Neural Networks 146 (2022) 238–255

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue on AI and Brain Science: Brain-inspired AI

Leveraging hierarchy in multimodal generative models for effective cross-modality inference

Miguel Vasco[a,*], Hang Yin[b], Francisco S. Melo[a], Ana Paiva[a]

[a] INESC-ID & Instituto Superior Técnico, University of Lisbon, Portugal
[b] Division of Robotics, Perception and Learning, EECS at KTH Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Article history:
Available online 24 November 2021

Keywords:
Multimodal representation learning
Cross-modality inference
Deep learning

ABSTRACT

This work addresses the problem of *cross-modality inference* (CMI), i.e., inferring missing data of unavailable perceptual modalities (e.g., sound) using data from available perceptual modalities (e.g., image). We overview single-modality variational autoencoder methods and discuss three problems of computational cross-modality inference, arising from recent developments in multimodal generative models. Inspired by neural mechanisms of human recognition, we contribute the Nexus model, a novel hierarchical generative model that can learn a multimodal representation of an arbitrary number of modalities in an unsupervised way. By exploiting hierarchical representation levels, Nexus is able to generate high-quality, coherent data of missing modalities given any subset of available modalities. To evaluate CMI in a natural scenario with a high number of modalities, we contribute the “Multimodal Handwritten Digit” (MHD) dataset, a novel benchmark dataset that combines image, motion, sound and label information from digit handwriting. We access the key role of hierarchy in enabling high-quality samples during cross-modality inference and discuss how a novel training scheme enables Nexus to learn a multimodal representation robust to missing modalities at test time. Our results show that Nexus outperforms current state-of-the-art multimodal generative models in regards to their cross-modality inference capabilities.

1. Introduction

Humans are provided with a complex cognitive framework that allows the creation of an internal representation of their external reality. This map of the external world is of a multimodal nature, composed from information provided by the environment (such as visual, auditory or somatosensory), captured by specific sensory organs. The biological mechanism behind the experience of such multimodal map remains an open question (Meyer & Damasio, 2009; Nanay, 2018). However, there is growing empirical evidence that suggests that perceptual information is processed hierarchically, in an unsupervised fashion, from lower-level sensory-specific cortices to higher-order multimodal cortices (Damasio, 1989; Meyer & Damasio, 2009).

The richness of such multimodal representation cannot be overstated. For example, reading lips in an environment with no sound induces activity in auditory cortices whose activity patterns are similar with those generated during the perception of actual spoken words (Bourguignon, Baart, Kapnoula, & Molinaro, 2020; Calvert et al., 1997). This remarkable ability to infer coherent information of absent modalities from information provided by available perceptions, regardless of their nature or complexity, is defined as *cross-modality inference* (CMI). CMI also plays a key role in allowing humans to overcome changes to the perceptual conditions during the execution of tasks (Maurer, Pathman, & Mondloch, 2006; Spence, 2011; Walker et al., 2010). In Fig. 1 we highlight the importance of CMI to successfully perform tasks in scenarios with forced absence of perceptual experience.

Despite the complexity of its biological origin, the challenges of processing single-modality observational data and of learning a multimodal representation can also be addressed computationally (Yan et al., 2021, 2020, 2020). Multimodal generative models are a natural solution to learn a multimodal representation due to their ability to encode and generate multimodal data. However, as presented in Section 3, current multimodal generative models fall short of the potential of computational cross-modality inference. To address this issue, we contribute the Nexus model, a novel unsupervised hierarchical multimodal generative model and propose a novel encoder solution and training scheme that allows Nexus to learn a multimodal representation of an arbitrary number of input modalities. Inspired by the Convergence–Divergence Zone (CDZ) framework for neural recognition (Damasio, 1989), we build Nexus by considering hierarchical representation levels: at a bottom level, modality-specific representations specialize in

[*] Corresponding author.
E-mail addresses: miguel.vasco@tecnico.ulisboa.pt (M. Vasco), hyin@kth.se (H. Yin), fmelo@inesc-id.pt (F.S. Melo), ana.s.paiva@tecnico.ulisboa.pt (A. Paiva).

https://doi.org/10.1016/j.neunet.2021.11.019
0893-6080/© 2021 Elsevier Ltd. All rights reserved.

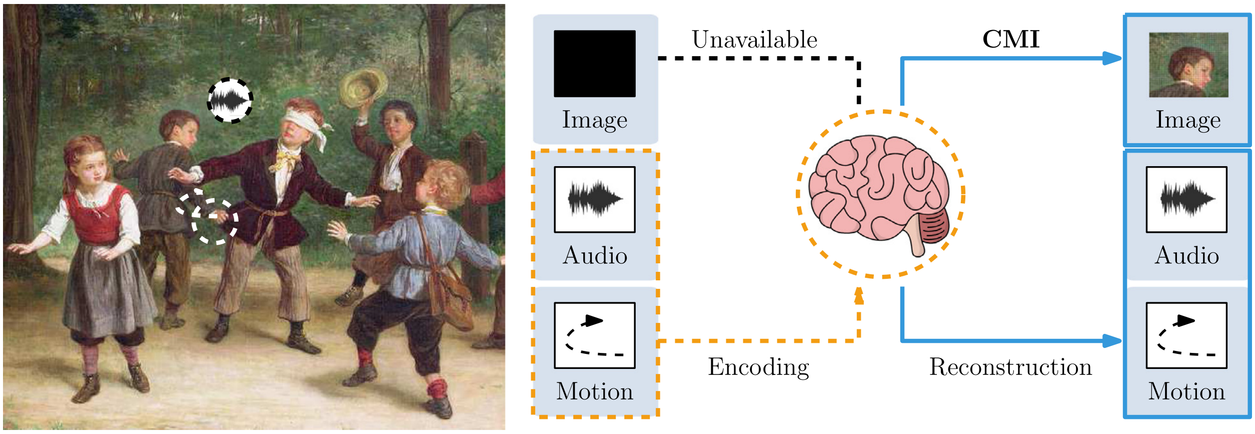
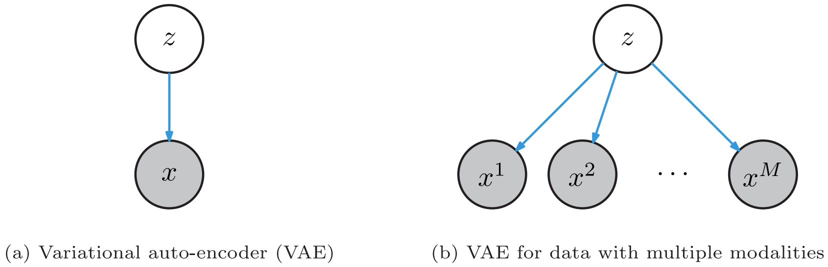


Fig. 1. The *cross-modality inference* (CMI) process: during the execution of tasks under incomplete perceptual experience, the agent encodes (dashed, orange) available perceptions (e.g. audio, motion) into a multimodal representation of the environment to generate (full, blue) absent modality information (e.g. image).
Source: Adapted from André-Henri Dargelas's painting "Blindman's Buff" (1828–1906).



(a) Variational auto-encoder (VAE)

(b) VAE for data with multiple modalities

Fig. 2. Bayesian network representation of the relation between the observed variable, *x*, and the latent variable, *z*.

generating data; at a top level, the multimodal representation generates samples of the modality-specific latent distributions. Nexus leverages these specialized hierarchical representations to generate high-quality, coherent cross-modality data.

We evaluate the cross-modality inference performance of Nexus. Such process should be evaluated in situations where multiple modalities provide information able to describe the same underlying reality. However, the majority of multimodal generative models are evaluated with standard two-modality datasets, often image and the associated label, falling short of accessing their CMI potential. To address the lack of a benchmark scenario for cross-modality generation with a large number of modalities, we contribute the "Multimodal Handwritten Digit" (MHD) dataset, combining image, motion, sound and semantic information from digit handwriting. We evaluate the key role of hierarchy in the generation of high-quality samples during CMI. In addition, we show that our novel training scheme allows Nexus to learn a multimodal representation robust to missing modalities at test time. The results reveal that Nexus outperforms state-of-the-art multimodal generative baseline models in regards to their CMI capabilities. To summarize, the contributions of this work are:

- In Section 3 we introduce three key issues of computational CMI that naturally arise from current multimodal generative models.
- In Section 5, inspired by the CDZ framework, we contribute Nexus, a novel unsupervised hierarchical generative model that learns a multimodal representation of an arbitrary number of modalities. Nexus successfully performs cross-modality inference, able to consider information provided by any set of input modalities and to generate high-quality coherent data for all target modalities. To do so, we introduce a new encoder solution for multimodal data (Section 5.1) and a new training scheme (Section 5.2).
- In Section 6.1, we contribute the "Multimodal Handwritten Digit" (MHD) dataset, a multimodal dataset that combines image, motion, sound and semantic information (i.e., class identity) from digit handwriting, thus providing a natural benchmark to evaluate CMI.
- In Section 6.2, we evaluate the key role of considering hierarchical representation spaces for high-quality, coherent, cross-modality generation.
- In Section 6.3 we evaluate Nexus in standard multimodal datasets and in Section 6.4 we evaluate Nexus in the challenging MHD scenario. The results reveal that our model outperforms all baseline methods in regards to their CMI capabilities.

2. Background

The variational auto-encoder (VAE) is a generative model originally introduced in the work of Kingma and Welling (2013). Given some data of interest, represented as a vector $x \in \mathbb{R}^w$, a VAE computes a representation of x (a "code") in the form of a vector $z \in \mathbb{R}^l$, such that x can be accurately reconstructed from z. Since, usually, $l \ll w$, the vector z can be seen as a compact representation of x, capturing the most relevant information to reconstruct x.

Formally, a VAE can be represented as the Bayesian network in Fig. 2a, where x is the observed data and z is a low-dimensional latent variable. Then, given a dataset $D = \{x_1, \ldots, x_N\}$, where each x_n is a sample of x, we want to compute a distribution p_θ that maximizes the (log-)likelihood of the data, given by

$$L(D) = \sum_{n=1}^{N} \log p_\theta(x_n)$$

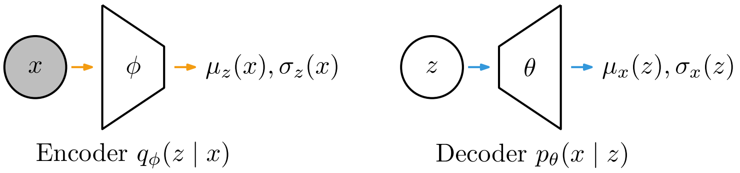


Fig. 3. Architecture of a variational auto-encoder. For a given input x, a neural network (the encoder) computes a mean $\mu_z(x)$ and variance $\sigma_z(x)$ describing the distribution $q_\phi(z|x)$. Conversely, given a "code" z, a neural network (the decoder) computes a mean $\mu_x(z)$ and variance $\sigma_x(z)$ describing the distribution $p_\theta(x|z)$. Often the variance $\sigma_x(z)$ is assumed to be constant.

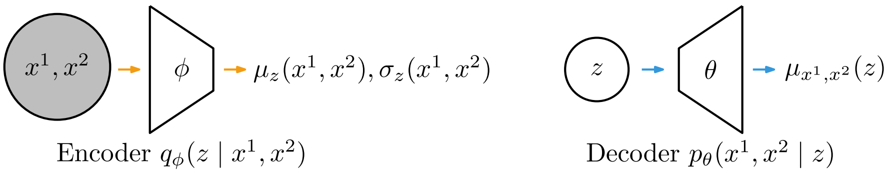

x^1, x^2
ϕ
$\mu_z(x^1, x^2), \sigma_z(x^1, x^2)$
Encoder $q_\phi(z \mid x^1, x^2)$
z
θ
$\mu_{x^1,x^2}(z)$
Decoder $p_\theta(x^1, x^2 \mid z)$

Fig. 4. Architecture of the naive multimodal extension of the variational auto-encoder. Note that, as this model ignores the decomposition of the input into distinct modalities, it is unable to perform *cross-modality inference*.

