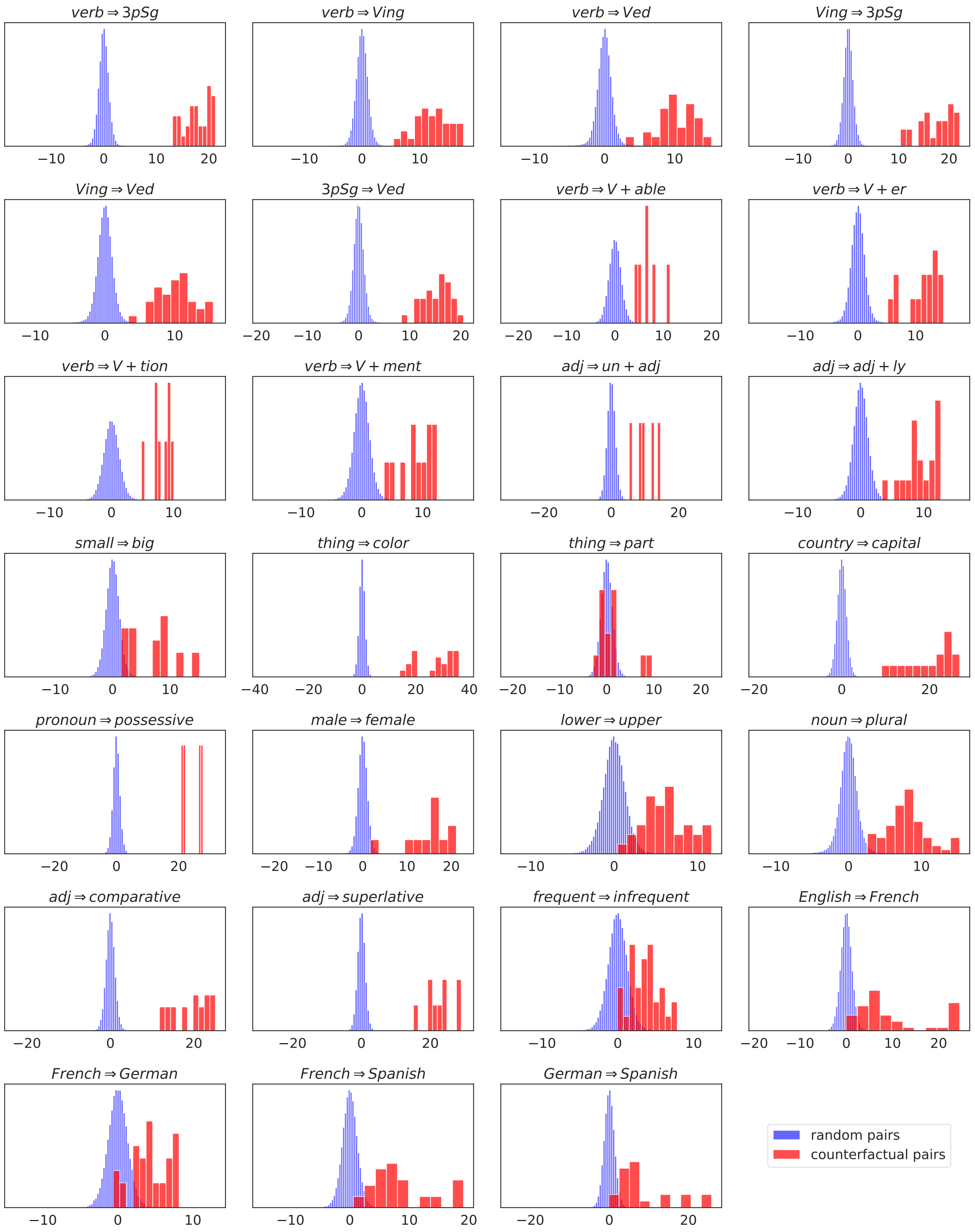

Figure 7. Histograms of the projections of the counterfactual pairs $\langle \bar{\gamma}_{W,(-i)}, \gamma(y_i(1)) - \gamma(y_i(0)) \rangle_{\mathrm{C}}$ (red), and the projections of the differences between 100K randomly sampled word pairs onto the estimated concept direction (blue). See Table 2 for details about each concept W (the title of each plot).

