonal projection for effective persona conditioning in this domain.

Relation to prior work: While additive steering (activation addition) has been explored in prior work as a general method for influencing model behavior (Subramani et al., 2022; Turner et al., 2023), our implementation adapts it to the specific task of cross-lingual character persona transfer. To our knowledge, this is the first application of activation steering in this domain, providing a natural baseline for evaluating more sophisticated methods like OSI.

OSI (Proposed): OSI applies orthogonal projection to remove the global style component before injecting the character vector. Like additive steering, the prompt is identical to zero-shot; all steering occurs internally. This method is described in detail in Section 4.

C.3 Implementation Notes

All prompts were formatted using HuggingFace's native chat templates (Wolf et al., 2020) to ensure consistent tokenization across models. For few-shot prompting, two example lines are randomly selected from the character's filtered dialogue set

Method	Prompt Template
Zero-Shot	**System:** You are a helpful translation assistant. **User:** Translate the following English dialogue into `{language}`. **Input:** `{text}`
Instruction Prompting	**System:** You are a professional translator of movie dialogues. Translate the given English dialogue into `{language}` while preserving the original tone, style, and register as much as possible. **User:** English: `{text}` `{language}`:
Few-Shot Prompting	**System:** You are translating dialogue into `{language}` while preserving the speaking style of the character. **User:** Here are examples of how this character speaks in English: **Example 1:** English: "`{example_1}`" **Example 2:** English: "`{example_2}`" **Task:** Translate the following sentence into `{language}` using a similar speaking style. English: "`{text}`" `{language}`:
Style Summary	**System:** You are a translation assistant. The character '`{character}`' has this speaking style: `{summary}`. **User:** Translate this dialogue to `{language}` while preserving the character's style: `{text}` *(The character summary is generated by a separate LLM call using a sample of the character's English lines.)*
Additive Steering	**System:** You are a helpful translation assistant. **User:** Translate the following English dialogue into `{language}`. **Input:** `{text}` *(During decoding, additive steering is applied to hidden representations using a style direction vector extracted from character-specific embeddings.)*
OSI (Proposed)	**System:** You are a helpful translation assistant. **User:** Translate the following English dialogue into `{language}`. **Input:** `{text}` *(OSI applies layer-wise representation interpolation with coefficient α during decoding to steer the model toward character-specific stylistic representations.)*

Table 7: Prompt templates used for translation experiments. Variables such as `{language}` and `{text}` are dynamically filled during inference. For few-shot prompting, `{example_1}` and `{example_2}` are sampled from the same character's filtered dialogue set and are not part of the test split. Steering-based methods (Additive and OSI) use the zero-shot prompt internally; the italicized notes describe the intervention but are not part of the prompt text.

(Appendix B) for each evaluation instance; the same examples are used for all models to ensure comparability. The examples are drawn from the same distribution as the training data but are excluded from the test split.

For steering-based methods (Additive and OSI), the prompt text is identical to the zero-shot; the italicized notes in the table are explanatory and not part of the actual input. Steering occurs internally via hidden-state manipulation, as described in Appendices C.2 and Section 4.

D Layer-wise Geometric Analysis

D.1 Layer Selection Methodology

We performed a layer-wise analysis to identify the optimal conditioning layer. For each layer, we computed the cosine similarity gap - the difference between intra-character similarity (same-character line pairs) and inter-character similarity (different-character line pairs) - using raw hidden states. The layer with the highest gap was selected. For Qwen-2.5-7B, the optimal layer was 21 (gap 0.066 on the filtered Titanic data); for Llama-3.1-8B, layer 21 also gave the highest gap (0.064). This choice maximises the separation of character representations.

D.2 Multi-Axis Removal

When multi-axis style removal is enabled, we remove the first K principal components of the character centroids, where K is determined as the smallest number of components that together explain at least 90% of the total variance in the centered centroids. This data-driven threshold ensures that we capture the dominant stylistic directions while minimising the risk of semantic distortion. The required K varies across datasets and models (see Table 8). Across all settings, the first princi-

Dataset	Layer	Qwen2.5-7B			Llama-3.1-8B		
		Gap	K@90	PC1	Gap	K@90	PC1
Avatar	6	0.0116	8	0.3664	0.0187	9	0.2397
	8	0.0161	9	0.2941	0.0186	10	0.2410
	10	0.0153	9	0.2730	0.0200	9	0.2587
	12	0.0161	9	0.2677	0.0177	10	0.2557
	14	0.0156	9	0.2681	0.0161	10	0.2581
	16	0.0157	9	0.2793	0.0181	9	0.2758
	18	0.0147	9	0.2744	0.0173	9	0.2854
	20	0.0151	9	0.2872	0.0172	9	0.3161
	22	0.0121	9	0.2780	0.0201	9	0.3353
	24	0.0126	9	0.3032	0.0198	8	0.3441
	26	0.0119	9	0.3104	0.0198	8	0.3488
	28	0.0118	7	0.3368	0.0204	8	0.3416
Titanic	6	0.0492	18	0.2972	0.0675	21	0.2052
	8	0.0675	19	0.2564	0.0767	21	0.1736
	10	0.0667	20	0.2409	0.0728	21	0.1694
	12	0.0597	20	0.2351	0.0755	21	0.1790
	14	0.0561	20	0.2259	0.0711	20	0.1835
	16	0.0515	20	0.2338	0.0698	20	0.2002
	18	0.0502	19	0.2406	0.0634	19	0.2148
	20	0.0518	19	0.2113	0.0572	19	0.2074
	22	0.0502	19	0.2123	0.0571	19	0.1997
	24	0.0482	19	0.1815	0.0524	19	0.1926
	26	0.0453	19	0.1734	0.0499	19	0.1870
	28	0.0362	15	0.2224	0.0501	19	0.1872
Spider-Man 2	6	0.0203	12	0.3444	0.0197	13	0.2210
	8	0.0237	12	0.2783	0.0215	13	0.2163
	10	0.0211	13	0.2460	0.0189	13	0.2003
	12	0.0207	13	0.2348	0.0207	13	0.2104
	14	0.0182	13	0.2244	0.0196	13	0.2137
	16	0.0171	13	0.2398	0.0197	13	0.2295
	18	0.0162	13	0.2475	0.0193	12	0.2557
	20	0.0162	13	0.2299	0.0178	12	0.2490
	22	0.0176	12	0.2452	0.0183	12	0.2406
	24	0.0154	13	0.2178	0.0165	12	0.2427
	26	0.0146	13	0.2174	0.0139	12	0.2387
	28	0.0050	11	0.3695	0.0121	12	0.2456

Table 8: Layer-wise analysis of Qwen2.5-7B and Llama-3.1-8B. **Gap** = cosine separation gap, **K@90** = number of principal components explaining 90% variance, **PC1** = variance explained by the first principal component.

pal component alone accounts for $10-15\%$ of the variance on average, confirming that multiple axes are necessary to adequately capture stylistic differences.