Training a VAE resorts to a variational approach. Letting

$$p_\theta(x) = \int p_\theta(x|z)p_0(z)dz, \tag{1}$$

where p_0 is some pre-specified prior (usually a unitary Gaussian distribution) and p_θ is a conditional distribution parameterized θ, it is possible to derive and *evidence lower bound* (or ELBO),

$$L(D) \geq \sum_{n=1}^{N} \mathbb{E}_{z\sim q_\phi(\cdot|x_n)}[\log p_\theta(x_n|z)] - \mathrm{KL}(q_\phi(z|x_n) \parallel p_0(z)), \tag{2}$$

where $q_\phi(z|x)$ is an arbitrary *posterior distribution*. The VAE is then trained by adjusting the parameters θ and ϕ such that the distributions p_θ and q_ϕ maximize the righthand side of (2) or, equivalently, minimize the loss

$$\ell(D) = \sum_{n=1}^{N} \mathrm{KL}(q_\phi(z|x_n) \parallel p_0(z)) - \mathbb{E}_{z\sim q_\phi(\cdot|x_n)}[\log p_\theta(x_n|z)]. \tag{3}$$

The diagram in Fig. 3 depicts the architecture of a VAE. The distributions q_ϕ and p_θ are usually taken as Gaussian distributions represented as neural networks. For a given input x, a first neural network – usually referred as the encoder – computes input-dependent mean $\mu_z(x)$ and variance $\sigma_z(x)$ that describes the distribution $q_\phi(z|x)$. Conversely, given a "code" z, a second neural network (the decoder) computes a mean $\mu_x(z)$ and variance $\sigma_x(z)$ describing the distribution $p_\theta(x|z)$. It is possible to interpret the first term on the righthand side of (2) as a *reconstruction term*, accounting for how well p_θ is able to reconstruct x from a code z generated by q_ϕ. The second term can be interpreted as a *regularization term*, striving to keep q_ϕ as simple as possible in order to allow for the generation of novel samples of x by sampling the code z from the prior $p_0(z)$.

The VAE model can easily be extended to deal with data with multiple modalities, as depicted in Fig. 2b. Suppose that x can be broken down as $x = (x^1, \ldots, x^M)$, where each x^m is an input "modality". Different modalities may correspond to different types of information (e.g., image, sound, etc.) and have different dimensionality.

A naive extension of the VAE model to this situation mostly ignores the individual modalities x^m, $m = 1, \ldots, M$, and treats x in an aggregated manner, as a single input. Fig. 2b depicts the Bayesian network for such approach, where

$$p_\theta(x|z) = \prod_{m=1}^{M} p_{\theta_m}(x^m|z), \tag{4}$$

and trivially leads to the loss

$$\ell(D) = \sum_{n=1}^{N} \mathrm{KL}(q_\phi(z|x_n) \parallel p_0(z)) - \mathbb{E}_{z\sim q_\phi(\cdot|x_n)}\left[\sum_{m=1}^{M} \log p_\theta(x_n^m|z)\right]. \tag{5}$$

where both the joint-modality encoder $q_\phi(z|x)$ and the modality-specific decoders $p_{\theta_k}(x^m|z)$ are once again instantiated as neural networks, as depicted in Fig. 4.

3. The problem of cross-modality inference

By ignoring the decomposition of the input into distinct modalities, the simple approach in Section 2 is unable to perform *cross-modality inference*, i.e., to reconstruct x from partial inputs.

One possible approach to enable the model to perform cross-modality inference – pioneered in the work of Yin et al. (2017) – is to consider an architecture akin to that depicted in Fig. 5. In this architecture, the Associative VAE (AVAE), a modality-specific encoder–decoder pair is trained to learn, respectively, the distributions $p_{\theta^m}(x^m|z)$ and $q_{\phi^m}(z|x^m)$, for $m = 1, \ldots, M$. These modality-specific models are combined by forcing the distributions over the latent space to *agree* for the same input.

For two modalities, i.e., an input $x = \{x^1, x^2\}$, Yin et al. introduce an *association loss* term of the form

$$\ell_{\text{assoc}}(D) = \sum_{n=1}^{N} d\left(q_{\phi^1}(\cdot|x_n^1), q_{\phi^2}(\cdot|x_n^2)\right) \tag{6}$$

where d is a distance metric between probability distributions.[1] The model in Fig. 5 is then trained as a single model, with a loss function

$$\ell_{\text{AVAE}}(D) = \ell_1(D) + \ell_2(D) + \lambda\ell_{\text{associative}}(D) \tag{7}$$

where

$$\ell_m(D) = \sum_{n=1}^{N} \mathrm{KL}(q_{\phi^m}(z|x_n^m) \parallel p_0(z)) - \mathbb{E}_{z\sim q_{\phi^m}(\cdot|x_n^m)}\left[\log p_{\theta^m}(x_n^m|z)\right]. \tag{8}$$

for $m = 1, 2$ and λ controls the relative importance of the association term. The trained model is now able, for example, to use the

[1] Yin et al. consider $d(p, q)$ to be the symmetric KL divergence between p and q, defined as $d(p, q) = \mathrm{KL}(p \parallel q) + \mathrm{KL}(q \parallel p)$.

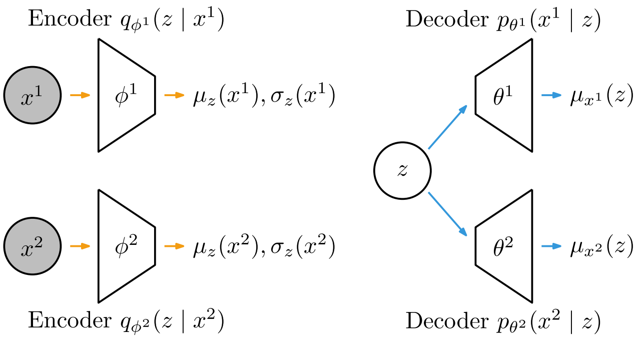

Encoder $q_{\phi^1}(z \mid x^1)$
x^1
ϕ^1
$\mu_z(x^1), \sigma_z(x^1)$
Decoder $p_{\theta^1}(x^1 \mid z)$
θ^1
$\mu_{x^1}(z)$
z
x^2
ϕ^2
$\mu_z(x^2), \sigma_z(x^2)$
θ^2
$\mu_{x^2}(z)$
Encoder $q_{\phi^2}(z \mid x^2)$
Decoder $p_{\theta^2}(x^2 \mid z)$

Fig. 5. Architecture of the Associational VAE (AVAE) from Yin, Melo, Billard, and Paiva (2017) for inputs with multiple modalities (see main text for details).

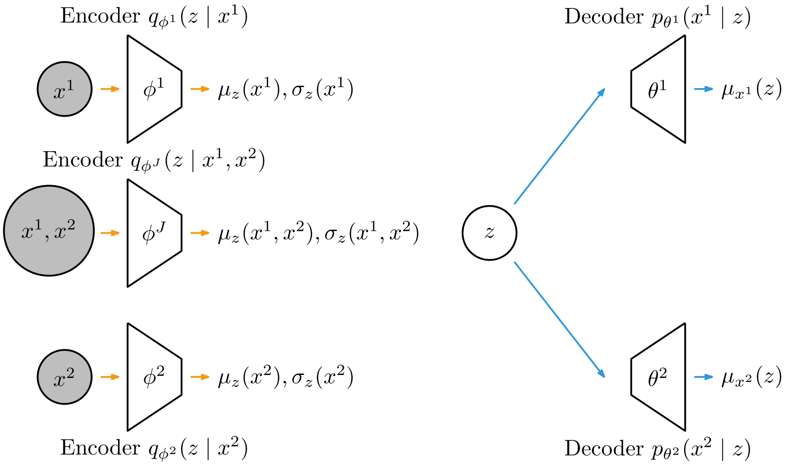

Encoder $q_{\phi^1}(z \mid x^1)$
x^1
ϕ^1
$\mu_z(x^1), \sigma_z(x^1)$
Decoder $p_{\theta^1}(x^1 \mid z)$
θ^1
$\mu_{x^1}(z)$
Encoder $q_{\phi^J}(z \mid x^1, x^2)$
x^1, x^2
ϕ^J
$\mu_z(x^1, x^2), \sigma_z(x^1, x^2)$
z
x^2
ϕ^2
$\mu_z(x^2), \sigma_z(x^2)$
θ^2
$\mu_{x^2}(z)$
Encoder $q_{\phi^2}(z \mid x^2)$
Decoder $p_{\theta^2}(x^2 \mid z)$

Fig. 6. Architecture for VAE supporting cross-modality inference from Suzuki, Nakayama, and Matsuo (2016). The model comprises a VAE for each of the two input modalities and a "joint" encoder for the combination of the two.

modality specific encoder q_{ϕ^1} to generate a latent vector z from the single modality input x^1, and then use this latent vector z as an input for the decoder p_{θ^2} to generate the missing modality x^2, successfully performing cross-modality inference. The approach of Yin et al. was limited to two modalities and unable to consider joint-modality information (having access to both x^1 and x^2) for the generation of modality data. To address these issues, Suzuki et al. (2016) proposed an extension of the model, depicted in Fig. 6. The architecture can be seen as a combination of multiple VAEs, one for every individual modality, and one "joint" encoder for every possible combination of modalities. This model, while not limited to two modalities, presents an obvious disadvantage: its dimension grows rapidly with the number of input modalities.

The models of Yin et al. and Suzuki et al. highlight a key difficulty in designing these models:

(i) **Scalability**: The model must "merge" the information from the different input modalities, in order to perform cross-modality inference. As more modalities are considered in the model, such merging should be efficient, scaling gracefully with the number (and dimensionality) of the input modalities. The use of multiple "sub-models", as in the work of Suzuki et al. (2016) does not scale well with the number of modalities and is, therefore, unsuited to deal with situations with a large number of modalities.

The difficulty identified above is common to other approaches that consider multiple modalities (Korthals, Rudolph, Leitner, Hesse, & Rückert, 2019; Ma, McDuff, & Song, 2019; Tian & Engel, 2019; Tsai, Liang, Zadeh, Morency, & Salakhutdinov, 2018; Vedantam, Fischer, Huang, & Murphy, 2017).

To address the **scalability** issue (i) in the design of multimodal generative models, Wu and Goodman (2018) proposed to employ an *implicit* joint-modality encoder, composed of some function f of single-modality distributions. In this work, the authors proposed a joint-modality encoder composed of a product-of-experts (POE) factorization of single-modality encoders with a prior-expert $q_\phi(z|x) \propto p(z)\prod_{m=1}^{M} q_{\phi^m}(z|x^m)$, as shown in Fig. 7. Denoted by Multimodal VAE (MVAE), the approach is able to scale to arbitrarily large number of modalities without requiring the creation of specific "sub-models" to account for combinations of modalities. In order to be able to perform cross-modality inference, the model of Wu et al. requires a *sub-sampling* training scheme that considers ELBO terms for complete (joint) observations, for single-modality observations and for partial observations with randomly chosen subsets of modalities. For scenarios with two modalities, i.e., an input $x = \{x^1, x^2\}$, this corresponds to a loss function:

$$\ell_{\text{MVAE}}(D) = \ell_{1,2}(D) + \ell_1(D) + \ell_2(D) \tag{9}$$

where the whole-observation term is defined as:

$$\ell_{1,2}(D) = \sum_{n=1}^{N} \mathrm{KL}(q_\phi(z|x_n) \parallel p_0(z)) - \mathbb{E}_{z\sim q_\phi(\cdot|x_n)}\Big[\log p_{\theta^1}(x_n^1|z) + \log p_{\theta^2}(x_n^2|z)\Big], \tag{10}$$

and the (partial) individual ELBO terms are now defined as:

$$\ell_m(D) = \sum_{n=1}^{N} \mathrm{KL}(q_\phi(z|x_n^m) \parallel p_0(z)) - \mathbb{E}_{z\sim q_{\phi(\cdot|x_n^m)}}\Big[\log p_{\theta^1}(x_n^m|z)\Big]. \tag{11}$$

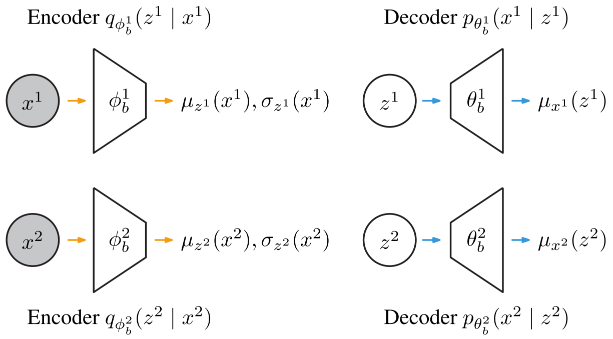

Encoder $q_{\phi_b^1}(z^1 \mid x^1)$
x^1
ϕ_b^1
$\mu_{z^1}(x^1), \sigma_{z^1}(x^1)$
Decoder $p_{\theta_b^1}(x^1 \mid z^1)$
z^1
θ_b^1
$\mu_{x^1}(z^1)$
x^2
ϕ_b^2
$\mu_{z^2}(x^2), \sigma_{z^2}(x^2)$
z^2
θ_b^2
$\mu_{x^2}(z^2)$
Encoder $q_{\phi_b^2}(z^2 \mid x^2)$
Decoder $p_{\theta_b^2}(x^2 \mid z^2)$

Fig. 7. Architecture for VAE supporting cross-modality inference with an *implicit* joint-modality encoder $q_\phi(z|x^1, x^2)$, composed of some function f of single-modality encoder distributions $q_{\phi^1}(z|x^1)$, $q_{\phi^2}(z|x^2)$. The MVAE model proposed by Wu and Goodman (2018) considers a product-of-experts (POE) factorization with a prior-expert $q_\phi(z|x) \propto p(z)q_{\phi^1}(z|x^1)q_{\phi^2}(z|x^2)$.