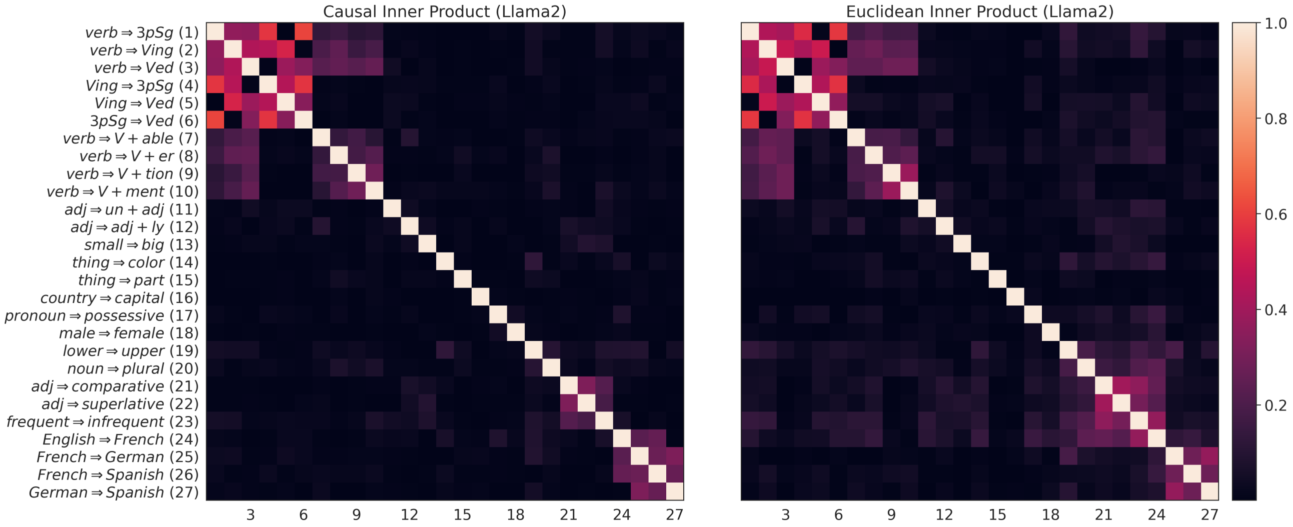

Figure 8. For the LLaMA-2-7B model, causally separable concepts are approximately orthogonal under the estimated causal inner product and, surprisingly, under the Euclidean inner product as well. The heatmaps show $|\langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle|$ for the estimated unembedding representations of each concept pair (W, Z). The plot on the left shows the estimated inner product based on (3.3), and the right plot represents the Euclidean inner product. The detail for the concepts is given in Table 2.

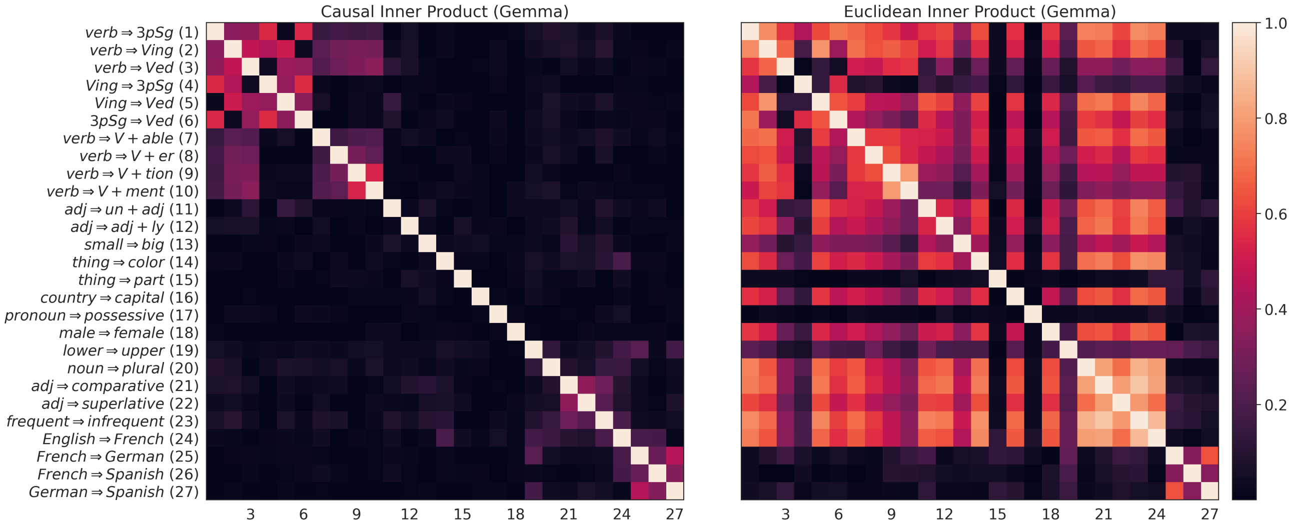

Figure 9. For the Gemma-2B model, causally separable concepts are approximately orthogonal under the estimated causal inner product; however, the Euclidean inner product does not capture semantics. The heatmaps show $|\langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle|$ for the estimated unembedding representations of each concept pair (W, Z). The plot on the left shows the estimated inner product based on (3.3), and the right plot represents the Euclidean inner product. The detail for the concepts is given in Table 2.