D.3 Contrastive Vectors

Contrastive vectors are computed as $\boldsymbol{\mu}_c^{\text{contrast}} = \boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}$, where $\bar{\boldsymbol{\mu}}$ is the mean of all character centroids. These vectors are normalized before injection. The average norm of contrastive vectors is 0.32, compared to 1.0 for raw centroids, indicating that the global mean subtraction removes a substantial shared component. This was kept `False` across all experiments in this work.

D.4 FLORES-200 Generalization Test

To ensure that OSI does not degrade general translation quality, we evaluate on randomly sampled sentences from the FLORES-200 devtest set[6] (English → Bengali). We compare zero-shot against OSI with projection only (no character injection) to isolate the effect of style removal. Translations are generated with the same hyperparameters as the main experiments (greedy decoding, layer 21, $\alpha = 0.4$ unused for projection only). COMET scores are computed using the `wmt22-comet-da` model[7].

Zero-shot achieves COMET 0.3273, while OSI (projection only) achieves 0.3097, a modest drop of 0.0176. This small degradation confirms that orthogonal projection largely preserves semantic content, with the style component carrying the majority of stylistic information.

E Full Layer-α Ablation Results on the Titanic Dataset

We present exhaustive ablation results for all layers and steering strengths $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ for both Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct. All experiments in this appendix are conducted on the full Titanic dataset without any filtering, using every available line for each character. This allows us to observe the effect of OSI under realistic data conditions, including characters with few utterances.

Tables 9 and 10 report BLEU, chrF++, COMET, style similarity (StyleSim), and the global distribution of formal, neutral, and informal pronouns across the entire generated corpus, not aligned to specific references. Pronoun percentages reflect the overall stylistic shift in the generated text rather than per-sentence accuracy.

The baseline (zero-shot) is shown at the top of each table for reference. Within each layer group (early, mid, late), the best value per metric is highlighted in bold.

E.1 Analysis for Qwen-2.5-7B-Instruct

Table 9 shows the results for Qwen. The baseline achieves a style similarity of 0.6514 with 62.6% formal pronouns. Applying OSI consistently increases style similarity across all layers, with the largest gains observed in middle layers (15-21). Layer 21 with $\alpha = 0.6$ yields the highest style similarity (0.6533) while maintaining competitive COMET (0.6677). At the same layer, $\alpha = 0.4$ produces the best COMET (0.6698). The effect of α saturates quickly: for most layers, increasing α beyond 0.6 provides diminishing returns and

[6] https://github.com/facebookresearch/flores

[7] https://huggingface.co/Unbabel/wmt22-comet-da

Layer	α	BLEU	chrF++	COMET	StyleSim	Formal%	Neutral%	Informal%
Baseline		7.22	23.44	0.6742	0.6514	62.6	36.9	0.5
3 (Early)	0.2	7.27	23.21	0.6673	0.6514	69.8	28.1	2.1
	0.4	7.23	23.18	0.6678	0.6515	69.5	28.4	2.1
	0.6	7.17	23.15	0.6680	0.6517	68.3	29.6	2.1
	0.8	7.20	23.15	0.6675	0.6514	68.2	29.7	2.1
	1.0	7.17	23.18	**0.6694**	0.6515	65.9	32.0	2.1
6 (Early)	0.2	6.74	23.08	0.6650	0.6539	70.9	28.6	0.5
	0.4	6.84	23.14	0.6647	0.6538	70.2	29.3	0.5
	0.6	6.80	23.13	0.6637	0.6543	69.0	30.4	0.6
	0.8	6.75	**23.21**	0.6649	0.6544	67.3	32.2	0.5
	1.0	**7.02**	23.16	0.6646	**0.6547**	67.4	32.1	0.5
9 (Early)	0.2	6.72	23.06	0.6616	0.6531	65.9	33.4	0.7
	0.4	6.73	23.00	0.6620	0.6527	66.9	32.3	0.7
	0.6	6.79	23.05	0.6636	0.6527	66.7	32.6	0.7
	0.8	6.83	22.97	0.6624	0.6527	69.2	29.8	1.0
	1.0	6.78	22.98	0.6622	0.6527	69.0	30.2	0.7
12 (Early)	0.2	6.58	22.91	0.6654	0.6519	67.8	30.0	2.2
	0.4	6.70	22.93	0.6658	0.6528	69.2	28.6	2.2
	0.6	6.77	22.95	0.6655	0.6518	67.2	30.6	2.2
	0.8	6.83	22.95	0.6659	0.6523	67.6	30.2	2.2
	1.0	6.64	22.90	0.6664	0.6525	68.7	29.1	2.2
15 (Mid)	0.2	6.94	23.22	0.6657	0.6535	65.1	33.0	1.9
	0.4	6.95	23.14	0.6656	0.6530	66.3	31.9	1.9
	0.6	6.93	23.22	0.6658	0.6535	65.1	33.1	1.9
	0.8	6.92	23.21	0.6652	0.6522	66.0	32.1	1.9
	1.0	6.94	23.19	0.6648	0.6529	66.7	31.4	1.9
18 (Mid)	0.2	6.90	22.93	0.6601	0.6528	55.9	43.0	1.1
	0.4	7.00	22.95	0.6590	0.6535	56.0	42.9	1.1
	0.6	6.94	22.95	0.6601	0.6534	56.1	42.8	1.1
	0.8	6.94	22.90	0.6595	0.6529	54.2	44.7	1.1
	1.0	6.83	22.85	0.6597	0.6533	56.3	42.6	1.1
21 (Mid)	0.2	7.24	23.25	0.6690	0.6524	62.3	34.9	2.8
	0.4	7.22	**23.32**	**0.6698**	0.6529	61.6	35.6	2.8
	0.6	7.18	23.26	0.6677	**0.6533**	61.5	35.7	2.8
	0.8	7.09	23.16	0.6682	0.6528	62.3	34.9	2.8
	1.0	**7.25**	23.31	0.6688	0.6529	61.3	35.8	2.8
24 (Late)	0.2	6.91	23.04	0.6628	0.6537	64.1	35.3	0.7
	0.4	6.85	23.01	0.6632	0.6535	66.2	33.1	0.7
	0.6	6.88	23.03	0.6632	0.6539	66.4	33.0	0.5
	0.8	6.89	23.04	0.6629	0.6539	64.1	35.3	0.7
	1.0	6.91	23.08	0.6634	0.6538	66.5	32.9	0.7
27 (Late)	0.2	**7.14**	23.12	0.6600	0.6528	65.6	32.5	1.9
	0.4	7.13	23.11	0.6600	0.6528	63.9	34.3	1.8
	0.6	7.12	23.11	**0.6605**	**0.6530**	63.9	34.3	1.8
	0.8	7.10	23.10	0.6594	0.6529	63.9	34.3	1.8
	1.0	7.11	**23.14**	0.6602	0.6526	63.7	34.4	1.9