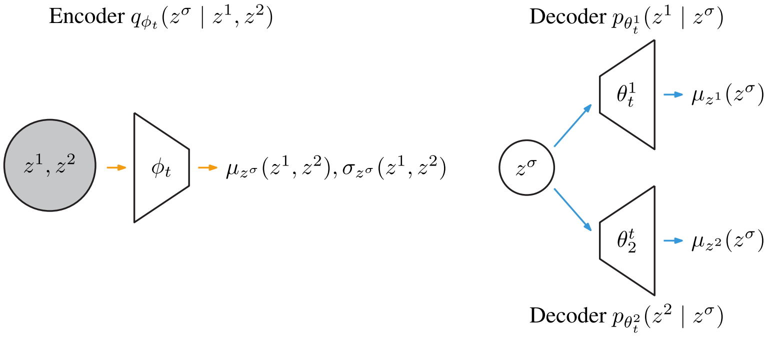

Encoder $q_{\phi_t}(z^\sigma \mid z^1, z^2)$
z^1, z^2
ϕ_t
$\mu_{z^\sigma}(z^1, z^2), \sigma_{z^\sigma}(z^1, z^2)$
Decoder $p_{\theta_t^1}(z^1 \mid z^\sigma)$
θ_t^1
$\mu_{z^1}(z^\sigma)$
z^σ
θ_2^t
$\mu_{z^2}(z^\sigma)$
Decoder $p_{\theta_t^2}(z^2 \mid z^\sigma)$

Fig. 8. Architecture for VAE supporting cross-modality inference by Shi et al. (2019). Employing an *implicit* joint-modality encoder, the MMVAE defines a mixture-of-experts (MOE) over the single-modality distributions $q_{\phi^1}(z|x^1)$, $q_{\phi^2}(z|x^2)$.

This training scheme presents an obvious disadvantage: the number of ELBO terms in the loss function grows rapidly with the number of input modalities. In addition, the model presents another less-obvious disadvantage: developed for weakly-supervised learning scenarios, where joint-modality information may not be fully available during training, the POE solution is prone to overconfident expert prediction, often of the higher-dimensional modality (e.g. images) (Shi, Siddharth, Paige, & Torr, 2019). As such, the model is able to infer missing low-dimensional information (e.g. label) from high-dimensional modalities (e.g. image), yet struggle with the inverse inference process.

Recently, Shi et al. (2019) proposed a novel model, denoted by Mixture-of-Experts MVAE (MMVAE), that employs an implicit mixture-of-experts (MOE) joint-modality encoder, $q_\Phi(z|x) = \sum_{m=1}^{M} \alpha_m \, q_{\phi^m}(z|x^m)$, with $\alpha_m = 1/M$ under the assumption that the modalities are of similar complexity. As shown in Fig. 8, the MOE solution incurs on some computational overhead due to the necessity of computing M^2 passes over the single-modality decoders, as each modality provides samples from its own encoding distribution to be evaluated by all generative models. For two modalities, i.e., $x = \{x^1, x^2\}$, this results in a loss function:

$$\ell_{\text{MMVAE}}(D) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{2}\mathbb{E}_{z\sim q_\phi(\cdot|x_n^m)} \log\left[\frac{p_\Theta(x_n^1, x_n^2, z)}{q_\Phi(z|x_n^1, x_n^2)}\right] \qquad (12)$$

However, the MOE solution proposed by Shi et al. (2019) presents a significant disadvantage for cross-modality inference: to infer missing modality data, the model is *only* able to consider the information provided by a single available modality. For example, given two available modalities x^1, x^2, the model randomly chooses a single modality to generate a latent representation and, subsequently, infer information regarding a third non-available modality x^3. This issue is significantly aggravating in scenarios where, instead of redundant information, the different modalities provide complementary information to characterize the underlying phenomena.

The models of Shi et al. (2019) and Wu and Goodman (2018) highlight two more key difficulties in designing multimodal generative models with implicit joint-modality encoders:

(ii) **Generalization**: The model must be able to infer missing modality information from provided available information, regardless of the nature and complexity of both target and input modalities.

(iii) **Compositionality** To perform cross-modality inference the model must be able to account for the information provided by *all* available modalities. As more modalities are made available, the model should be able to consider the redundant and complementary information they provide in order to generate a more adequate multimodal latent representation.

Other approaches proposed the factorization of the multimodal representation into separate, independent, representations (Hsu & Glass, 2018; Lee & Pavlovic, 2020; Tsai et al., 2018). Tsai et al. (2018) proposed a factorized model that encodes a multimodal representation separated into multimodal discriminative factors and modality-specific generative factors. Hsu and Glass (2018) proposed to disentangle (modality-specific) style and semantic generative factors in a factorized way. However, by factorizing the multimodal representations into independent generative factors, both models are unable to perform cross-modality inference when some modalities are unavailable, for example when label information is unavailable. By requiring explicit semantic (label) information to encode the discriminative factors, the models are unable to address the **generalization issue**.

Other recent approach by Sutter, Daunhawer, and Vogt (2020) proposed the use of a MoE joint encoder with a novel Jensen–Shannon-Divergence loss, in order to overcome the significant computational cost of training the original MMVAE model, and the use of factorized modality-specific representation spaces to improve the generative capabilities of the MMVAE model. However, similarly to the MMVAE model, this approach is unable to address the **compositionality** issue.

In parallel, some solutions consider the problem of learning a *disentangled* multimodal representation, either in a single-latent representation (Daunhawer, Sutter, Marcinkevičs, & Vogt, 2021) or in multiple factorized representations (Lee & Pavlovic, 2020): Daunhawer et al. (2021) propose a *single-latent representation* model that learns to disentangle modality-specific and invariant factors in a self-supervised way. Lee and Pavlovic (2020), quite similarly to the work of Tsai et al. (2018), explore the disentanglement problem considering the factorization of modality-specific and multimodal factors, using a Product-of-Experts solution to merge multimodal information. In this work we explore learning multimodal representations without requiring disentanglement, as the full assessment of the benefits of enforcing disentanglement in representation learning remains an open question (Locatello et al., 2019).

We address the issues brought up by computational approaches to the cross-modality inference process. In a recent work, four criteria for the successful learning of a multimodal representation were posited (Shi et al., 2019). The issues presented in this work can also be considered as desiderata for the learning process of multimodal generative models and, as such, we can naturally establish associations with those criteria: for example, the **generalization** issue shares the same concerns as the "Coherent Cross Generation" criteria. In this sense, the issues presented here can also be employed as evaluation criteria of the quality of multimodal generative models. We instantiate such evaluation in Section 6.

In this work, we develop a model that graciously scales with an arbitrary number of modalities, addressing the **scalability** (i) issue, and is able to successfully perform *robust* cross-modality inference considering all information provided to the model and regardless of the target and available modalities nature and complexity, addressing both the **generalization** (ii) and the **compositionality** (iii) issues.

4. Human representation learning

We now turn the attention to the case of human representation learning. Humans are provided with a complex cognitive framework that allows for multimodal perception. Several regions of the brain are responsible for the convergence of multimodal information, even in areas once thought to process only unimodal information (Ghazanfar & Schroeder, 2006). These regions contain multimodal neurons that respond to stimuli from multiple sensory modalities, whose behaviour begins to be uncovered with the development of novel brain imaging techniques (Burianová et al., 2013; Calvert, 2001; Man, Kaplan, Damasio, & Damasio, 2013; Marstaller & Burianová, 2014).

The *Convergence–Divergence Zone* (CDZ) framework is widely employed to explain the neural mechanisms of human multimodal perception (Damasio, 1989; Meyer & Damasio, 2009). In the CDZ model, depicted in Fig. 9, two different sets of neuron ensembles are proposed: (i) lower-level ensembles in early sensory and motor cortices, responsible for the processing of modality-specific information; and (ii) higher-level ensembles in association cortices, responsible for the processing of multimodal information. According to the framework, the high-level neuron ensembles (multimodal) do not hold a composite version of the original perceptual information, but instead hold a record of the arrangement of the low-level neural ensemble activity generated by the perception of a given object (Damasio, 1989). The existence of higher-level multisensory convergence zones can be observed experimentally. The superior colliculus of the human midbrain contains multimodal neurons that respond to visual and auditory stimuli, in part responsible for the orienting behaviour of moving one's gaze towards the source of a sound (Bell, Meredith, Van Opstal, & Munoz, 2005; Edwards, Ginsburgh, Henkel, & Stein, 1979).

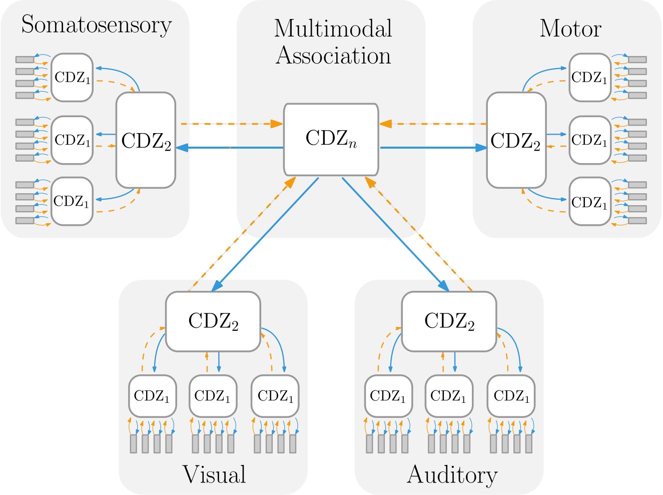


Fig. 9. The CDZ framework, proposed by Damasio (1989). The model distinguishes between modality-specific neuron ensembles in early sensorimotor cortices and higher-order neural ensembles in multimodal association cortices. In the CDZ framework, information is propagated from the modality-specific cortices (orange, dashed arrows) to first-order CDZs, which, in turn, project back information (blue, full arrows) to the early cortical sites. Modality-specific information from low-order CDZs is propagated forward in order to encode a multimodal representation of the observed phenomena in higher-level association cortices (CDZ$_n$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

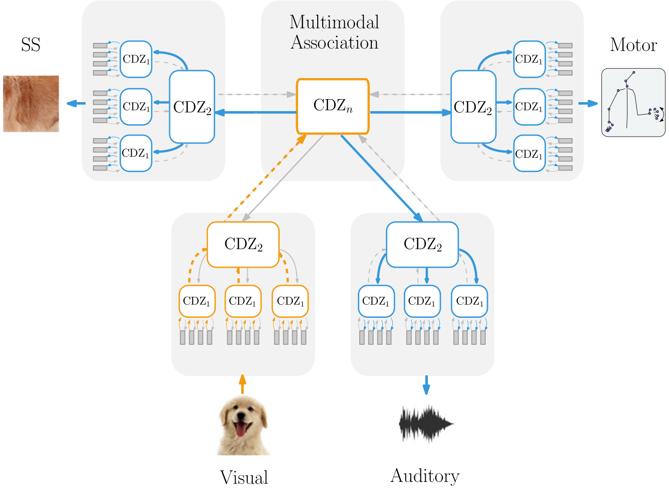


Fig. 10. The cross-modality inference process in the CDZ framework, proposed by Damasio (1989): in this example, available information (image of a dog) is collected by the visual sensors of the human and forward processes in order to encode a multimodal representation, from which information is back-propagated to the remaining (absent) perceptual modalities.

The CDZ framework also provides a graceful explanation of the human cross-modality inference process. As depicted in Fig. 10, the available perceptual information results in the activation

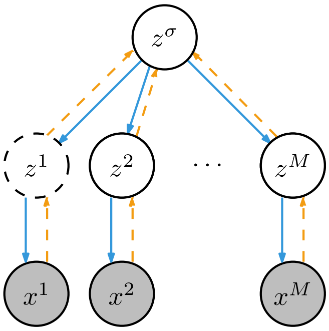


Fig. 11. Network representation of the Nexus model, highlighting the two-level architecture and the relation between the observed variables $x^1, \ldots, x^M$, the *modality-specific* latent variables $z^1, \ldots, z^M$ and the multimodal *context* latent variable z^σ.

of modality-specific low-level neural ensembles, whose activity patterns are forward projected to the high-level multimodal ensembles. Subsequently, the high-level ensembles propagate information back to the modality-specific neural ensembles, inducing activity that is coherent to the (absent) perceptual phenomena. Such cross-modality activations have been observed experimentally as well. For example, the visual observation of lip movement (e.g. when observing a muted video clip) results in the retro-activation of early auditory cortices, whose activity pattern resembles that of the expected sound, despite sound not being provided in the current sensory environment (Bourguignon et al., 2020; Calvert et al., 1997). Cross-modality activations have been observed for other sensory modalities, such as the activation of auditory and olfactory cortices by reading words with auditory or olfactory meaning, respectively (González et al., 2006; Kiefer, Sim, Herrnberger, Grothe, & Hoenig, 2008).

The parallel between biological and computational representation learning further motivates our work in multimodal representation learning. In particular, in this work we take inspiration from the CDZ model and leverage hierarchy to design a novel hierarchical generative model able to perform efficient cross-modality generation.