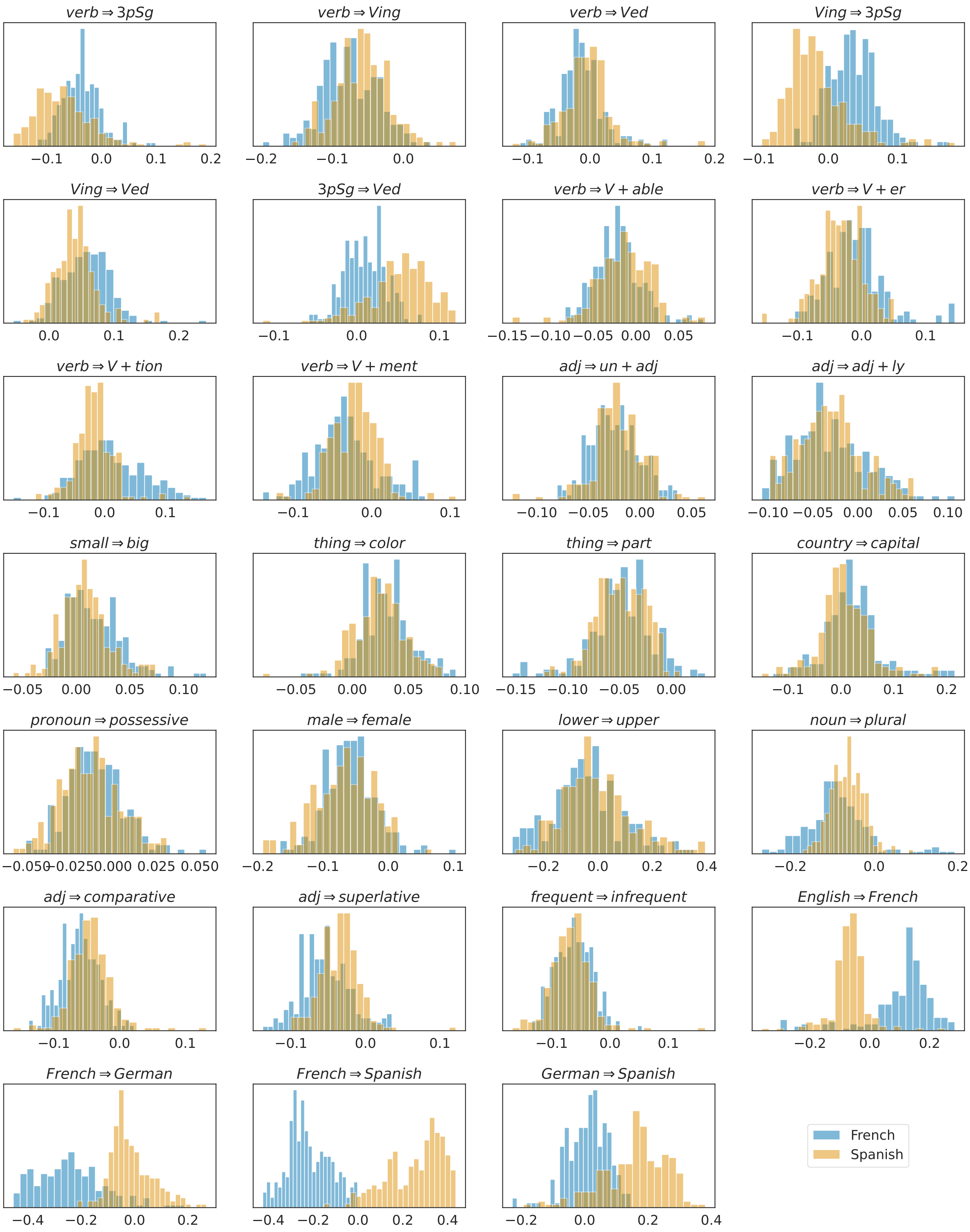

Figure 10. Histogram of $\bar{\gamma}_C^\top \lambda(x_j^{\mathrm{fr}})$ vs $\bar{\gamma}_C^\top \lambda(x_j^{\mathrm{es}})$ for all concepts C, where $\{x_j^{\mathrm{fr}}\}$ are random contexts from French Wikipedia, and $\{x_j^{\mathrm{es}}\}$ are random contexts from Spanish Wikipedia.