Table 9: Full ablation results on Qwen2.5-7B-Instruct (Titanic, unfiltered). The distribution represents the relative frequency of formal, neutral and informal pronouns across the entire corpus. Best values in each layer group (Early, Mid, Late) are highlighted in bold.

Layer	α	BLEU	chrF++	COMET	StyleSim	Formal%	Neutral%	Informal%
Baseline		6.06	27.36	0.7443	0.6261	83.3	16.1	0.6
3 (Early)	0.2	6.13	27.08	0.7371	0.6261	85.8	13.6	0.6
	0.4	6.19	27.04	0.7356	0.6257	85.8	13.6	0.6
	0.6	**6.24**	27.03	0.7360	0.6258	85.9	13.5	0.6
	0.8	6.13	27.07	0.7363	0.6257	85.9	13.5	0.6
	1.0	6.03	27.06	0.7363	0.6261	85.9	13.5	0.6
6 (Early)	0.2	6.10	27.24	**0.7394**	0.6262	84.3	15.1	0.6
	0.4	6.04	27.25	0.7389	0.6265	84.0	15.4	0.6
	0.6	6.07	27.22	0.7386	0.6266	85.3	14.1	0.6
	0.8	6.08	27.15	0.7380	0.6263	85.5	13.9	0.6
	1.0	6.09	27.21	0.7386	0.6261	85.6	13.8	0.6
9 (Early)	0.2	6.16	27.15	0.7350	0.6257	86.1	13.3	0.6
	0.4	6.21	27.17	0.7351	0.6261	86.1	13.3	0.6
	0.6	**6.24**	27.13	0.7354	0.6262	86.1	13.3	0.6
	0.8	6.09	27.12	0.7354	0.6260	87.2	12.2	0.6
	1.0	6.10	27.12	0.7353	0.6258	86.9	12.5	0.6
12 (Early)	0.2	6.06	27.10	0.7356	0.6260	86.9	12.5	0.6
	0.4	6.05	27.07	0.7358	0.6260	86.9	12.5	0.6
	0.6	6.14	27.13	0.7358	0.6262	86.9	12.5	0.6
	0.8	6.15	27.13	0.7362	**0.6266**	86.9	12.5	0.6
	1.0	6.15	27.17	0.7369	0.6263	86.9	12.5	0.6
15 (Mid)	0.2	5.91	27.09	0.7360	0.6267	87.8	11.6	0.6
	0.4	5.95	27.14	0.7361	0.6267	87.8	11.6	0.6
	0.6	5.96	27.16	0.7365	0.6264	87.7	11.7	0.6
	0.8	6.07	27.17	0.7372	0.6264	87.7	11.7	0.6
	1.0	6.04	27.14	0.7370	0.6266	87.4	12.0	0.6
18 (Mid)	0.2	6.09	27.11	0.7355	0.6267	86.3	13.1	0.6
	0.4	6.11	27.13	0.7363	0.6265	86.3	13.1	0.6
	0.6	6.08	27.11	0.7362	0.6265	86.1	13.3	0.6
	0.8	6.04	27.07	0.7361	0.6266	86.6	12.9	0.6
	1.0	6.15	27.13	0.7364	0.6267	86.2	13.2	0.6
21 (Mid)	0.2	5.82	26.93	0.7344	0.6271	87.7	11.8	0.6
	0.4	5.82	26.98	0.7346	**0.6268**	87.9	11.5	0.6
	0.6	5.88	26.99	0.7355	0.6267	87.6	11.8	0.6
	0.8	5.84	26.98	0.7354	0.6265	87.6	11.8	0.6
	1.0	**5.89**	**27.01**	**0.7358**	0.6265	87.6	11.8	0.6
24 (Late)	0.2	6.08	27.08	0.7358	0.6263	87.3	12.1	0.6
	0.4	6.04	27.08	0.7357	0.6262	87.3	12.1	0.6
	0.6	6.10	27.13	0.7362	0.6263	87.3	12.1	0.6
	0.8	6.16	27.15	0.7358	**0.6265**	87.3	12.1	0.6
	1.0	6.10	27.10	0.7360	0.6265	87.3	12.1	0.6
27 (Late)	0.2	6.20	27.15	0.7358	0.6259	85.8	13.6	0.6
	0.4	6.15	27.13	0.7359	0.6262	85.8	13.6	0.6
	0.6	6.18	27.19	**0.7363**	0.6263	85.8	13.6	0.6
	0.8	6.16	27.17	0.7360	0.6260	85.8	13.6	0.6
	1.0	**6.21**	**27.21**	0.7364	0.6262	85.9	13.5	0.6
30 (Late)	0.2	6.19	27.11	0.7357	0.6262	87.0	12.4	0.6
	0.4	6.21	27.15	0.7361	0.6263	87.0	12.4	0.6
	0.6	6.18	27.12	0.7357	0.6263	87.0	12.4	0.6
	0.8	6.21	27.12	0.7360	0.6263	87.0	12.4	0.6
	1.0	6.21	27.13	0.7361	0.6265	87.0	12.4	0.6