5. The Nexus model

The Nexus model can be described by the network architecture depicted in Fig. 11. We consider a scenario with a set of M modalities of arbitrary nature, $x = \{x^1, x^2, \ldots, x^M\}$. Inspired by the CDZ model (Damasio, 1989; Meyer & Damasio, 2009), we build Nexus considering an hierarchical structure, composed of two representation levels: at a bottom level we assume that modality data is generated by a stochastic process mediated by *modality-specific* latent variables $z = \{z^1, z^2, \ldots, z^M\}$. The training of each modality-specific representation, depicted in Fig. 12, follows the VAE loss function of Eq. (3), resulting in a total bottom-level loss $\ell_b(D)$,

$$\ell_b(D) = \sum_{n=1}^{N}\sum_{m=1}^{M} \mathrm{KL}\left[q_{\phi_b^m}(z_n^m \mid x_n^m) \;\|\; p(z_n^m)\right] - \mathbb{E}_{q_{\phi_b^m}(z_n^m|x_n^m)}\left[\log p_{\theta_b^m}(x_n^m \mid z_n^m)\right], \tag{13}$$

where $q_{\phi_b^m}(z^m \mid x^m)$ and $p_{\theta_b^m}(x^m \mid z^m)$, correspond to the bottom-level modality-specific encoder and decoder networks. Thus, each modality-specific latent variable encodes a representation specialized in the generation of the corresponding modality-specific data. By considering separate modality-specific latent representations, each latent space can have an adequate dimensionality to the complexity of the underlying modality.

As shown in Fig. 13, at the top-level Nexus learns a multimodal representation z^σ responsible for the generation of samples of the *modality-specific* latent distributions $z^1, z^2, \ldots, z^M$. Following the multimodal VAE loss of Eq. (5), the training of the multimodal representation follows the top-level loss $\ell_t(D)$,

$$\ell_t(D) = \sum_{n=1}^{N} \mathrm{KL}\left[q_{\phi_t}(z_n^\sigma \mid \bar{z}_n^{1:M}) \;\|\; p_0(z_n^\sigma)\right] - \sum_{m=1}^{M} \mathbb{E}_{\substack{q_{\phi_b}(\bar{z}_n^{1:M}|x_n^{1:M}) \\ q_{\phi_t}(\; z_n^\sigma \;|\bar{z}_n^{1:M})}}\left[\log p_{\pi_m}(\bar{z}_n^m \mid z_n^\sigma)\right], \tag{14}$$

where $q_{\phi_t}(z^\sigma \mid \bar{z}^{1:M})$ and $p_{\theta_t^m}(\bar{z}^m \mid z^\sigma)$, correspond to the top-level joint-modality encoder and modality-specific decoders, respectively. The bar symbol over the modality-specific latent samples $\bar{z}^{1:M}$ denotes that no gradients are propagated through this value back to the lower-level networks, which we found to improve the model's performance.[2]

The total loss objective of the Nexus model $\ell(D)$ is given by,

$$\begin{aligned}\ell(D) &= \ell_b(D) + \ell_t(D) \\ &= \sum_{n=1}^{N}\Bigg(\mathrm{KL}\left[q_{\phi_t}(z_n^\sigma \mid z_n^{1:M}) \;\|\; p_0(z_n^\sigma)\right] \\ &+ \sum_{m=1}^{M}\Bigg(\mathrm{KL}\left[q_{\phi_b^m}(z_n^m|x_n^m) \;\|\; p_0(z_n^m)\right] \\ &- \mathbb{E}_{q_{\phi_b^m}(z_n^m|x_n^m)}\left[\log p_{\theta_b^m}(x_n^m \mid z_n^m)\right] \\ &- \mathbb{E}_{\substack{q_{\phi_b^m}(z_n^{1:M}|x_n^{1:M}) \\ q_{\phi_t}(z_n^\sigma|z_n^{1:M})}}\left[\log p_{\theta_t^m}(z_n^m|z_n^\sigma)\right]\Bigg)\Bigg)\end{aligned} \tag{15}$$

It is important to note that learning the multimodal representation z^σ should be a simpler task than learning a multimodal representation in non-hierarchical models: in Nexus, z^σ does not concern itself with the generation of high-dimensional and complex modality data but only with the generation of low-dimensional modality-specific latent samples. The modality-decoders will then decode such samples, specialized in the generation of high-quality modality-specific data.

Contrary to other works that consider distinct modality-specific and discriminative representation spaces to generate multimodal data (requiring label information in the process), the hierarchical design of Nexus allows for cross-modality inference regardless of the nature of the provided modalities: modality-specific data is generated only from its corresponding modality-specific latent variable, which in turn is generated from the multimodal context latent variable. By leveraging hierarchy, we address directly the **generalization** issue presented in Section 3. We show the fundamental role of hierarchy to generate coherent, high-quality, data in Section 6.2.

5.1. Joint-modality encoder

We now turn to the question of how to define the multimodal joint proposal distribution $q_{\phi_t}(z^\sigma|z^{1:M})$ in order to encode a representation able to tackle the remaining issues discussed in Section 3: how to learn such representation in a way that is (graciously) extendable to an arbitrary number of modalities (**scalability** issue) and that is able to consider the information provided by all available modalities (**compositionality** issue).

[2] For simplicity of notation, we will not employ the bar notation throughout the remaining of the paper.

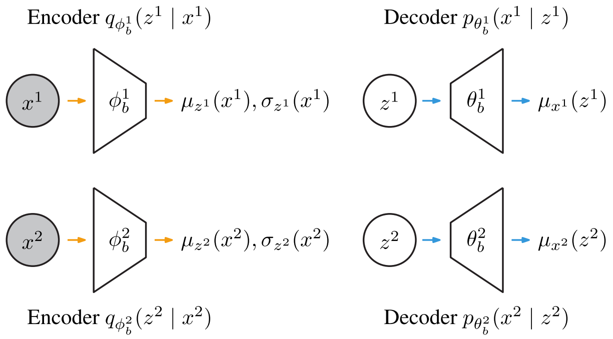

Encoder $q_{\phi_b^1}(z^1 \mid x^1)$
Decoder $p_{\theta_b^1}(x^1 \mid z^1)$
x^1
ϕ_b^1
$\mu_{z^1}(x^1), \sigma_{z^1}(x^1)$
z^1
θ_b^1
$\mu_{x^1}(z^1)$
x^2
ϕ_b^2
$\mu_{z^2}(x^2), \sigma_{z^2}(x^2)$
z^2
θ_b^2
$\mu_{x^2}(z^2)$
Encoder $q_{\phi_b^2}(z^2 \mid x^2)$
Decoder $p_{\theta_b^2}(x^2 \mid z^2)$

Fig. 12. Architecture of the modality-specific networks of the Nexus model, instantiated for a scenario with two modalities, highlighting the relation between the observed variables x^1, x^2, the *modality-specific* latent variables z^1, z^2.

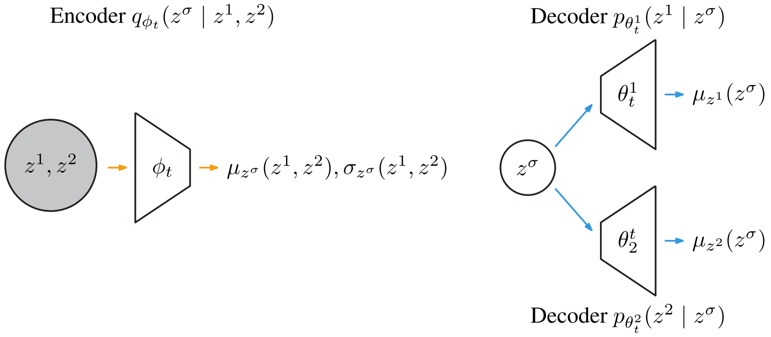

Encoder $q_{\phi_t}(z^\sigma \mid z^1, z^2)$
Decoder $p_{\theta_t^1}(z^1 \mid z^\sigma)$
z^1, z^2
ϕ_t
$\mu_{z^\sigma}(z^1, z^2), \sigma_{z^\sigma}(z^1, z^2)$
z^σ
θ_t^1
$\mu_{z^1}(z^\sigma)$
θ_2^t
$\mu_{z^2}(z^\sigma)$
Decoder $p_{\theta_t^2}(z^2 \mid z^\sigma)$

Fig. 13. Architecture of the multimodal networks of the Nexus model for a scenario with two modalities, highlighting the relation between the *modality-specific* latent variables z^1, z^2 and the multimodal *contextual* latent variable z^σ.

We introduce the *aggregator* joint-modality encoder, a novel solution to approximate the joint-modality proposal distribution, as depicted in Fig. 14. Following recent work in Graph Neural Networks (Hamilton, Ying, & Leskovec, 2017), we approach the encoding process of multimodal data as a Directed Acyclical Graph (DAG), in which the nodes of the graph correspond to the modality-specific latent representations $z^{1:M}$ and the multimodal latent representation z^σ. Each modality-specific representation has a single directed edge towards the multimodal note z^σ. We can define the flow of information in the graph as:

$$z^\sigma \leftarrow f(z^1, z^2, \ldots, z^M), \tag{16}$$

where we define an *aggregator* function $f^{(M)} : \{k^1, \ldots, k^M\} \rightarrow k^\sigma \in \mathbb{R}^k$, responsible for aggregating the information provided by each modality-specific representation. As the *aggregator* function requires that the samples provided are of common dimensionality, we preprocess those samples using modality-specific *mapping* networks $q_{\phi_t^m}(z^m)$ to reduce all samples to a common dimensionality d_k. We can define the multimodal encoder $q_{\phi_t}(z^\sigma \mid z^{1:M})$:

$$q_{\phi_t}(z^\sigma \mid z^{1:M}) := q_{\phi_t}\left(z^\sigma \mid f\left(k^1, k^2, \ldots, k^M\right)\right). \tag{17}$$

Several choices of an *aggregator* function can be employed, from simple concatenation to more complex recurrent networks, such as RNNs and LSTM. Empirically, we found a simple *mean* function to be suitable for the aggregation procedure. By employing an *aggregator* function, we allow the multimodal encoding procedure to consider an arbitrary number of modalities, thus addressing the **scalability** issue presented in Section 3. Moreover, the multimodal representation is able to be encoded considering information provided by all available modalities (or any subset of available modalities), thus addressing the **compositionality** issue.

Aggregator functions have been extensively employed in machine learning literature. Recently, of particular interest, these methods became a core component of Neural Processes (Garnelo et al., 2018) and of Generative Query Networks (Eslami et al., 2018). Even in such complex networks, simple aggregator functions have shown remarkable performance in merging information: a mean function for Neural Processes (similarly to Nexus) is used to summarize image encoded inputs and a additive function for Generative Query Networks is used to summarize image observations. However, to the best of our knowledge, Nexus is the first model to employ an aggregator function as a multimodal encoder to merge information provided by different modalities.

5.2. Forced Perceptual Dropout (FPD) training scheme

While the *aggregator* function is able to consider any subset of available modalities to encode the multimodal representation z^σ, by always providing all modalities during its training, the model might lack robustness to missing modalities at test-time and, as such, not be able to perform CMI. To address this issue, we propose a novel training scheme for the multimodal encoder which we denote by *Forced Perceptual Dropout* (FPD), whose pseudo-code is presented in Algorithm 1.

During training, given a complete latent set $z_n^{1:M} = \{z_n^1, \ldots, z_n^M\}$, we define a smaller “available” subset $z_n^d \in z_n^{1:M}$ of modality-specific latent representations by sampling latent samples from an Uniform distribution (without replacement) over the complete set $z_n^{1:M}$. The number of representations to consider is also sampled from an Uniform distribution $U\{1, M-1\}$. For each sample in the batch, the model determines if the dropout mechanism is to be applied by sampling from a Bernoulli distribution,

$$\mathbb{1}_n^d \sim \text{Bern}(\rho), \tag{18}$$

where the user-defined parameter ρ defines the probability of dropout occurring. Finally, we define the multimodal encoder

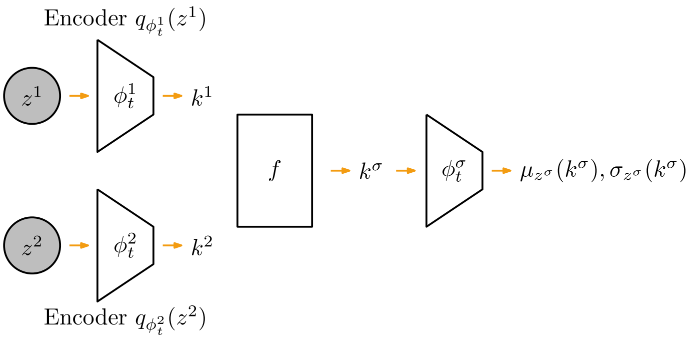

Encoder $q_{\phi_t^1}(z^1)$
z^1
ϕ_t^1
k^1
f
k^σ
ϕ_t^σ
$\mu_{z^\sigma}(k^\sigma), \sigma_{z^\sigma}(k^\sigma)$
z^2
ϕ_t^2
k^2
Encoder $q_{\phi_t^2}(z^2)$

Fig. 14. The *aggregator* joint-modality encoder $q_{\phi_t}(z^\sigma \mid z^{1:M})$ employed by Nexus, instantiated in a scenario with two modalities x^1, x^2: the modality-specific latent samples z^1, z^2 are initially mapped to a common-dimensionality latent *code* representation $k^1, k^2 \in \mathbb{R}^k$. The information provided by the codes is merged using an *aggregator* function f, resulting in an aggregated code k^σ which is encoded to generate the top-representation latent space z^σ.

