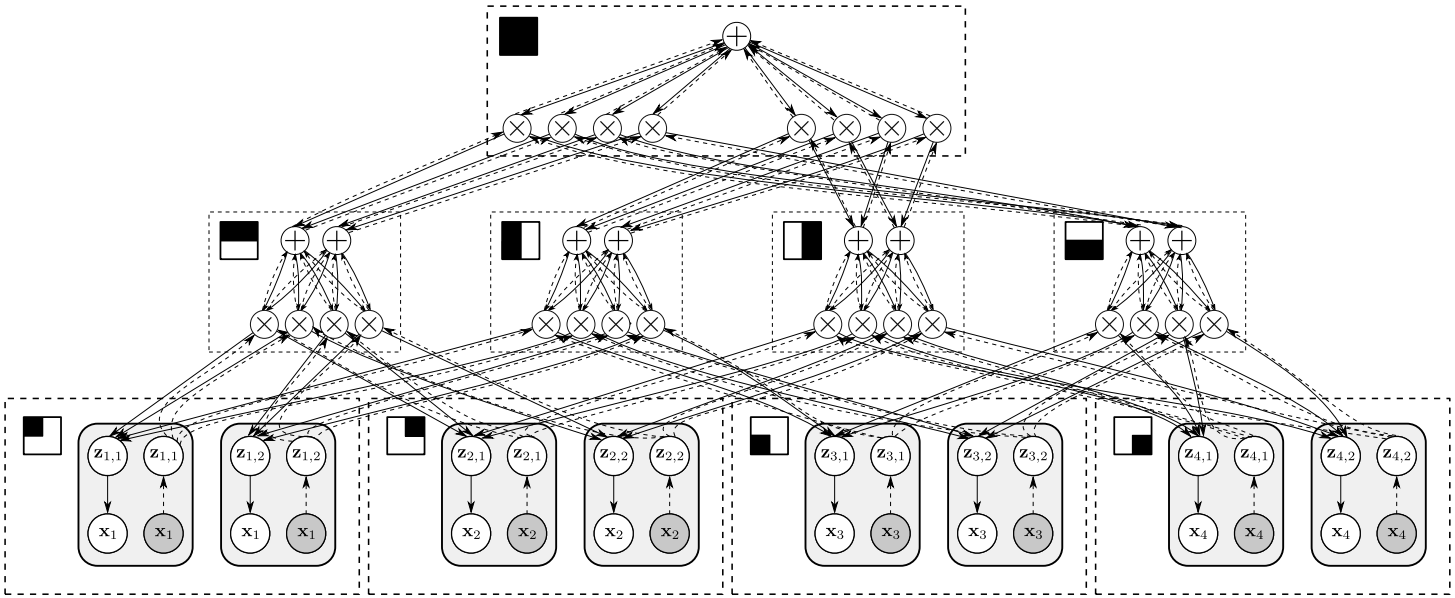

Figure 1: *(extended)* **Supplementary: Hierarchical Decompositional Mixtures of Variational Autoencoders**
In this paper, we use the so-called Poon-Domingos SPN structure [1], which recursively decomposes an image using axis-aligned splits with step size Δ. In this figure, to keep the illustration uncluttered, Δ is chosen such that only one split is performed in each axis: e.g. the overall image could be 32×32, in which case Δ would be 16. Also, for simplicity, we use only two sum nodes for each internal region, and two VAE leaves per leaf region.
In SPNs, every node is a distribution over a sub-set of the modeled random variables – this sub-set is called the *scope* of the respective node. In this figure, all nodes which share the same scope are surrounded by dashed boxes, and their respective scope is indicated by black regions in top left corner. The general principle in SPNs is to recursively arrange 3 simple techniques: i) pre-specified distributions, ii) factorized distributions (products), and iii) mixture distributions (weighted sums). This recursive arrangement is captured by an acyclic directed graph, where all leaves are pre-specified distributions and all internal nodes are either sums or products. **The main innovation in this paper, is to use expressive but intractable leaf distributions, namely VAEs.** This yields a hybrid model which combines exact SPN inference with approximate VAE inference.
The **generative process** of our model (depicted with solid lines) follows the usual SPN latent variable semantics [1, 2]: each sum node is interpreted as discrete latent variable, randomly selecting one of its children with probability given by the corresponding sum weight. By starting from the root, we follow the randomly selected child when meeting a sum node, and follow all children, when meeting a product node. In this way, the SPN probabilistically selects a combination of VAE experts, each of which generates its part of the scope. Note that the usual structural requirements in SPNs – completeness and decomposability [1] – guarantee that this generative process is well-defined: In particular, the scopes of selected VAEs will be a proper partitioning of the overall scope. The **inference process** (depicted with dashed lines) utilizes local inference networks for each VAE, computing local estimates for the evidence-lower-bound (ELBO). The local estimates are fed into the SPN, yielding an ELBO estimate for the entire model (see Theorem 2 in the main paper).
[1] Poon and Domingos, *Sum-product networks: A new deep architecture.* UAI, pp. 337–346, 2011.
[2] Peharz et al., *On the latent variable interpretation in sum-product networks.* TPAMI, 39(10):2030–2044, 2017.