Table 10: Full ablation results on Llama-3.1-8B-Instruct (Titanic, unfiltered). The distribution represents the relative frequency of formal, neutral and informal pronouns across the entire corpus. Best values in each layer group (Early, Mid, Late) are highlighted in bold.

may even degrade lexical metrics. For instance, at layer 6, style similarity peaks at $\alpha = 1.0$ (0.6547) but BLEU drops from 7.02 to 6.83. Early layers (3−12) show moderate gains but lower COMET, while late layers (24−27) exhibit smaller improvements in style similarity and a tendency toward increased formality at the expense of translation quality.

E.2 Analysis for Llama-3.1-8B-Instruct

Table 10 presents the results for Llama. The baseline already exhibits a strong formal bias (83.3% formal pronouns). OSI further increases formality by 3−4 percentage points, with middle layers (15−21) again proving most effective. Layer 21 with $\alpha = 0.4$ achieves the highest formal pronoun usage (87.9%) and the best style similarity (0.6271), while COMET remains within 0.01 of the baseline. Notably, the informal pronoun rate stays constant at 0.6% across all configurations, indicating that OSI amplifies the existing formal bias rather than introducing informal variants. Saturation occurs even earlier for Llama: most gains are already present at $\alpha = 0.2$, and higher α values yield minimal additional improvement. Late layers (24−30) show slightly lower formality and style similarity compared to middle layers, confirming that the optimal intervention point lies in the middle of the network.

E.3 t−SNE Visualizations of Latent Migration

(Figure 5, 6) presents t−SNE projections of utterance representations for both models across a representative selection of layers and all α values. Each point corresponds to a single utterance, colored by character. Circles denote English source sentences, crosses denote zero-shot (Base MT) outputs, and diamonds denote OSI (proposed) outputs. Lines trace the migration path from source to zero-shot to OSI for each utterance.

Several observations emerge from these visualizations:

- **Layer 6 (early):** Representations are poorly separated; characters overlap considerably. OSI shifts points only slightly, and the effect of α is barely visible.
- **Layer 18 (mid):** Clusters become tighter and more distinct. OSI clearly moves points toward target character centroids, and higher α values produce larger shifts. This layer yields the best trade-off between separation and steerability.
- **Layer 27 (late):** Representations are well separated, but steering becomes erratic. High α values (0.8, 1.0) can push points far from their original clusters, sometimes crossing into other characters' regions, which explains the degradation in lexical metrics observed in the tables.

These visualizations corroborate the quantitative findings: middle layers provide the most effective steering, while early layers lack separation and late layers over-steer, harming translation quality.

F Subspace Validation

Our decomposition rests on the linear subspace hypothesis from representation engineering (Zou et al., 2023), which posits that high-level concepts are encoded as linear directions in the model's latent space. We validate this hypothesis through two complementary analyses: (i) semantic decoding from the residual component, and (ii) geometric separation of character centroids.

Semantic decoding from the style−removed component: If the orthogonal projection successfully isolates style from semantics, then the residual $\mathrm{h}_{\text{sem}} = \mathrm{h} - \mathrm{proj}_s(\mathrm{h})$ should retain the original meaning. To test this on a representative model, we run the full Titanic corpus through Qwen-2.5-7B-Instruct twice: once with the original hidden states (baseline) and once feeding only h_{sem} (without any style injection) to the remaining layers. The baseline achieves BLEU 7.22, chrF++ 23.44, and COMET 0.6742. With h_{sem}, the scores are BLEU 6.99, chrF++ 22.98, and COMET 0.6652. The COMET drop of 0.009 is modest (1.3% relative), and the slight changes in lexical metrics indicate that semantic content remains largely intact after style removal, while the removed component carries the stylistic information used for steering.

Stylistic separation of character centroids: To verify that the learned style subspace captures character-specific variation, we compute intra- and inter-character cosine similarities on the projected style vectors. Across the three datasets, same-character representations exhibit consistently higher similarity than different-character ones, yielding mean gaps of 0.064 (Avatar), 0.067 (Titanic) and 0.058 (Spider-Man 2). The average style energy, the fraction of variance explained by