Algorithm 1 Forced Perceptual Dropout (FPD)

```
Input: Dropout parameter $\rho$; batch-size $N$; batch of modality-specific latent codes $z^{1:M} = \{z^1, \dots, z^M\}$
for each sample in batch $n = 1, 2, \dots, N$ do
    Define $z_n = z_n^{1:M}$
    Sample Dropout indicator $\mathbb{1}_n^d \sim \text{Bern}(\rho)$
    if $\mathbb{1}_n^d = 1$ then
        Sample available subset $z_n^d \in z_n$
        Encode $z_n^\sigma \sim q_{\phi_t}(\cdot \mid z_n^d)$
    else
        Encode $z_n^\sigma \sim q_{\phi_t}(\cdot \mid z_n)$
    end if
end for
```

$q_{\phi_t}(z^\sigma \mid z_n^{1:M})$ accordingly to the following rule:

$$q_{\phi_t}(z^\sigma \mid z_n^{1:M}) = \begin{cases} q_{\phi_t}(z^\sigma \mid z_n^d), & \text{if } \mathbb{1}_n^d = 1 \\ q_{\phi_t}(z^\sigma \mid z_n^{1:M}), & \text{otherwise} \end{cases} \tag{19}$$

We repeat the FPD procedure for each sample in a training batch. Other multimodal models employ similar mechanisms to improve robustness to missing modalities, such as the subsampling training scheme of MVAE (Wu & Goodman, 2018). However, that scheme requires multiple forward passes of the whole batch of data through the model (and multiple gradient computations), proportional to the number of possible combinations of modalities, and, as such, is computationally intensive in scenarios with large number of modalities. On the other hand, FPD is applied in a single forward pass, reducing the computational training overhead.

By employing FPD to train Nexus, we are forcing the model to learn to encode a multimodal representation, able to generate all modality-specific representations, despite not being given complete multimodal information. In this way, we explicitly account for the CMI process during training and promote the robustness of the model to missing-modality information at test-time. We evaluate the role of FPD and the dropout parameter ρ for the robustness of the multimodal representation encoded in Appendix A.

6. Evaluation

We evaluate quantitatively and qualitatively the performance of Nexus in performing CMI, accordingly to three criteria brought up by the issues presented in Section 3: able to consider an arbitrary number of modalities (**scalability**), able to improve the quality of the multimodal representation considering the (redundant and complementary) information that multiple modalities provide (**compositionality**) and able to generate coherent, high-quality, data regardless of the target modality (**generalization**).

Likelihood-based metrics are often employed to evaluate the generative capabilities of multimodal models (Suzuki et al., 2016; Wu & Goodman, 2018). However, such metrics are not suitable to evaluate cross-modality generation due to the lack of a defined target sample, required to compute the probabilistic likelihood. Another recent approach employs accuracy-based metrics (Shi et al., 2019). However, accuracy metrics are not able to evaluate the quality of the generated samples, only their semantic coherence given the provided data.

To address the lack of quantitative metrics suitable for evaluating the computational CMI process, we propose two complementary evaluation metrics, represented in Fig. 15:

- **Accuracy** — evaluates if the samples generated by CMI are semantically coherent with the available modality data that was provided to the model, *e.g.* generated images from label 7 should be classified as image samples of that digit. Higher is better.
- **Modality Frechet Distance (MFD)** — evaluates if the samples generated by CMI are similar to the original samples in the dataset, *e.g.* generated images from label 7 should account for the different ways an image of the digit 7 can be handwritten. This metric is not considered for symbolic modalities (e.g. discrete labels). Lower is better.

To evaluate accuracy, in each dataset, we employ pretrained modality-specific classifiers and evaluate the accuracy of generated samples (higher is better). To evaluate MFD, in each dataset, we employ pretrained class-based and modality-specific autoencoders, responsible for encoding a representation of the provided samples. For each class $k \in [0, K]$ in the dataset, we encode representations of samples both from the dataset and generated by cross-modality inference, resulting in a distribution of real dataset representations $\mathcal{N}(\mu_r^k, \Sigma_r^k)$ and of generated representations $\mathcal{N}(\mu_g^k, \Sigma_g^k)$. The encoding procedure of both distributions is performed by pretrained class-and-modality-specific auto-encoders. The MFD score F is then given by the Frechét distance between the two distributions, averaged over all classes (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017):

$$F = \frac{1}{K} \sum_k \left\| \mu_r^k - \mu_g^k \right\|^2 + \text{Tr}\left(\Sigma_r^k + \Sigma_g^k - 2 \left(\Sigma_r^k \Sigma_g^k \right)^{1/2} \right). \tag{20}$$

The same classifiers and class-and-modality-specific autoencoders are employed in the evaluation of all considered models. In Appendix B.2 we present the architecture and training hyperparameters of the classifiers and autoencoders.

In order to understand the role of the hierarchical extension and the proposed aggregation joint-modality encoder for efficient

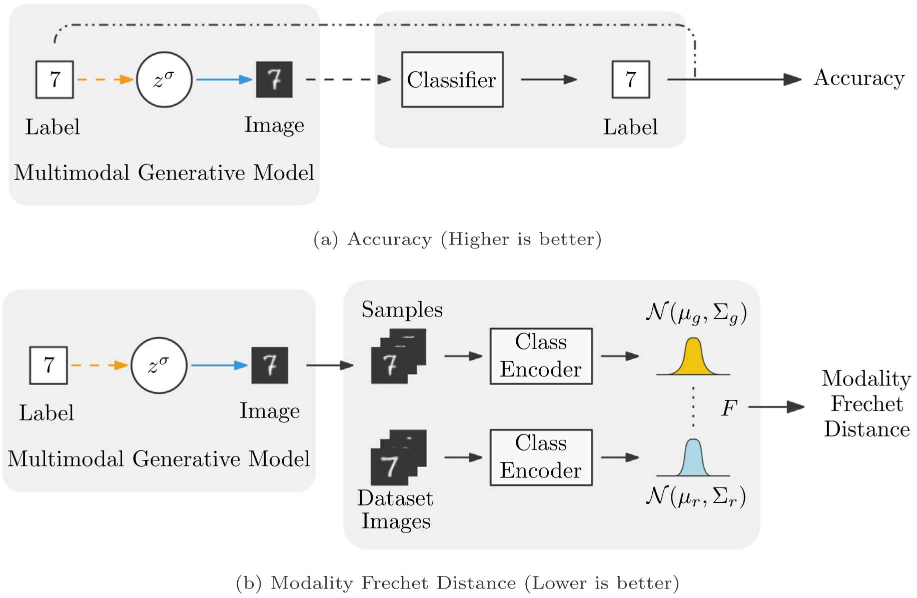

7
z^σ
Label
Image
Multimodal Generative Model
Classifier
7
Label
Accuracy
(a) Accuracy (Higher is better)
7
z^σ
Label
Image
Multimodal Generative Model
Samples
Class Encoder
$\mathcal{N}(\mu_g, \Sigma_g)$
F
Modality Frechet Distance
Dataset Images
Class Encoder
$\mathcal{N}(\mu_r, \Sigma_r)$
(b) Modality Frechet Distance (Lower is better)

Fig. 15. Pictorial description of the CMI evaluation metrics employed for Multimodal Generative models in this work. We evaluate both the semantic coherence of the generated samples (*accuracy*), as well as their quality (*MFD*) A robust CMI performance should provide samples in the high accuracy and low MFD regime regardless of the complexity or nature of the target modality.

cross-modality inference, we consider two different variations of the Nexus model:

- Nexus — The full proposed model employing the joint-modality aggregation encoder introduced in Section 5.1 and the FPD training scheme;
- Nexus-0 — The Nexus model employing a naive concatenation of the modality-specific codes and the FPD training scheme, allowing the evaluation of the role of considering hierarchical representation spaces;

We evaluate Nexus against state-of-the-art variational-autoencoder-based multimodal generative models that are able to address the fundamental conditions for computational cross-modality inference: **(i)** able to consider an arbitrary number of modalities; **(ii)** able to consider modalities of arbitrary nature, e.g., without requiring semantic information (labels). As such, we select the two baselines able to account for such restrictions: the MVAE model and the MMVAE model, employing the authors' publicly available code[3]. For evaluation purposes we employ the same training hyperparameters and network architecture across all models, presented in the Appendix. For fairness, we train all the models according to the standard loss functions made available by the authors without importance sample weighting. The prior and likelihood distributions of all modalities are assumed to be Gaussian, except for the label modality which is assumed to follow a Bernoulli distribution.

Our evaluation aims at addressing the issues of computational CMI, presented in Section 3. We evaluate Nexus in multiple scenarios with different number (and nature) of modalities (**scalability**). In Section 6.4, we show that the cross-modality *accuracy* metric improves as more modalities are made available to Nexus (**compositionality**). Finally, we show that Nexus is the only model able to generate samples in the high accuracy and low MFD regimes, regardless of the target modality (**generalization**). The computational code employed in this work can be downloaded from https://github.com/miguelsvasco/nexus_pytorch.

6.1. Multimodal handwritten digits dataset

To provide a natural scenario to evaluate CMI and to address the lack of a dataset with a large number of modalities, we contribute the "Multimodal Handwritten Digits" (MHD) dataset, a benchmark dataset containing images, motion trajectories, sounds and labels associated with handwritten digits. The MHD dataset contains 6000 samples per digit class of images, trajectories, sounds and labels, partitioned in 50,000 training and 10,000 testing samples.

To generate the image and trajectory data, examples of which are presented in Fig. 16, we resort to the "UJI Char Pen 2" dataset, from which only one-stroke-formed digits are processed (Llorens et al., 2008). To address the small number of digit samples presented in that dataset, we learn a probabilistic model of each character and re-sample with perturbations constrained in a kinematics feature space, following the procedure described in Yin et al. (2017). This way, we generate 60,000 samples of 28 × 28 greyscale images and 200-dimensional representations of the associated trajectories. Moreover, we normalize all trajectories to the unit interval.

To obtain the sound modality we extract from the "Speech Commands" dataset the samples belonging to the digit classes (Warden, 2018). We process the original sound waves, with sample-rate of 16 384 Hz, by truncating their duration to 1 s and construct a Mel Spectrogram representation, considering a 512 ms hopping window and 128 mel bins. This results in a 128 × 32 representation per audio sample, whose values we normalize to a 0–1 range. As the number of sounds per class is less than the required 6000, we divide proportionally the representations into training and testing partitions and associate each representation with a unique image–trajectory pair by sampling without replacement. We restart the sampling procedure until all pairs have a corresponding sound representation associated.

6.2. Hierarchical representation

In a preliminary evaluation, we appraise the role of hierarchy for CMI. For these initial evaluations we consider a subset of the

[3] The MVAE model is taken from https://github.com/mhw32/multimodal-vae-public and the MMVAE model is taken from https://github.com/iffsid/mmvae.

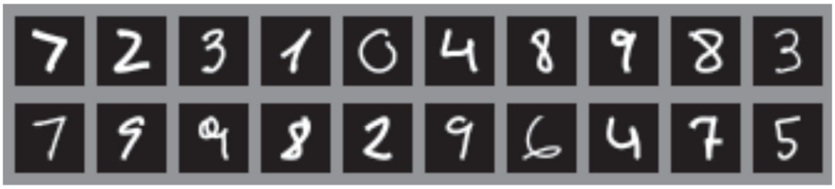


Fig. 16. Samples of images retrieved from the "Multimodal Handwritten Digits" dataset.

Table 1
Results of the hierarchical evaluation of cross-modality accuracy (higher is better), averaged over both modalities, and cross-modality image MFD (lower is better), averaged over all classes. Results averaged over 5 independent runs.

(a) Hierarchical models			(b) Single-level models		
Model	Accuracy (%)	Image MFD	Model	Accuracy (%)	Image MFD
Nexus	93.4 ± 05.3	202.4 ± 44.5	Nexus	96.5 ± 03.6	260.7 ± 54.6
Nexus-0	98.2 ± 01.8	198.0 ± 40.8	Nexus-0	93.0 ± 06.9	272.9 ± 65.8
MVAE	62.9 ± 31.5	237.0 ± 41.1	MVAE	16.5 ± 04.1	150.4 ± 42.2
MMVAE	91.8 ± 03.3	286.9 ± 59.8	MMVAE	91.9 ± 09.0	217.0 ± 75.5

MHD dataset, concerning only the image, x^i, and label, x^l, associated with handwritten digit samples. We show the modality-specific representation spaces of Nexus play a fundamental role in the generation high-quality image samples through CMI. In Appendix A we evaluate the key role of the FPD training scheme in allowing Nexus to learn a multimodal representation robust to missing modalities at test time.

