Gravitation-Driven Semantic Alignment for Text Video Retrieval

Supplementary Material

7. From an Entangled Integral to Decoupled Factors: The Evolution of GraviAlign

Our approach to multimodal similarity originates from a profound physical intuition: semantic similarity is not a mere geometric distance but a dynamic interaction between two "probability clouds," each laden with uncertainty. This is the genesis of our **S**emantic **G**ravitational **I**nteraction (SGI) model. The initial formulation is captured by the SGI integral:

$$I_{\text{SGI}} = \iint p_v(x)p_t(y) \cdot \exp\left(\frac{S(x,y)}{T}\right) \mathrm{d}x\,\mathrm{d}y, \quad (11)$$

where $p_v(x)$ and $p_t(y)$ represent the probability distributions of the video and text semantics, and $S(x, y)$ is an interaction term inspired by gravitational potential, modulated by a temperature T via a Boltzmann-like weight.

However, the elegance of this formulation comes with a challenge: it is a computationally intractable entangled integral. The distributions and the interaction term are deeply coupled, precluding a closed-form solution. This impasse led us to a core insight: the SGI integral implicitly captures two distinct, orthogonal aspects of similarity:

1. **Geometric Overlap (Ψ_{Overlap}):** This factor measures the degree to which the two distributions intersect in the embedding space, reflecting a static, geometric alignment.
2. **Semantic Attraction ($\Phi_{\text{Attraction}}$):** This component models the attractive potential energy between the semantic 'centers of mass' (centroids) of the two distributions, reflecting a dynamic, potential-based alignment.

Our key methodological leap was to decouple these two fundamental factors. This led to our novel **GraviAlign** Dual-Factor framework, where the total alignment score is conceptualized in the log-domain as a sum of the two factors. This directly corresponds to the additive form $S_{\text{align}} = A+B+C$ used in our main paper, where Semantic Attraction is captured by Term A, and Geometric Overlap is jointly represented by Term B and Term C.

8. Theoretical Justification for the Geometric Score (Term B and Term C)

The geometric part of our alignment score, composed of Term B and Term C, is not an ad-hoc choice but a principled measure of alignment that emerges naturally from multiple theoretical perspectives, which we detail below.

8.1. Perspective 1: Zero-Difference Likelihood

A strong alignment implies that the random difference vector $Z_{\text{diff}} = Z_v - Z_t$ is highly likely to be zero. Let $Z_v \sim \mathcal{N}(\mu_v, \Sigma_v)$ and $Z_t \sim \mathcal{N}(\mu_t, \Sigma_t)$ be independent Gaussian embeddings. Their difference is also a Gaussian. For brevity, let $\Delta\mu = \mu_v - \mu_t$. Then:

$$Z_{\text{diff}} \sim \mathcal{N}(\Delta\mu, \Sigma_v + \Sigma_t) = \mathcal{N}(\Delta\mu, \Sigma_{\text{joint}}). \quad (12)$$

The probability density function (PDF) of a general multivariate Gaussian $z \sim \mathcal{N}(\mu, \Sigma)$ is:

$$p(z) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \exp\left(-\tfrac{1}{2}(z-\mu)^\top\Sigma^{-1}(z-\mu)\right) \quad (13)$$

We evaluate this PDF for Z_{diff} at the point $z = 0$. The quadratic form in the exponent becomes $(-\Delta\mu)^\top\Sigma_{\text{joint}}^{-1}(-\Delta\mu) = \Delta\mu^\top\Sigma_{\text{joint}}^{-1}\Delta\mu$. Thus:

$$p(Z_{\text{diff}} = 0) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_{\text{joint}}|}} \exp\left(-\tfrac{1}{2}\Delta\mu^\top\Sigma_{\text{joint}}^{-1}\Delta\mu\right) \quad (14)$$

Taking the logarithm gives us the log-likelihood:

$$\begin{aligned}\log p &= -\frac{1}{2}\log\left((2\pi)^D|\Sigma_{\text{joint}}|\right) - \frac{1}{2}\Delta\mu^\top\Sigma_{\text{joint}}^{-1}\Delta\mu \\ &= -\frac{1}{2}\Delta\mu^\top\Sigma_{\text{joint}}^{-1}\Delta\mu - \frac{1}{2}\log|\Sigma_{\text{joint}}| - \frac{D}{2}\log(2\pi).\end{aligned} \quad (15)$$

Ignoring the constant term, the two remaining components correspond exactly to Term B and Term C, respectively.

8.2. Perspective 2: Direct Geometric Overlap

This perspective directly measures the volume of intersection between the two PDFs via the overlap integral, Ψ_{Overlap}. For two Gaussian distributions, this integral has a known closed-form solution:

$$\Psi = \int_{\mathbb{R}^D} p_v(x)p_t(x)\,\mathrm{d}x = \mathcal{N}(\mu_v \mid \mu_t, \Sigma_v + \Sigma_t). \quad (16)$$

Using the same notation $\Delta\mu = \mu_v - \mu_t$, evaluating the PDF and taking the logarithm yields:

$$\begin{aligned}\log\Psi &= \log\left(\mathcal{N}(\mu_v \mid \mu_t, \Sigma_{\text{joint}})\right) \\ &= -\frac{1}{2}\log\left((2\pi)^D|\Sigma_{\text{joint}}|\right) - \frac{1}{2}\Delta\mu^\top\Sigma_{\text{joint}}^{-1}\Delta\mu \\ &= -\frac{1}{2}\Delta\mu^\top\Sigma_{\text{joint}}^{-1}\Delta\mu - \frac{1}{2}\log|\Sigma_{\text{joint}}| - \frac{D}{2}\log(2\pi).\end{aligned} \quad (17)$$

Again, this expression is equivalent to the sum of Term B, Term C, and a constant.