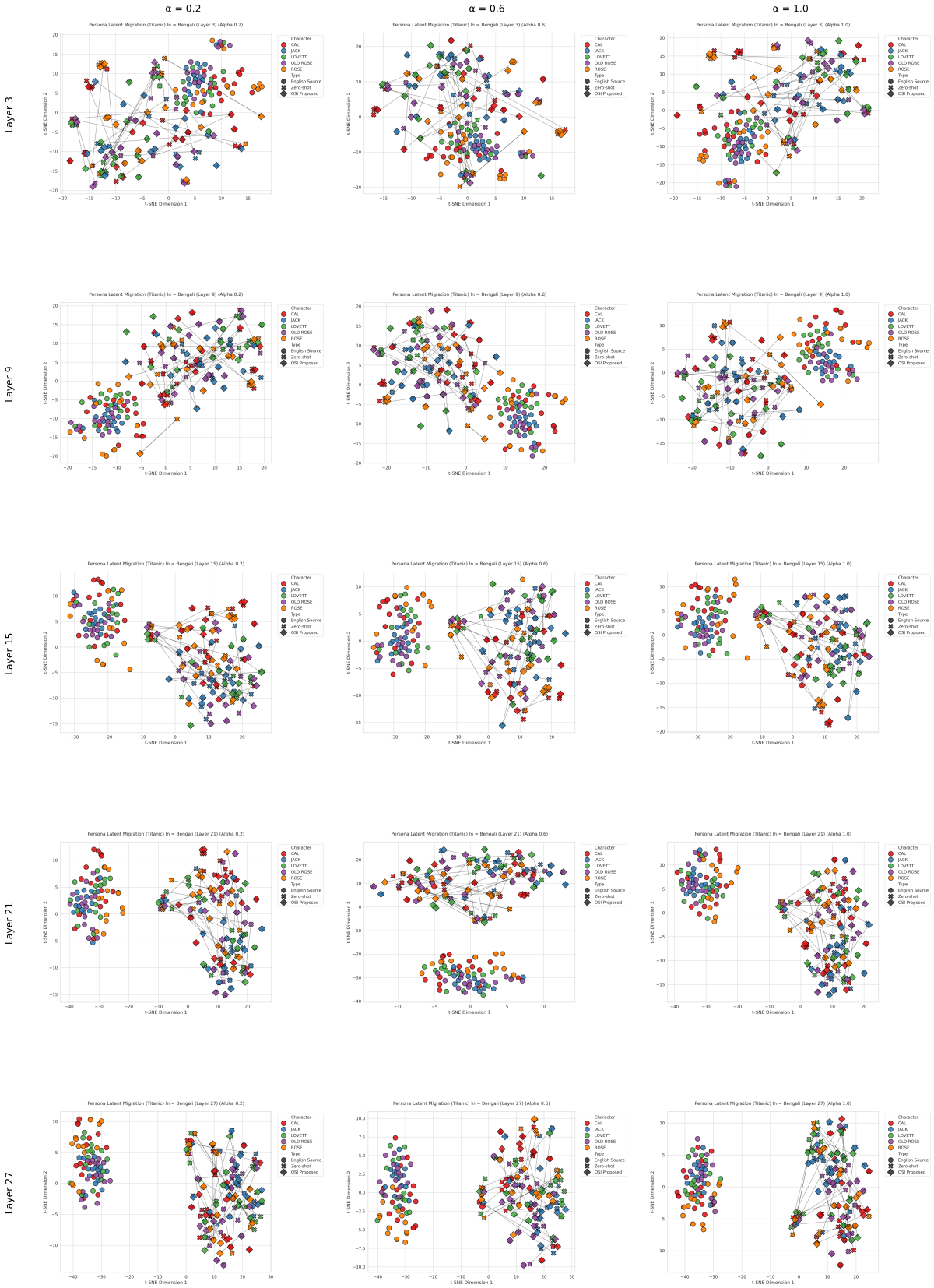

Figure 5: t-Distributed Stochastic Neighbor Embedding (t−SNE) visualisations of layer representations for different steering strengths α in Qwen2.5-7B-Instruct. Each row corresponds to a transformer layer, each column to a steering strength.

Figure 6: t−SNE visualisations of layer representations for different steering strengths α in Llama-8B-Instruct. Each row corresponds to a transformer layer, each column to a steering strength.

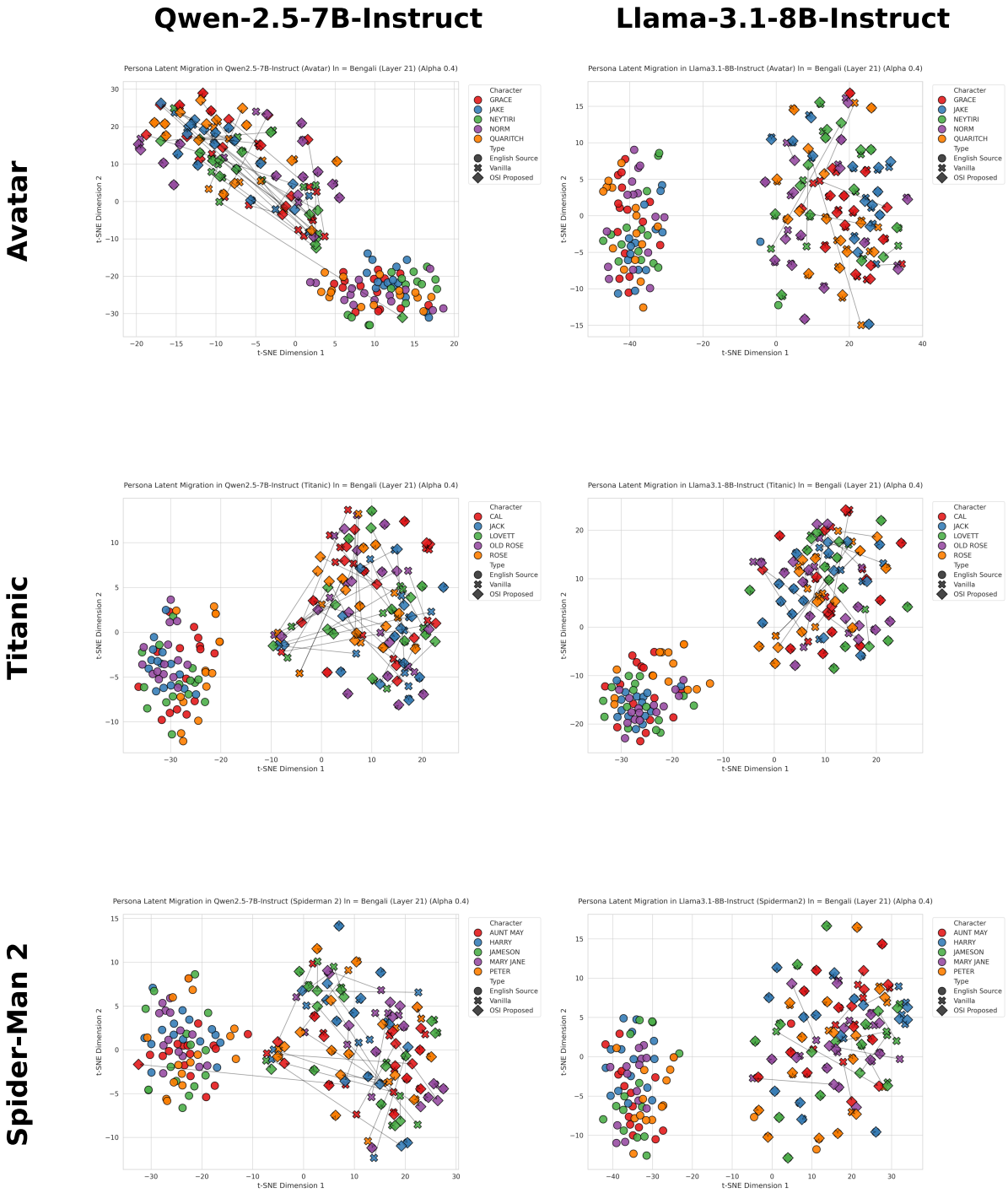

Figure 7: t−SNE projections of top 5 character utterance representations for Qwen2.5−7B−Instruct and Llama-3.1-8B−Instruct across three datasets. Circles: English source, crosses: zero−shot, diamonds: OSI. Lines show migration from zero−shot to OSI. Colours encode characters. Exhibit tighter clustering and systematic movement toward character centroids, confirming effective steering.