We implement non-hierarchical versions of the Nexus models where we input the modality observations directly into the joint-modality encoder $q_{\phi_t}(z^\sigma \mid x)$. Moreover, to evaluate the potential of hierarchical architectures regardless of the base model, we also extend the baseline models with two representation levels following the Nexus architecture. The hierarchical version of MVAE employs a top-level POE multimodal encoder $q_{\phi_t}(z^\sigma \mid z^{1:M}) \propto p_0(z^\sigma) \prod_{m=1}^{M} q_{\phi_t^m}(z^\sigma \mid z^m)$. The hierarchical version of MMVAE employs a top-level MOE multimodal encoder $q_{\phi_t}(z^\sigma \mid z^{1:M}) = \sum_{m=1}^{M} \frac{1}{M} q_{\phi_t^m}(z^\sigma \mid z^m)$.

All versions share the same encoder–decoder architectures (presented in Appendix B.1), except for the bottom-representation Gaussian sampling layer which is absent in the non-hierarchical versions of the models. We selected $\rho = 0.1$ empirically, for the Nexus models. We consider a 16-dimensional multimodal latent space z^σ, a 64-dimensional image latent space z^i and a 5-dimensional label latent space z^l. The non hierarchical models employ a single 64-dimensional multimodal latent space z^σ.

The quantitative results, averaged over 5 independently-seeded runs, concerning the cross-modality accuracy and image MFD metrics are presented in Table 1 and image samples resulting from cross-modality generation are shown in Fig. 17.

The results show that Nexus is the only model able to perform CMI with both high accuracy and low image MFD. For Nexus the extension to a hierarchical architecture results in a significant decrease on the MFD metric, resulting in higher-quality samples as shown in Fig. 17a. This decrease demonstrates the potential of considering hierarchical representation levels in the architecture of multimodal generative models: the top multimodal representation learns a representation able to generate coherent modality-specific latent samples, of lower dimension and complexity than the modality data itself. The modality-specific generators interpret these latent samples in order to generate high quality modality data. Without hierarchy, the same multimodal representation must be able to encode and generate the modality-data itself, a more complex task than the former. As shown in Table 1, while the accuracy of the generated data is not significantly affected by hierarchy, the image MFD decreases significantly with the hierarchical extension. Regarding the different multimodal fusion solutions, the results show that the naive concatenation solution (Nexus-0) performs on par with the more complex aggregation solution in this simple scenario, attesting once again to the importance of considering hierarchical representation spaces to learn multimodal representations.

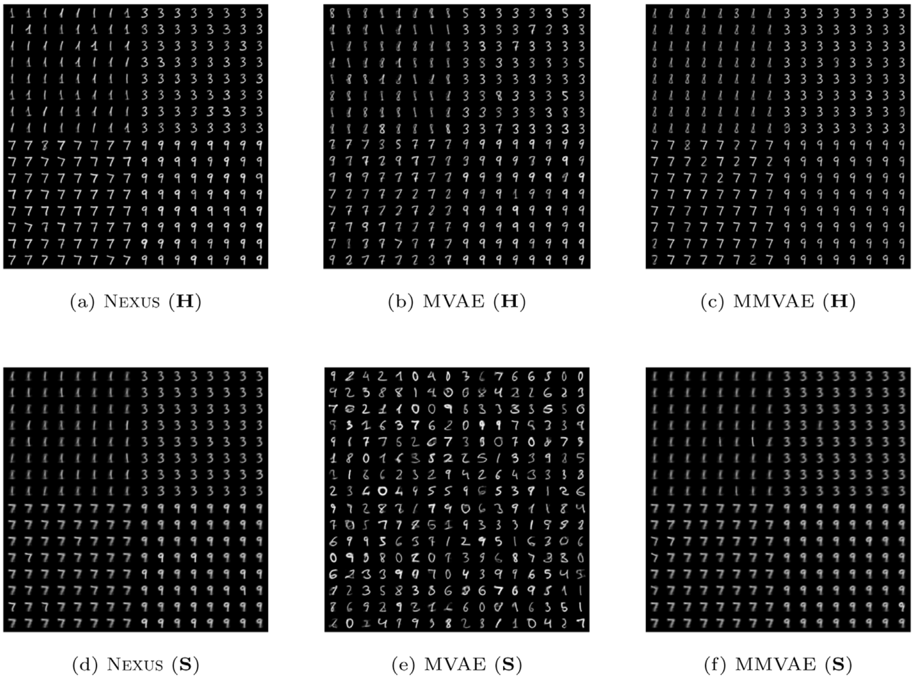
Regarding the MMVAE baseline, a direct comparison of Figs. 17c and 17f shows that the hierarchical extension of the MMVAE is able to generate higher-quality image samples than the single-level version. Such visual inspection seems contrary to the results regarding image MFD in Table 1. This seemingly contradiction is due to the computation of the MFD score: the blurriness in the batches of images generated by MMVAE are interpreted by the auto-encoders as variability. With the hierarchical extension, the images become much higher-quality and the lack of variability becomes evident, as seen in Fig. 17c. Thus, the image MFD score, on average, increases. Nonetheless, despite the hierarchical extension, the generated samples still present high MFD, suggesting that the MoE solution employed by the model is unable to learn a suitable representation of the modalities.

For the MVAE baseline, the same hierarchical extension results in a significant increase in the accuracy metric, but also in the image MFD score. The latter increase is explained by the overconfident expert problem of the single-level MVAE: the model learns a multimodal representation that disregards the information provided by the lower-dimensional modality in scenarios with modalities of distinct complexities (784-dimensional images against 10-dimensional labels). As shown in Fig. 17e, the MVAE model learns a representation that is able to generate high-quality images (low image MFD) at the cost of low accuracy. By extending MVAE with hierarchy, the imbalance between the modalities decreases as the difference in dimensionality of their modality-specific representation spaces is smaller than the difference in dimensionality of the original data. The generation procedure of the hierarchical version MVAE loses the quality provided by the overconfident expert but gains in accuracy, as shown in Fig. 17b. However, the high variance of accuracy reveals that, despite the hierarchical extension, the CMI generation procedure is still not robust across all target modalities.

6.3. Standard datasets

In this section we evaluate the performance of Nexus on two literature-standard datasets: MNIST and FashionMNIST. We evaluate Nexus against the MVAE and MMVAE baselines, regarding single-modality reconstruction accuracy, joint-modality reconstruction accuracy, cross-modality generation accuracy and cross-modality MFD. We employ the same model architectures and training hyperparameters of the prior evaluation. The quantitative results of the evaluation of the MNIST and FashionMNIST datasets are presented in Table 2. All results are averaged over 5 independent runs. In addition, we present image samples generated from label information in Fig. 18.

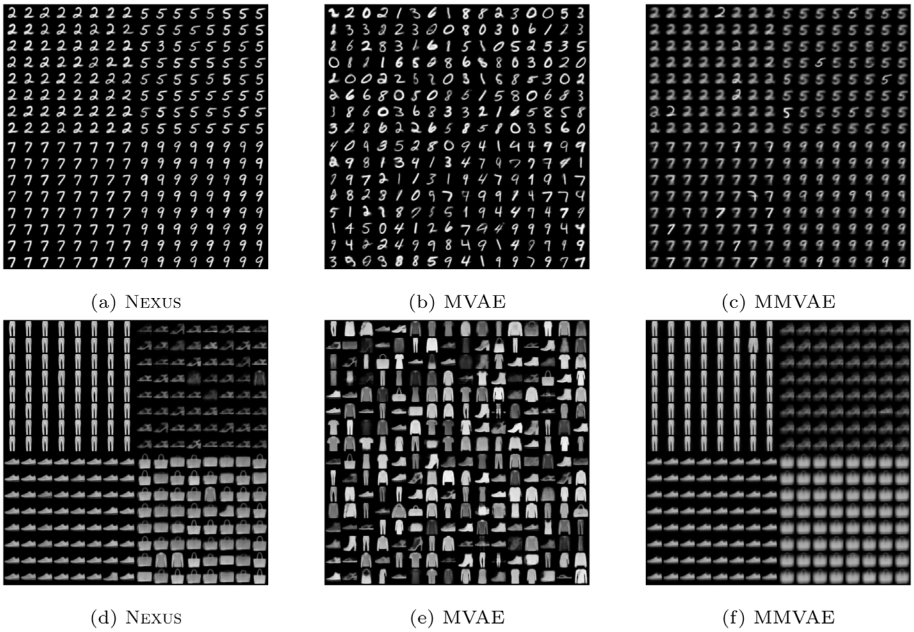
The results on Table 2 show that Nexus is the only model able to encode a multimodal representation capable of generating



(a) Nexus (**H**)　(b) MVAE (**H**)　(c) MMVAE (**H**)

(d) Nexus (**S**)　(e) MVAE (**S**)　(f) MMVAE (**S**)

Fig. 17. Samples of Cross-modality generated images from available label information x^l = {"0", "4", "7", "9"}, provided by hierarchical (a–c) and single-level (d–f) multimodal generative models. Nexus is the only model able to perform CMI and generate samples with high *accuracy* and low *MFD* (best viewed with zoom).



(a) Nexus　(b) MVAE　(c) MMVAE

(d) Nexus　(e) MVAE　(f) MMVAE

Fig. 18. Examples of cross-modality image samples considering the MNIST dataset, from labels x^l = {"2", "5", "7", "9"} (a–d), and considering the FashionMNIST dataset, from labels x^l = {"Trouser", "Sandal", "Sneaker", "Bag"} (e–h). Nexus is the only model able to perform CMI and generate samples with high *accuracy* and low *MFD* (best viewed with zoom).

modality data in the high accuracy and low image MFD regimes, regardless of given single-modality or joint-modality observations. Moreover, in this two-modality scenario, the concatenation solution performs on-par with the aggregator solution, evidence of the fundamental importance of the hierarchical design for the result of these models.

Once again, the results of the MVAE model reveal that the PoE solution employed is the overconfident expert prediction issue: the model unable to generate semantically coherent modality data (low cross-modality accuracy), as seen in Fig. 18. Moreover, the minor increase in accuracy from single-modality observations to joint-modality observations suggests that the model is unable to consider the information provided by the two modalities, hinting once again to the overconfident expert issue. Finally, the MMVAE model is also able to generate semantically-coherent modality data, even outperforming Nexus in the FashionMNIST dataset. It does so, however, at the cost of the quality of the images generated, as shown in Table 2 and Fig. 18, having the highest image MFD in both datasets.

We can further understand the impact of the hierarchical configuration of Nexus on the generation of modality-information

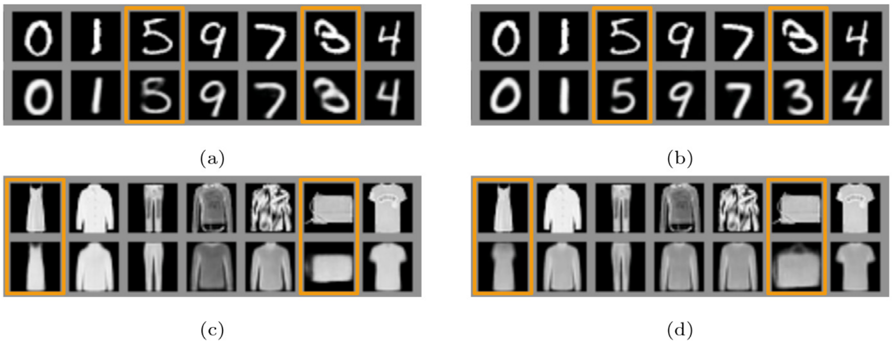

(a) (b) (c) (d)

Fig. 19. Images reconstructed from the image-specific latent space z^i (a, c) and multimodal latent space z^σ (b, d) employing the Nexus model, presenting the original image data (top row) and the reconstructed data (bottom row). We highlight samples (in orange) where the abstraction provided by the contextual multimodal latent space (b, d) allows the generation of more prototypical information, in comparison with the modality-specific reconstructions (a, c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Accuracy and image-specific modality Frechet distance metrics for (a) the MNIST dataset and (b) the FashionMNIST dataset. All results averaged over 5 independent runs.

(a) MNIST				
Model	Single-modality accuracy (%)	Joint-modality accuracy (%)	Cross-modal accuracy (%)	Cross-modal image MFD
Nexus	97.03 ± 02.97	99.98 ± 00.02	97.19 ± 02.88	372.7 ± 91.9
Nexus-0	97.48 ± 02.54	99.96 ± 00.08	97.38 ± 02.67	365.5 ± 85.9
MVAE	96.10 ± 03.61	97.09 ± 02.82	13.48 ± 03.31	285.1 ± 61.7
MMVAE	72.39 ± 28.75	–	68.06 ± 30.83	493.2 ± 134.3
(b) FashionMNIST				
Model	Single-modality accuracy (%)	Joint-modality accuracy (%)	Cross-modal accuracy (%)	Cross-modal image MFD
Nexus	82.00 ± 18.01	88.59 ± 11.43	74.09 ± 02.96	120.2 ± 34.1
Nexus-0	84.14 ± 15.87	89.49 ± 10.64	74.59 ± 03.22	120.7 ± 34.1
MVAE	85.24 ± 14.47	87.79 ± 11.99	20.91 ± 07.95	112.8 ± 40.2
MMVAE	85.46 ± 14.50	–	83.46 ± 06.54	133.4 ± 52.7

from the multimodal representation by considering the modality-information reconstruction procedure. Modality-information can be reconstructed directly at the lower-level modality-specific representation space or be reconstructed from the top-level multimodal representation space.

