1. Length and depth of theoretical foundation (From the ground up?)
2. GPUs (Uni or GCP?)
3. Cross-Modal Autoencoder
   1. Latent-space arithmetic -> Book
   2. Feasible?
4. Visual Tokenizer (Vector-Quantized Knowledge Distillation) vs. Image Masking
5. Writing paper(s) next to the Thesis?