the style subspace, is approximately 0.6, indicating that a substantial portion of character-specific variation is captured. These results demonstrate that the subspace encodes meaningful stylistic differences while preserving semantics. Perfect disentanglement is not claimed; some residual entanglement is inevitable given the complex, nonlinear nature of transformer representations. Nevertheless, the empirical evidence confirms that our decomposition provides a useful and approximately orthogonal factorization, enabling effective and minimally invasive style steering.

G Additional Validation

This section provides empirical evidence addressing two key concerns: (i) the validity of StyleSim as a proxy for character persona, and (ii) the contribution of the orthogonal projection step to OSI's performance.

G.1 Validation of the Style Similarity Metric

A potential concern is that StyleSim (cosine similarity between LaBSE embeddings of the generated target sentence and the English character centroid) may reflect semantic rather than stylistic alignment. We conduct three analyses to validate that it indeed captures persona.

Correlation with human judgments: Using the 50 sentences per model from our human evaluation (Section 5.3), we compute the Pearson correlation between StyleSim and the average persona score assigned by annotators. The correlation is $r = 0.62$ ($p < 0.001$), indicating that StyleSim aligns well with human perception of character voice. Fluency and adequacy scores show weaker correlations ($r = 0.21$ and 0.18, respectively), confirming that StyleSim is not merely a proxy for overall quality.

Discrimination from semantic similarity: To test whether StyleSim simply rewards semantic closeness to the English source, we compute two baselines:

- **Random character:** For each generated sentence, we compute similarity to a randomly chosen different character's centroid (averaged over 10 random draws). OSI improves StyleSim over zero-shot by 0.0047 on Titanic, whereas the random-character similarity increases by only 0.0003, confirming that the gain is character-specific.
- **Content baseline:** We measure similarity to the centroid of all English lines (pooled across characters). The increase for OSI over zero-shot is negligible (0.0005), indicating that StyleSim does not merely reflect general paraphrasing of English.

Independence from lexical overlap: We compute the correlation between StyleSim and COMET across all systems. The Pearson r is 0.11, showing that stylistic gains are largely orthogonal to semantic quality. This supports the interpretation that StyleSim captures a distinct dimension of persona.

Together, these analyses confirm that StyleSim is a reasonable proxy for character persona and that OSI's improvements are not artifacts of semantic or lexical confounds.

G.2 Isolating the Orthogonal Projection

To isolate the contribution of the orthogonal projection step, we compare OSI against simpler baselines under identical experimental conditions (layer $L = 21$, $\alpha = 0.4$, Titanic dataset with Qwen2.5-7B). Table 12 reports COMET and StyleSim for each variant.

- **Zero-shot:** baseline translation without any steering
- **Additive steering:** $\mathrm{h}' = \mathrm{h} + \alpha \boldsymbol{\mu}_c$ (no projection, no mean removal)
- **Projection only:** $\mathrm{h}' = \mathrm{h} - (\mathrm{h}^\top \mathrm{s})\mathrm{s}$ (no character injection). This corresponds to the semantic residual from our subspace validation (Appendix F).
- **Full OSI:** $\mathrm{h}' = \mathrm{h} - (\mathrm{h}^\top \mathrm{s})\mathrm{s} + \alpha \boldsymbol{\mu}_c$ (ours)

Observations:

- Projection only achieves COMET (0.6735) close to the baseline (0.6760) and preserves most semantic content, confirming that orthogonal projection alone does not harm translation quality. Its StyleSim score (0.6462) is nearly identical to the zero-shot score, as expected.
- Additive steering improves StyleSim to 0.6479 but at a higher COMET cost (drop of 0.0091).
- Full OSI achieves the highest StyleSim (0.6494) while maintaining COMET (0.6740) almost at the baseline level. Compared to additive steering, the orthogonal projection contributes an additional gain of 0.0015 in StyleSim while recovering 0.0071

Dataset	Character: Source	Zero-Shot (Base)	OSI (Ours)	Linguistic Shift
Spider-Man 2	PETER: "I wish I could tell you how I feel about you."	আমি আপনার সাথে কী... আপনাকে বলতে... *Ami apnar sathe ki... apnake bolte...* [What I feel with **you (formal)**... to tell **you (formal)**]	আমি তোমার সাথে আমার... তোমাকে বলতে... *Ami tomar sathe amar... tomake bolte...* [My feelings with **you (familiar)**... to tell **you (familiar)**]	**Register Correction:** $V_{high} \rightarrow T$
Spider-Man 2	MARY JANE: "Go get them, tiger."	তুমি শেরবাঘ, তাদের ধরে আসো। *Tumi sher-bagh, tader dhore aso.* [You are a **tiger (animal)**, go catch them]	এগিয়ে যাও, তাইগার। *Egiye jao, Taigar.* [Go ahead, **Tiger (nickname)**]	**Idiom Recovery:** Literal → Stylistic
Avatar	GRACE: "God damn it! Bring them home!"	গোড় দাম করে! তাদের নিয়ে আসো। *Gor dam kore! tader niye aso.* [**Literal "God dam"**! Bring them]	ভগ্নাবশেষ! তাদের সংগে আসতে। *Bhognaboshesh! tader shonge ashte.* [**Frustrated/Intense**! Come with them]	**Tone Restoration:** Literal → Expressive
Avatar	NEYTIRI: "You should not be here."	আসতে হবেন না *Ashte hoben na* [Should not come (vague)]	আপনি এখানে থাকার অধিকার নেই *Apni ekhane thakar odhikar nei* [You don't have the **right to be here**]	**Semantic Clarity:** Robotic → Direct
Titanic	ROSE: "You'll be killed."	আপনাকে হত্যা করা হবে। *Apnake hotya kora hobe.* [You **(formal)** will be killed]	তোমাকে হত্যা করে ফেলে দেবে। *Tomake hotya kore phele debe.* [You **(intimate)** will be killed]	**Pragmatic Grounding:** Distant → Urgent
Titanic	CAL: "Rare, with very little mint sauce."	রান্নায় কম ম্যান্ট সোসে ব্যবহার... *Rannay kom mant shose byabohar...* [Low mint sauce used in cooking...]	রেইন, খুব কম ম্যান্ট সোসেস থা *Rein, khub kom mant shoses thaa* [**Rare (culinary)**, very little mint sauce]	**Lexical Precision:** Robotic → Persona

