Are vision and language models able to reason across time? We evaluate the performance of vision and language models (VLMs) on the task of video question answering, with a particular focus on their temporal reasoning abilities. We probe the STAR video QA dataset on two VLMs with data perturbation methods of text and video inputs, and find that models are generally unable to identify the meaning of before and after in sequential questions. We then ask how a model can effectively learn these temporal relations, and design a new dataset drawn from videos and annotations from the Charades dataset. We create annotations that include targeted hard negative examples for the contrastive loss objective of one VLM, Merlot Reserve, such that the model must adapt to learn temporal relations. We further explore how to model fine-grained temporal relationships, and evaluate the benefits. We find that our approach shows promising signs of improvement on tasks that require temporal understanding, although it gains little sensitivity to temporal relations when probed.