We present reconstructed samples from both representation spaces for the MNIST and FashionMNIST datasets in Fig. 19. The samples show that the images reconstructed from the multimodal representation z^σ are more prototypical than the samples reconstructed from the modality-specific representation z^i. We can understand such abstraction given the hierarchical nature of the representation spaces in Nexus: the modality-specific representations encode an abstraction of high-dimensional modality information, generating low-dimensional codes. Such low-dimensional codes are encoded and generated by the top multimodal representation space, which learns to generate coherent modality-specific codes, providing another layer of abstraction. While resulting in a lower variability of samples, as shown in the higher image MFD evaluation of Table 2, the abstraction provided by the generation procedure from the multimodal representation provides a significant advantage to Nexus in comparison with non-hierarchical models: the samples generated accentuate the features that unequivocally define the observed phenomena (in this case, digit class correspondence). This allows Nexus to generate coherent modality-information regardless of the target modality and even in scenarios with a high number of modalities.

6.4. Multimodal handwritten digits

We evaluate Nexus in a challenging cross-modality generation scenario that considers the complete set of modalities provided by the MHD dataset: image x^i, trajectory x^t, sound x^s and label x^l. In this task, depicted in Fig. 20, we show that Nexus is the only model able to perform CMI regardless of the target modality and considering the information provided by any subset of available modalities.

Due to the high complexity of the sound modality, we pretrain a SigmaVAE model to learn a modality-specific representation of sound. We resort to the authors' optimal training scheme and consider a regularization hyperparameter $\beta = 10$ (Rybkin, Daniilidis, & Levine, 2020). We employ the pretrained SigmaVAE model as the bottom-level sound-specific encoder and decoder. For a fair comparison, we evaluate our model against the hierarchical versions of the baselines, sharing the same network architectures and training hyperparameters of our own. For this scenario, the dimensionality of the multimodal latent space z^σ is increased to 32, to account for the higher number of modalities. We consider a 64-dimensional image-space z^i, a 128-dimensional sound-space z^s, a 16-dimensional trajectory-space z^t and a 5-dimensional label-space z^l.

We evaluate the cross-modality generation performance for each target modality, as a function of the number of modalities provided to the model. The results are shown in Table 3, averaged over all possible combinations of provided modalities and 5 independent runs. The results show that the Nexus model outperforms the other baselines in accuracy across all target modalities. The accuracy results also show that, in this challenging scenario with a large number of modalities, the naive concatenation approach is unable to encode a suitable representation when provided with incomplete perceptual data. On the other hand, the proposed aggregation joint-modality encoder allows the model to generate semantically correct data (as shown by the high accuracy), regardless of the number of modalities provided. As the number of modalities provided to the model increases, so does the accuracy of the respective cross-modality generated samples, addressing the **compositionality** issue presented in Section 3. This is to be expected as the confidence of the model in generating samples of the correct class increases as more information is provided. Contrary to the MMVAE baseline model, the Nexus is able to take advantage of this additional information provided by multiple modalities.

Regarding the MFD of the generated samples, Table 3 shows that Nexus outperforms the other baselines for the target image modality. For the other modalities, the MVAE baseline is able to generate samples with lower MFD, once again, at the cost

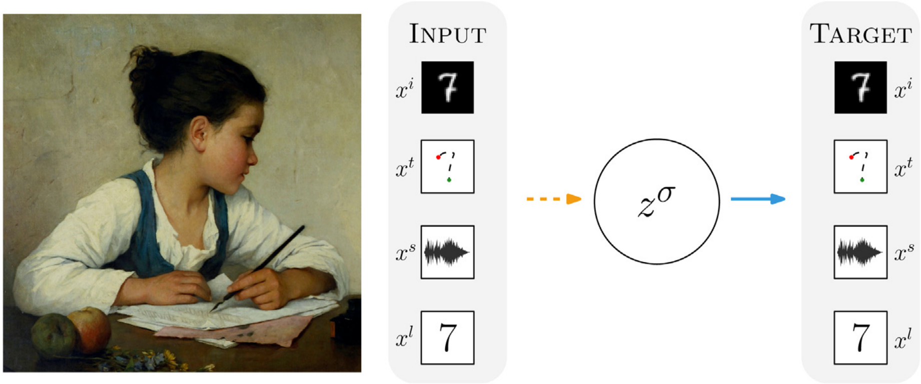


Fig. 20. The cross-modality generation task in the Multimodal Handwritten Digit dataset, considering the image x^i, trajectory x^t, sound x^s and label x^l associated with handwritten digits: given any subset of *input* modalities, the multimodal generative model must be able to generate high-quality, coherent, samples of all *target* modalities. We show that Nexus is the only model able to perform cross-modality inference, generating modality-specific information in the with high accuracy and low MFD regimes.
Source: Image adapted from "A Girl Writing; The Pet Goldfinch" by Henriette Browne (1874).

Table 3
Evaluation in the MHD dataset, as a function of the number of the observed modalities provided to the models and the target cross-modality generated modality (I = Image, T = Trajectory, S = Sound, L = Label). Results for one modality are averaged over all possible combinations of input modalities (excluding the target modality). The results for three modalities consider all modalities (excluding the target modality) in the encoding process. All results are also averaged over 5 independent runs.

Model	Target	Accuracy (%)		MFD	
		1 Modality	3 Modalities	1 Modality	3 Modalities
Nexus	I	84.9 ± 12.4	99.0 ± 00.2	203.9 ± 86.3	76.6 ± 02.8
	T	81.0 ± 10.4	93.8 ± 01.6	618.3 ± 264.0	444.4 ± 36.7
	S	77.2 ± 08.9	94.4 ± 04.2	22393 ± 1659	20239 ± 3411
	L	82.0 ± 06.6	96.9 ± 00.5	NA	NA
Nexus-0	I	49.0 ± 39.3	99.3 ± 00.4	265.2 ± 144.6	94.0 ± 19.2
	T	46.1 ± 37.3	75.1 ± 06.8	625.4 ± 341.9	539.0 ± 220.1
	S	64.8 ± 10.1	60.0 ± 07.3	15452 ± 1152	15778 ± 1196
	L	67.8 ± 04.3	99.2 ± 00.3	NA	NA
MVAE	I	28.6 ± 05.2	80.9 ± 07.2	228.4 ± 61.8	201.3 ± 45.2
	T	13.7 ± 04.6	17.8 ± 03.7	399.3 ± 179.1	391.0 ± 178.7
	S	33.6 ± 14.2	88.6 ± 09.7	6608 ± 1471	8133 ± 1751
	L	23.4 ± 13.4	39.9 ± 07.7	NA	NA
MMVAE	I	66.1 ± 39.8	–	236.9 ± 62.7	–
	T	63.8 ± 38.1	–	547.8 ± 235.4	–
	S	70.4 ± 05.4	–	14998 ± 1325	–
	L	66.0 ± 39.6	–	NA	NA

of the accuracy of these lower dimensional modalities. This is a result of the overconfident expert problem of this model. While the hierarchical extension reduced this effect in the previous scenario, the same extension is less effective in a scenario where the differences in dimensionality of the modality-specific latent spaces are significantly greater.

Overall, the results show that the Nexus model outperforms the baselines by being the only model considered that is able to perform CMI with both high accuracy and low rank, regardless of the target modality considered and the subset of modalities available to the process, addressing the **generalization** issue. The results across all evaluations show that Nexus is able to satisfy all conditions required for effective computational cross-modality inference.

7. Conclusion

In this work, we presented the Nexus model, a hierarchical generative model able to learn a multimodal representation of an arbitrary number of modalities. We have identified three issues in computational CMI that naturally arise from current multimodal generative models. We have shown that, by considering hierarchical representation levels and a novel training scheme, Nexus is able to address simultaneously all those criteria, thus providing a computational model that performs *effective* cross-modality inference. Furthermore, we introduced a novel multimodal benchmark dataset of images, trajectories, sounds and symbols associated with handwritten digits and showed that the Nexus model outperforms the baseline models in this challenging scenario. We attested the importance of leveraging hierarchy for cross-modality generation.

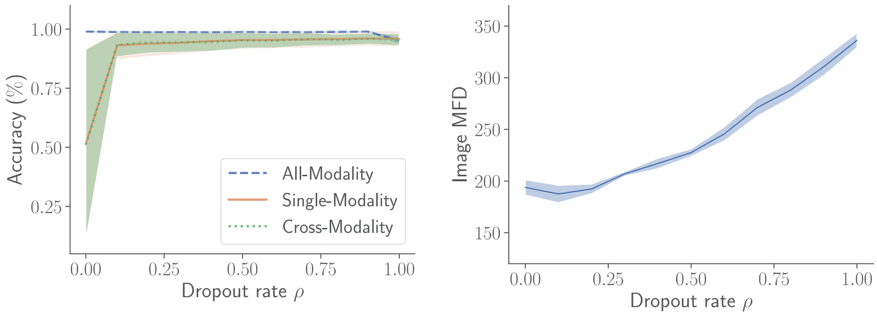
For future work we will address current issues brought up by the hierarchical nature of Nexus. We will explore the modular potential that hierarchy provides to Nexus, by considering pretrained representation models, in scenarios with multiple modalities. We wish to explore the role of disentanglement for multimodal representation and other learning techniques, such as adversarial and contrastive learning, applied to multimodal hierarchical learning. In addition, we will investigate Nexus as a perceptual model of reinforcement learning agents equipped with multiple sensors, in order to explore the robustness of learned task policies in scenarios with changing perceptual conditions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by national funds through the Portuguese Fundação para a Ciência e a Tecnologia under project UIDB/50021/2020 (INESC-ID multi annual funding) and project PTDC/CCI-COM/5060/2021. This work was also suported by funds from Europe Research Council under project BIRD 884887. The first author acknowledges the Fundação para a Ciência e a Tecnologia PhD grant SFRH/BD/139362/2018.



(a)

(b)

Fig. A.21. Results of the evaluation regarding the role of the FPD training mechanism in the CMI performance of the Nexus model, regarding (a) the accuracy and (b) image modality Frechet distance of samples generated by the Nexus model. Results are averaged over 5 independent runs.

Appendix A. Evaluation of the FPD training scheme

We evaluate the role of the Forced Perceptual Dropout (FPD) training scheme, introduced in Section 5.2, in allowing Nexus to perform effective CMI. To be more precise, we consider the effect of the value of the dropout parameter ρ of the FPD training scheme, Eq. (18), on the CMI performance of the Nexus model. We evaluate, for different values of ρ, the accuracy of the samples generated from information provided by both modalities (all-modality generation), provided solely by its own modality (single-modality generation), and provided by the complementary modality (cross-modality generation). In addition we evaluate the modality Frechet distance of the image samples generated from label information. The results are presented in Fig. A.21.

Fig. A.21a shows that Nexus learns a multimodal representation robust to different values of dropout: the accuracy of all-modality generation remains constant until $\rho = 1$, in which case dropout is always performed and the model has never been shown complete multimodal information. However, the occurrence of dropout ($\rho > 0$) during training appears to be fundamental for coherent single-modality and cross-modality generation: if no dropout is applied the model is provided only with complete information during training. This results in a lower average accuracy value with higher variance, as the models learn to prioritize the generation of the higher-dimensional modality (similarly to the MVAE model), regardless of the modality given as input.

Fig. A.21b reveals that ρ also has a significant effect on the MFD of images generated by CMI. As ρ increases, the variability of image information observed and encoded in z^σ decreases. This results in the generation of increasingly prototypical images, with limited variability, increasing the corresponding MFD score.

Appendix B. Network architectures

We present the network architectures and training hyperparameters employed throughout the evaluation. All models were built using Pytorch and evaluated in a machine running Ubuntu 16.04, equipped with a Titan RTX containing 24 GB of dedicated GPU memory and 96 GB of RAM memory. The computational code can be downloaded from https://github.com/miguelsvasco/nexus_pytorch.

B.1. Multimodal generative model architecture

We present the generative model networks employed in this work, which are maintained across all evaluation scenarios. The modality-specific network architectures are presented in Table B.4. The joint-modality *aggregator* encoder and the top-level modality-specific decoders are presented in Table B.5. The baseline models use the bottom-level encoder–decoder models, forcing the different latent spaces to be of common dimensionality.

B.2. Autoencoder and classifier architecture

We present the architecture of the class-specific, modality-specific autoencoders, required to compute the *modality Frechet distance* metric, in Table B.6. We present the architecture of the modality-specific classifiers, required to compute the *accuracy* metric, shown in Table B.7.