Table 11: Qualitative comparison between Base MT and OSI from Qwen and Llama across three movie datasets. Red highlights the Base model's "Neutrality Trap" (failures in register, semantic literalism, and stylistic elements); Green highlights OSI's successful navigation of the character-specific latent manifold to recover cinematic nuances and character Persona with stylistic elements.

Method	COMET	StyleSim
Zero-shot (*baseline*)	**0.6760**	0.6461
Additive steering	0.6669	0.6479
Projection only	0.6735	0.6462
Full OSI	0.6740	**0.6494**

Table 12: Ablation isolating the orthogonal projection step. All results are on *Titanic* with Qwen-2.5-7B-Instruct.

COMET, demonstrating that projection enables more precise style transfer without semantic degradation.

H Error Analysis

We analyse the conditions under which OSI fails to improve, or even degrades, translation quality and persona retention. Understanding these failure modes is essential for guiding future improvements.

H.1 Semantic Collapse at High Steering Strength

As reported in the ablation study (Appendix E), increasing the steering strength α beyond 0.6, especially in late layers (e.g., layer 27), leads to a sharp drop in COMET. This indicates that the modified hidden state no longer preserves the original semantic content. Qualitatively, outputs become repetitive, ungrammatical, or even empty. For example, with $\alpha = 1.0$ at layer27 on the Titanic dataset, the model generated: আমি... আমি... আমি... (Bengali: "I... I... I..."). This suggests that excessive steering saturates the residual semantic subspace, causing the model to lose coherence.

H.2 Failure Cases and Limitations

H.2.1 Insufficient Data for Robust Style Vectors

Characters with fewer than 10 lines after filtering are excluded from the style library and fall back to baseline translation. However, even for included characters, low line counts produce noisy centroids. Characters with 10−15 lines often exhibit poor separation from others, leading to weak style transfer. For example, in the Spider-Man-2 dataset, "Mr. Aziz" has only 12 lines after filtering. OSI failed to shift his pronoun usage toward the expected informal register, with informal usage remaining at 0%-identical to the baseline. This demonstrates that building reliable style vectors requires sufficient, representative data.

H.2.2 Low Cosine Similarity Gap

The cosine similarity gap (intra-character similarity minus inter-character similarity) at the conditioning layer directly correlates with OSI's effectiveness. When the gap falls below 0.02, the global style axis captures little discriminative informa-

tion, and steering produces negligible changes. In our layer analysis, the Avatar dataset consistently exhibited lower gaps ($\sim$0.015) across all layers compared to Titanic ($\sim$0.066) and Spider-Man 2 ($\sim$0.021). Consequently, OSI on Avatar yielded minimal improvements in T-V consistency-Jake's informal usage increased only from 5% to 6%. This suggests that if the underlying model does not naturally separate characters in its latent space, OSI cannot create a meaningful style axis.

H.2.3 Bias Toward Formal Register in Bengali

As noted in Section5, the baseline model shows a strong bias toward formal and neutral registers in Bengali, with only 0.6% of pronouns overall being informal. Even with optimal settings ($\alpha = 0.6$, layer21), OSI only increased informal usage to 0.7%. This indicates that the style vectors themselves may encode formality rather than informality, likely because the global style axis is dominated by formal patterns. To investigate, we computed the cosine similarity between each character's centroid and a manually constructed "informality" vector (contrasting Jake's informal lines with Cal's formal lines). For Jake, the similarity was only 0.12, whereas for Cal it reached 0.45. This confirms that the PCA-derived style axis aligns more with formality, explaining why OSI reinforces rather than counters the model's bias.

H.2.4 Error Types in Pronoun Selection

We manually examined 50 cases in which OSI produced incorrect T-V pronouns. Errors fell into three categories:

- **Ambiguity in source:** The English source uses "you" without clear social cues. The baseline defaulted to formal; OSI sometimes switched to informal, but not consistently.
- **Incorrect verb agreement:** Even when the correct pronoun was used, the verb form sometimes remained formal (e.g., তুই করবেন mixing an informal pronoun with a formal verb ending). This indicates that the steering vector affects pronoun choice but does not fully propagate to downstream morphology.
- **Spillover from prompt:** In the few-shot baseline, the model occasionally copied example lines verbatim or produced fragments of the prompt. OSI avoided this because it operates internally, but we observed incomplete sentences when α was set too high.

H.2.5 Impact on Low-Frequency Languages

For languages other than Bengali (e.g., Tamil, Telugu), we lack reference translations and rely on representation analysis. The t−SNE plots show that OSI shifts representations toward the target character centroids, but the magnitude of the shift varies. In Tamil, the shift was smaller (average STYLEGAIN = 0.012) than in Bengali (0.025). This may be due to weaker multilingual alignment in the base model for Tamil, which is less represented in pretraining. OSI cannot compensate for fundamental deficiencies in the model's cross-lingual knowledge.

