tion $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ reconstructs the image from the latent, giving $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$, where $z \in \mathbb{R}^{h \times w \times c}$. Importantly, the encoder *downsamples* the image by a factor $f = H/h = W/w$, and we investigate different downsampling factors $f = 2^m$, with $m \in \mathbb{N}$.

In order to avoid arbitrarily high-variance latent spaces, we experiment with two different kinds of regularizations. The first variant, *KL-reg.*, imposes a slight KL-penalty towards a standard normal on the learned latent, similar to a VAE [46, 69], whereas *VQ-reg.* uses a vector quantization layer [96] within the decoder. This model can be interpreted as a VQGAN [23] but with the quantization layer absorbed by the decoder. Because our subsequent DM is designed to work with the two-dimensional structure of our learned latent space $z = \mathcal{E}(x)$, we can use relatively mild compression rates and achieve very good reconstructions. This is in contrast to previous works [23, 66], which relied on an arbitrary 1D ordering of the learned space z to model its distribution autoregressively and thereby ignored much of the inherent structure of z. Hence, our compression model preserves details of x better (see Tab. 8). The full objective and training details can be found in the supplement.

3.2. Latent Diffusion Models

Diffusion Models [82] are probabilistic models designed to learn a data distribution $p(x)$ by gradually denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov Chain of length T. For image synthesis, the most successful models [15,30,72] rely on a reweighted variant of the variational lower bound on $p(x)$, which mirrors denoising score-matching [85]. These models can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(x_t, t);\ t = 1 \dots T$, which are trained to predict a denoised variant of their input x_t, where x_t is a noisy version of the input x. The corresponding objective can be simplified to (Sec. B)

$$L_{DM} = \mathbb{E}_{x,\epsilon \sim \mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\Big]\,, \qquad (1)$$

with t uniformly sampled from $\{1, \dots, T\}$.

Generative Modeling of Latent Representations With our trained perceptual compression models consisting of $\mathcal{E}$ and $\mathcal{D}$, we now have access to an efficient, low-dimensional latent space in which high-frequency, imperceptible details are abstracted away. Compared to the high-dimensional pixel space, this space is more suitable for likelihood-based generative models, as they can now (i) focus on the important, semantic bits of the data and (ii) train in a lower dimensional, computationally much more efficient space.

Unlike previous work that relied on autoregressive, attention-based transformer models in a highly compressed, discrete latent space [23, 66, 103], we can take advantage of image-specific inductive biases that our model offers. This includes the ability to build the underlying UNet primarily from 2D convolutional layers, and further focusing the objective on the perceptually most relevant bits using the reweighted bound, which now reads

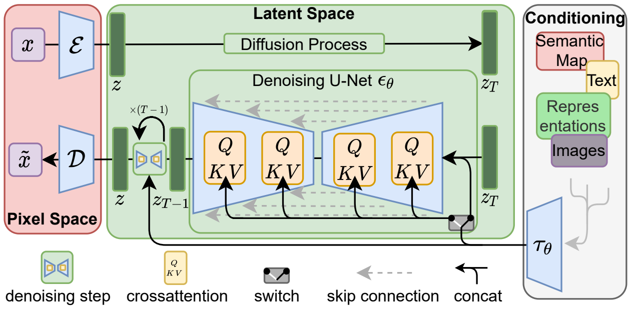

Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon \sim \mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\Big]\,. \qquad (2)$$

The neural backbone $\epsilon_\theta(\circ, t)$ of our model is realized as a time-conditional UNet [71]. Since the forward process is fixed, z_t can be efficiently obtained from $\mathcal{E}$ during training, and samples from $p(z)$ can be decoded to image space with a single pass through $\mathcal{D}$.

3.3. Conditioning Mechanisms

Similar to other types of generative models [56, 83], diffusion models are in principle capable of modeling conditional distributions of the form $p(z|y)$. This can be implemented with a conditional denoising autoencoder $\epsilon_\theta(z_t, t, y)$ and paves the way to controlling the synthesis process through inputs y such as text [68], semantic maps [33, 61] or other image-to-image translation tasks [34].

In the context of image synthesis, however, combining the generative power of DMs with other types of conditionings beyond class-labels [15] or blurred variants of the input image [72] is so far an under-explored area of research.

We turn DMs into more flexible conditional image generators by augmenting their underlying UNet backbone with the cross-attention mechanism [97], which is effective for learning attention-based models of various input modalities [35,36]. To pre-process y from various modalities (such as language prompts) we introduce a domain specific encoder τ_θ that projects y to an intermediate representation $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$, which is then mapped to the intermediate layers of the UNet via a cross-attention layer implementing $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$, with

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y).$$

Here, $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ denotes a (flattened) intermediate representation of the UNet implementing ϵ_θ and $W_V^{(i)} \in$