B.3. Training hyperparameters

The total loss objective of the Nexus model $\ell(D)$ is given by,

$$\begin{aligned}
\ell(D) &= \ell_b(D) + \ell_t(D) \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M}\Big(\alpha\,\mathrm{KL}\left[q_\eta(z_n^c \mid z_n^{1:M}) \;\|\; p(z_n^c)\right] \\
&+ \beta^m \mathrm{KL}\left[q_{\phi_m}(z_n^m|x_n^m) \;\|\; p(z_n^m)\right] \\
&- \mathbb{E}_{q_{\phi_m}(z_n^m|x_n^m)}\left[\lambda^m \log p_{\theta_m}(x_n^m \mid z_n^m)\right] \\
&- \mathbb{E}_{\substack{q_{\theta_m}(z_n^{1:M}|x_n^{1:M}) \\ q_\eta(z_n^c|z_n^{1:M})}}\left[\gamma^m \log p_{\pi_m}(z_n^m|z_n^c)\right]\Big)
\end{aligned} \tag{21}$$

where we introduce specific training hyperparameters to balance the reconstruction and regularization terms of the different modalities: the λ and γ parameters balance the reconstruction of the modality-specific data and representations; the β and α parameters balance the regularization of the modality-specific and multimodal latent spaces.

We present the training hyperparameters employed in this work for the standard evaluation (Section 6.3) in Table B.8 and multimodal evaluation (Section 6.4) in Table B.9.

Table B.4
Architecture of the bottom-level, modality-specific, networks (best viewed with zoom).

(a) x^i - Image

Encoder	Decoder
Input $\mathbb{R}^{1+28+28}$	Input $\mathbb{R}^D$
Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish	FC, 128 + Swish
Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish	FC, 3136 + Swish
Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish	Transposed Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish
FC, 128 + Swish	Transposed Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish
FC, 128 + Swish	Transposed Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Sigmoid
FC, D, FC, D	–

(b) x^t - Trajectory

Encoder	Decoder
Input $\mathbb{R}^{200}$	Input $\mathbb{R}^D$
FC, 512 + Batchnorm + Leaky ReLU	FC, 512 + Batchnorm + Leaky ReLU
FC, 512 + Batchnorm + Leaky ReLU	FC, 512 + Batchnorm + Leaky ReLU
FC, 512 + Batchnorm + Leaky ReLU	FC, 512 + Batchnorm + Leaky ReLU
FC, D, FC, D	FC, 200 + Sigmoid

(c) x^s - Sound

Encoder	Decoder
Input $\mathbb{R}^{1+128+32}$	Input $\mathbb{R}^D$
Convolutional, 1 × 128 kernel, (1,1) stride, (0,0) padding + Batchnorm + Leaky ReLU	FC, 2048 + Batchnorm + Leaky ReLU
Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU	Transposed Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU
Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU	Transposed Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU
FC, D, FC, D	Transposed Convolutional, 1 × 128 kernel, (1,1) stride, (0,0) padding + Sigmoid

(d) x^l - Label

Encoder	Decoder
Input $\mathbb{R}^{10}$	Input $\mathbb{R}^D$
FC, 128 + Batchnorm + Leaky ReLU	FC, 128 + Batchnorm + Leaky ReLU
FC, 128 + Batchnorm + Leaky ReLU	FC, 128 + Batchnorm + Leaky ReLU
FC, 128 + Batchnorm + Leaky ReLU	FC, 128 + Batchnorm + Leaky ReLU
FC, D, FC, D	FC, 10 + Softmax

Table B.5
Architecture of the (a) joint-modality aggregator encoder and (b) top-level modality-specific decoder networks (best viewed with zoom).

(a)

Encoder			
Input $\mathbb{R}^{D_i}$	Input $\mathbb{R}^{D_t}$	Input $\mathbb{R}^{D_s}$	Input $\mathbb{R}^{D_l}$
FC, d_k	FC, d_k	FC, d_k	FC, d_k
Aggregator function f			
FC, 512 + Batchnorm + Leaky ReLU			
FC, 512 + Batchnorm + Leaky ReLU			
FC, D			

(b)

Decoder
Input $\mathbb{R}^D$
FC, 512 + Batchnorm + Leaky ReLU
FC, 512 + Batchnorm + Leaky ReLU
FC, 512 + Batchnorm + Leaky ReLU
FC, D_m

Table B.6
Architecture of the modality-specific autoencoder models (best viewed with zoom).

(a) Image autoencoder, with $|h| = 128$.

Encoder	Decoder
Input $\mathbb{R}^{1+28+28}$	Input $\mathbb{R}^{D}$
Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish	FC, 128 + Swish
Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish	FC, 3136 + Swish
Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish	Transposed Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish
FC, 128 + Swish	Transposed Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Swish
FC, 128 + Swish	Transposed Convolutional, 4 × 4 kernel, 2 stride, 1 padding + Sigmoid
FC, D, FC, D	–

(b) Trajectory autoencoder, with $|h| = 64$.

Encoder	Decoder
Input $\mathbb{R}^{200}$	Input $\mathbb{R}^{D}$
FC, 512 + Batchnorm + Leaky ReLU	FC, 512 + Batchnorm + Leaky ReLU
FC, 512 + Batchnorm + Leaky ReLU	FC, 512 + Batchnorm + Leaky ReLU
FC, 512 + Batchnorm + Leaky ReLU	FC, 512 + Batchnorm + Leaky ReLU
FC, D, FC, D	FC, 200 + Sigmoid

(c) Sound autoencoder, with $|h| = 512$.

Encoder	Decoder
Input $\mathbb{R}^{1+128+32}$	Input $\mathbb{R}^{D}$
Convolutional, 1 × 128 kernel, (1,1) stride, (0,0) padding + Batchnorm + Leaky ReLU	FC, 2048 + Batchnorm + Leaky ReLU
Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU	Transposed Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU
Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU	Transposed Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU
FC, D, FC, D	Transposed Convolutional, 1 × 128 kernel, (1,1) stride, (0,0) padding + Sigmoid

Table B.7
Architecture of the modality-specific classifier models (best viewed with zoom).

(a) Image classifier
Input $\mathbb{R}^{1+28+28}$
Convolutional, 5 × 5 kernel, 1 stride, 0 padding + ReLU + Dropout($p = 0.2$)
MaxPool (2,2)
Convolutional, 5 × 5 kernel, 1 stride, 0 padding + ReLU + Dropout($p = 0.2$)
MaxPool (2,2)
FC, 128 + Dropout($p = 0.2$)
FC, 64 + Dropout($p = 0.2$)
FC, 10

(b) Trajectory classifier
Input $\mathbb{R}^{200}$
FC, 512 + Batchnorm + Leaky ReLU
FC, 512 + Batchnorm + Leaky ReLU
FC, 128 + Batchnorm + Leaky ReLU
FC, 10

(c) Sound classifier
Input $\mathbb{R}^{1+128+32}$
Convolutional, 1 × 128 kernel, (1,1) stride, (0,0) padding + Batchnorm + Leaky ReLU
Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU
Convolutional, 4 × 1 kernel, (2,1) stride, (1,0) padding + Batchnorm + Leaky ReLU
FC, 128 + Batchnorm + Leaky ReLU
FC, 64 + Batchnorm + Leaky ReLU
FC, 10

Table B.8
Training hyperparameters for the standard evaluation scenario, presented in Section 6.3.

Parameter	Value
Training Epochs	100
Learning rate	10^{-3}
Batch-size	64
Optimizer	Adam
λ^{i}	1.0
λ^{l}	50.0
β^{i}	1.0
β^{l}	1.0
γ^{i}	1.0
γ^{l}	50.0
β^{c}	1.0

Table B.9
Training hyperparameters for the multimodal evaluation scenario, presented in Section 6.4.

Parameter	Value
Training Epochs	100
Learning rate	10^{-3}
Batch-size	64
Optimizer	Adam
λ^{i}	1.0
λ^{t}	50.0
λ^{s}	1.0
λ^{l}	50.0
β^{i}	1.0
β^{t}	1.0
β^{s}	1.0
β^{l}	1.0
γ^{i}	1.0
γ^{t}	50.0
γ^{s}	1.0
γ^{l}	50.0
β^{c}	1.0

References

Bell, A. H., Meredith, M. A., Van Opstal, A. J., & Munoz, D. P. (2005). Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. *Journal of Neurophysiology*, *93*(6), 3659–3673.

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, *40*(5), 1053–1065.

Burianová, H., Marstaller, L., Sowman, P., Tesan, G., Rich, A. N., Williams, M., et al. (2013). Multimodal functional imaging of motor imagery using a novel paradigm. *Neuroimage*, *71*, 50–58.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex*, *11*(12), 1110–1123.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1–2), 25–62.

Daunhawer, I., Sutter, T. M., Marcinkevičs, R., & Vogt, J. E. (2021). Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models. In *Pattern recognition: 42nd dagm german conference, dagm gcpr 2020, tübingen, germany, september 28–october 1, 2020, proceedings 42* (pp. 459–473). Springer.

Edwards, S. B., Ginsburgh, C. L., Henkel, C. K., & Stein, B. E. (1979). Sources of subcortical projections to the superior colliculus in the cat. *Journal of Comparative Neurology*, *184*(2), 309–329.

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., et al. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204–1210.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., et al. (2018). Neural processes. arXiv preprint arXiv:1807.01622.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–285.

González, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuán, A., Belloch, V., et al. (2006). Reading cinnamon activates olfactory brain regions. *Neuroimage*, *32*(2), 906–912.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in neural information processing systems* (pp. 1024–1034).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems* (pp. 6626–6637).

Hsu, W.-N., & Glass, J. (2018). Disentangling by partitioning: A representation learning framework for multimodal sensory data. arXiv preprint arXiv:1805.11264.

Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The sound of concepts: four markers for a link between auditory and conceptual brain systems. *Journal of Neuroscience*, *28*(47), 12224–12230.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Korthals, T., Rudolph, D., Leitner, J., Hesse, M., & Rückert, U. (2019). Multi-modal generative models for learning epistemic active sensing. In *2019 ieee international conference on robotics and automation*.

Lee, M., & Pavlovic, V. (2020). Private-shared disentangled multimodal VAE for learning of hybrid latent representations. arXiv preprint arXiv:2012.13024.

Llorens, D., Prat, F., Marzal, A., Vilar, J. M., Castro, M. J., Amengual, J.-C., et al. (2008). The ujipenchars database: a pen-based database of isolated handwritten characters.. In *Lrec*.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., et al. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International conference on machine learning* (pp. 4114–4124). PMLR.

Ma, S., McDuff, D., & Song, Y. (2019). Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proceedings of the ieee international conference on computer vision* (pp. 7598–7607).

Man, K., Kaplan, J., Damasio, H., & Damasio, A. (2013). Neural convergence and divergence in the mammalian cerebral cortex: from experimental neuroanatomy to functional neuroimaging. *Journal of Comparative Neurology*, *521*(18), 4097–4111.

Marstaller, L., & Burianová, H. (2014). The multisensory perception of co-speech gestures–a review and meta-analysis of neuroimaging studies. *Journal of Neurolinguistics*, *30*, 69–77.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316–322.

Meyer, K., & Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Neurosciences*, *32*(7), 376–382.

Nanay, B. (2018). Multimodal mental imagery. *Cortex*, *105*, 125–134.

Rybkin, O., Daniilidis, K., & Levine, S. (2020). Simple and effective VAE training with calibrated decoders. arXiv preprint arXiv:2006.13202.

Shi, Y., Siddharth, N., Paige, B., & Torr, P. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in neural information processing systems* (pp. 15692–15703).

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995.

Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2020). Multimodal generative learning utilizing jensen-Shannon-divergence. arXiv preprint arXiv:2006.08242.

Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint multimodal learning with deep generative models. arXiv preprint arXiv:1611.01891.

Tian, Y., & Engel, J. (2019). Latent translation: Crossing modalities by bridging generative models. arXiv preprint arXiv:1902.08261.

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., & Salakhutdinov, R. (2018). Learning factorized multimodal representations. arXiv preprint arXiv:1806.06176.

Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2017). Generative models of visually grounded imagination. arXiv preprint arXiv:1705.10762.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., et al. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21–25.

Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209.

Wu, M., & Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. In *Advances in neural information processing systems* (pp. 5575–5585).

Yan, C., Hao, Y., Li, L., Yin, J., Liu, A., Mao, Z., et al. (2021). Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yan, C., Li, Z., Zhang, Y., Liu, Y., Ji, X., & Zhang, Y. (2020). Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(4), 1–17.

Yan, C., Shao, B., Zhao, H., Ning, R., Zhang, Y., & Xu, F. (2020). 3D room layout estimation from a single RGB image. *IEEE Transactions on Multimedia*, *22*(11), 3014–3024.

Yin, H., Melo, F. S., Billard, A., & Paiva, A. (2017). Associate latent encodings in learning from demonstrations. In *Thirty-first aaai conference on artificial intelligence*.