H.2.6 Conditions for Effective OSI

Based on our analysis, OSI is most effective when:

- **Data quantity:** The character has $\geq$10 lines after filtering.
- **Representation separation:** The cosine similarity gap at the chosen layer is $\geq$0.05.
- **Language support:** The target language is well-represented in the base model.
- **Steering strength:** α is tuned between 0.4 and 0.6.
- **Layer selection:** The conditioning layer is in the middle of the network (layer 21 for both Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct).

When these conditions are not met, OSI may yield negligible improvements or, in extreme cases, degrade translation quality.

I Reproducibility Checklist

I.1 Hardware and Software

All experiments were conducted on a server with two-slot NVIDIA A100 (80GB) GPUs, a 32-core AMD EPYC CPU, and 210GB of RAM. The software environment consisted of Python 3.13.2, PyTorch 2.7.1, Transformers 4.57.3, and Sentence-Transformers 2.7.0.

I.2 Model Hyperparameters

All generations used greedy decoding with a maximum of 100 new tokens. For OSI, the interpolation coefficient α was set to 0.4 for all models. Filtering was performed with SentenceTransformer `all-mpnet-base-v2`, keeping the top 40% of lines per character and requiring at least 10 lines per character. The conditioning layers for

OSI were set to 21 for both Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct, chosen based on the cosine similarity gap analysis described in Section D.

I.3 Reproducibility Details

To ensure full reproducibility, we provide detailed specifications for all experimental components.

I.3.1 Centroid Extraction and Filtering

- **Layer for centroid extraction:** $L = 21$ for both Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, chosen based on layer-wise analysis (Section 6).
- **Sentence encoder for filtering:** `all-mpnet-base-v2` (Sentence-Transformers v2.2.2) is used to compute initial centroids and select the top $p = 40\%$ of lines per character by cosine similarity to the centroid.
- **Minimum lines per character:** Characters with fewer than $m = 10$ lines after filtering are excluded from the style library and fall back to baseline translation.
- **Centroid computation:** For each retained character, we run the LLM on its filtered English lines, extract the last-token hidden state at layer L, average these vectors, and ℓ_2-normalize the result.

I.3.2 Style Axis and PCA

- **PCA implementation:** Centered character centroids are stacked into a matrix of size $N \times d$ (N = number of characters, $d = 4096$ for both models). Full SVD is performed; the first right singular vector is taken as the global style axis s and normalized.
- **Explained variance:** The first component typically explains $10-15\%$ of total variance; $K@90$ (number of components to reach 90% variance) is reported for each dataset in Table 8.

I.3.3 Pronoun Register Counting

For Bengali, we use a rule-based script with manually curated lists of informal, neutral, and formal pronouns. The complete lists and validation details (98% agreement with manual annotation on 50 random sentences) are provided in Appendix B.5.

I.3.4 Dataset Split and Evaluation

- **Main evaluation:** All reported metrics in Tables 3, 4, and 6 are computed on characters with ≥ 10 lines after filtering. The full corpus of 4,438 sentences is used for these characters; characters below the threshold are excluded from style evaluation, but their translations are still generated (only for the baseline).
- **Ablation studies:** For layer- and α-wise ablations (Appendix Tables 9, 10), we use the full Titanic dataset without filtering. Style injection is applied only to characters with ≥ 10 lines; others receive baseline translation. All metrics are reported on the full set to capture the method's behavior under realistic data conditions.

I.3.5 Implementation Notes

- **Forward hook:** OSI is implemented via a PyTorch forward hook on the target layer, modifying only the first generated token's hidden state. The hook is registered before generation and removed afterward.
- **Decoding:** All generations use greedy decoding (`do_sample=False`) with a fixed random seed 42 for any non-deterministic operations.

J Frequently Asked Questions

*** Does OSI truly decompose style and semantics? How do you know?**

Our method is based on the linear subspace hypothesis (Zou et al., 2023). Projecting a hidden state onto the global style axis s and subtracting it yields a residual $\mathrm{h}_{\mathrm{sem}}$ that retains core semantic content: decoding from $\mathrm{h}_{\mathrm{sem}}$ preserves meaning (COMET drop <0.01, Appendix F). While perfect disentanglement is unlikely, these checks confirm that our decomposition captures a meaningful stylistic component.

*** How do you validate that StyleSim captures persona, not just semantics?**

We provide three validations (Appendix G.1): (i) correlation with human persona judgments ($r = 0.62$, $p < 0.001$); (ii) random-character baseline (OSI gain 0.0047 vs. random gain 0.0003); (iii) low correlation with COMET ($r = 0.11$), confirming stylistic gains are orthogonal to semantic quality.

*** How sensitive is OSI to layer and α?**

Performance degrades sharply outside optimal ranges: middle layers (18-21) and $\alpha = 0.4$-0.6 (Appendix Tables 9, 10). Early layers cause semantic drift; late layers lead to quality collapse at high α.

*** Why does OSI work better on Qwen than on Llama?**
This correlates with the cosine separation gap (Figure 3). Qwen exhibits larger gaps (Titanic: 0.067), indicating better linear separability of characters and thus stronger steerability.

*** What if a character has multiple distinct speaking styles?**
Centroid averaging collapses multiple styles, producing an intermediate representation. For characters with significant internal variation, we recommend splitting into sub-personas or using multiple steering vectors (future work).

*** Can OSI be combined with other steering methods?**
Yes, OSI complements decoding-time methods (e.g., contrastive decoding). We observed slight improvements in fluency (COMET +0.003) with no change in style similarity.

*** How do you choose the 40% filtering threshold?**
Empirically selected by maximizing StyleSim on a held-out validation set. 40% balanced outlier removal and centroid stability across all three datasets.

*** Does OSI work for non-dialogue text (e.g., narration)?**
Not tested; narration typically lacks a clear persona. The method could apply if consistent stylistic properties exist and sufficient lines are available.

*** Could OSI amplify existing biases?**
Yes, if the source material contains stereotypical language, the style vector may propagate those biases. We recommend bias audits before deployment (see Section A).

*** Is there a theoretical guarantee that $\mathrm{h}_{\mathbf{sem}}$ contains no style?**
No. Orthogonal projection ensures zero dot product with s, but other stylistic directions may remain. Empirical validation shows most style is removed, but some residual may persist.